Computer Science and Information Systems 21(4):1865–1885 https://doi.org/10.2298/CSIS240418061G

MFE-Transformer: Adaptive English Text Named Entity Recognition Method Based on Multi-feature Extraction and Transformer

Liuxin Gao

School of Foreign Languages, Zhengzhou University of Science and Technology
450064 Zhengzhou, China
publicgj@163.com

Abstract. English text named entity recognition aims to alleviate the problem of insufficient labeling data in the target domain. Existing methods usually use feature representation or model parameter sharing to realize cross-domain transfer of entity recognition capability, but there is still a lack of full utilization of structured knowledge in text sequences. Therefore, this paper proposes an adaptive English named text entity recognition method based on multi-feature extraction and transformer. Firstly, a bidirectional long term memory conditional random field entity recognition model based on BERT pre-trained language model is constructed on a generic domain dataset. In the training process, the weights of two character vectors of text words are dynamically calculated and combined, which makes the model make full use of the information in the character granularity, and the parts-of-speech information and block analysis are added as additional features. The word vectors, character-level features and additional features are spliced into the BiLSTM-CRF neural network model for training. Finally, experiments are carried out on five English datasets and specific cross-domain named entity recognition datasets respectively. The results show that the average performance of the proposed model is improved by 0.43% and 1.47% compared with the current cross-domain model, indicating that the structured knowledge in feature representation can effectively improve the entity recognition capability of the target domain.

Keywords: Adaptive English named text entity recognition, multi-feature extraction, transformer, domain invariant knowledge.

1. Introduction

Named entity recognition (NER) is a fundamental information extraction task that plays a crucial role in natural language processing applications such as information retrieval, automatic text summarization, intelligent question answering, machine translation4, and knowledge graphs. NER is to extract some predefined specific entity from the sentence and identify its correct type, such as person, place, organization [1-3].

Early NER methods can be divided into two types: rule-based methods and statistics-based methods. The rule-based approach is to manually design a large number of rules for specific fields to match named entities according to the task, and generalize and restrict them to other fields. Therefore, a rule-based approach is time-consuming and expensive. A statistics-based approach transforms NER tasks into sequential labeling tasks and trains

them using a corpora of manual labeling. Because the cost of labeling is much lower than the cost of designing rules, statiscy-based methods became mainstream before the explosion of deep learning, such as hidden Markov models (HMM) [4] or conditional random fields (CRF) [5].

NER has gone through three major stages of development, rule-based recognition methods, statistics-based machine learning methods and depth-based learning methods. In NER, the appearance of BiLSTM-CRF has opened the prelude of named entity recognition in deep learning. Its emergence makes the model more concise and robust, and becomes the deep learning benchmark for solving NER problems. The representation layer is based on character information, adding radicals, glyphs, parts of speech, pronunciation, dictionaries and other information on the basis of character information. Text encoding can use neural networks to obtain context dependencies. The label decoding layer predicts and labels the input sequence. These models have their own advantages and disadvantages when dealing with entities [6-8]. For example, CNN can parallelize data, so the calculation speed is faster, but there is a problem of memory loss of context information. LSTM is a variant of RNN that is efficient at learning long-haul dependent information, but there are still gradient problems. GNN can mine the relationship between entities more efficiently by constantly mining the graph data model, but the model structure is too large, so the flexibility and expansion is poor. Transformer is often pre-trained in combination with BERT to generate deep language features, but it consumes a lot of computing resources. Therefore, the selection model should be analyzed according to the specific situation.

Character representation is the focus of NER research. Simple external features such as parts of speech, radicals, and strokes, which contain less information, are effective in the absence of enough information, and perform well before the proposed BERT pre-trained model. However, with the development of pre-trained language models such as BERT, pre-trained character representations can capture most of the semantic information of Chinese characters, so it is difficult for NER models to benefit from simple external features when using these pre-trained language models. In addition, it is necessary and valuable to learn from the existing English NER method to solve the Chinese NER problem. Uzair et al. [9] showed that the randomly initialized output layer performed better in the target domain than the output layer in the migrated source domain. The word embedding layer and the hidden layer had better generalization ability, and could be effectively mapped to the semantic vector space of the target domain after migrating to the target domain. Liang et al. [10] used BERT-WWM (whole word mask) model in the Chinese field. The biggest difference between Ernie and the original model was that it adopted the entity-level masking training strategy and combined external knowledge to train the BERT model. The performance of NER experiment conducted in the Chinese field is better than that of the original BERT. Although these models have reached the state of the art (SOTA) in the general field, whether the pre-trained model trained on a large number of unstructured corpus can improve the NER task effect in a specific field still needs to be verified by relevant research [11,12].

To solve these problems, this paper proposes an adaptive English named text entity recognition method based on multi-feature extraction and transformer. The main contributions of this paper are as follows:

1. Firstly, a bidirectional long term memory conditional random field entity recognition model based on BERT pre-trained language model is constructed on a generic domain dataset.
2. In the training process, the weights of two character vectors of text words are dynamically calculated and combined, which makes the model make full use of the information in the character granularity, and the parts-of-speech information and block analysis are added as additional features.
3. The word vectors, character-level features and additional features are spliced into the BiLSTM-CRF neural network model for training.

2. Related Works

2.1. Pre-trained Model

In NER deep learning model, word embedding is a common data preprocessing method that can learn word vector representations and capture the corresponding semantic and syntactic information of sentences.

Pokharel et al. [13] provided a comprehensive review of the pre-trained language model of NLP, and divided PTM into pre-trained word embeddings and pre-trained context encoders. On the basis of classification, pre-trained character embeddings are divided into static embeddings and dynamic contextual embeddings. Static embeddings are trained to look up tables in which the embeddings of each character are fixed, such as NNLM, Word2vec, FastText, Glove, etc. However, because the training result of static embedding is a fixed word vector matrix, which cannot be dynamically modified, it cannot solve the polysemy phenomenon and realize the real semantic understanding of text.

Dynamic context embedders are also known as pre-trained language models, which generate representations that change depending on the context, such as ELMo, BERT, ERNIE, ALBERT, NEZHA, etc. BERT is the most commonly used. For a given character, BERT takes its character position embedding, sentence position embedding and character embedding as input, and then uses mask language model (MLM) to pre-train the input sentence with deep bidirectional representation to obtain robust contextual character embedding. Because of BERT excellent performance, some scholars began to explore how to achieve similar results with fewer training resources [14-16]. Therefore, RoBERTa, SpanBERT and other improved models based on BERT model have been proposed successively. Such models limit the length of input sequences and thus do not perform well on generative tasks such as automatic text summarization. To solve this problem, XLNet extracts long text features by applying Transformer-XL. In addition, the ERNIE model [17] also focuses more on word vector training in the Chinese field.

2.2. The NER Correlation Between Chinese and English

In recent years, the superior nature of deep learning in the field of NLP, which can learn feature representations directly from data, has brought significant breakthroughs in the field. In terms of English named entity recognition, deep learning models have also significantly improved their performance. At the same time, compared with Chinese named entity recognition technology, NER technology of English text has reached a relatively

mature stage due to the unique English word segmentation rules, that is, the natural space barrier between words. Therefore, in recent years, scholars at home and abroad began to apply English NER technology to Chinese named entity recognition, because both Chinese and English texts have obvious grammatical and lexical characteristics. Secondly, both Chinese and English are context-important languages, and the context information of entities has an important impact on entity recognition. In addition, the problems faced by Chinese NER and English NER are similar, for example, the problem of unknown words. With the development of various fields and the era of big data, there will be a large number of new entities, but these new entities lack uniform naming rules in the dictionary. Therefore, named entity recognition (NER) requires strong contextual reasoning ability to identify nested entities in both English and Chinese, including both outer and inner entities. This is one of the hot topics in current NER research. At the same time, both Chinese and English NER have the problem of text ambiguity, and the same text may represent different entity types in different positions, so it is necessary to conduct entity disambiguation before NER [18,19].

To sum up, due to the particularity of Chinese language and the late start of NER in Chinese, some deep learning methods directly applied to Chinese named entity recognition tasks cannot achieve the same good results as English named entity recognition. Therefore, it is difficult to recognize Chinese named entities. The particularity and difficulty of NER in Chinese are reflected in the following points:

1. The boundary of Chinese words is fuzzy. Unlike English texts, Chinese texts do not have displayed separators (such as spaces) and obvious parts-of-speech changes (for example, place names and personal names in English are capitalized) as boundary markers. Therefore, it is difficult to determine the segmentation boundary.
2. Nested entities. Entity contains other entities or is contained by other entities, and it is one of the current research hotspots to identify both internal and external entities.
3. Entity ambiguity. In the results of entity recognition, there may be different references to the same entity, or there may be multiple meanings of one word, which will lead to inaccurate and ambiguous results of entity recognition. Therefore, it is necessary to disambiguate the entity recognition results before obtaining accurate and unambiguous information.
4. Low resource NER. At present, for limited domains and limited entity types, named entity recognition can achieve good recognition results in these places, and cannot be transferred to other specific domains.

2.3. Flat Solid Boundary Problem

The process of named entity recognition usually includes two parts: (1) entity boundary recognition; (2) Identify the type of entity (person, place, institution or other) [20-22]. So determining the entity boundary plays an important role in named entity recognition. NER entities can often be identified by some obvious formal sign, such as a place or person entity capitalized in a word. Therefore, entity boundary identification is relatively easy in English. However, compared with English, Chinese named entity recognition task is more complicated. This is because Chinese entities often do not have obvious formal signs, and the composition of entities is more complex. Compared with entity category labeling subtask, entity boundary recognition is more difficult in Chinese NER task. Therefore,

in the Chinese NER task, it is necessary to adopt more complex and fine algorithms for entity recognition and boundary recognition to achieve higher accuracy and recall rate.

Some researchers have used character-based approach to solve NER, which has achieved good performance, but can not use word boundary and word order information to determine the entity boundary. In recent years, with the introduction of deep learning, NER research mainly focuses on the feature that there is no clear boundary between Chinese words. In the process of research, it is found that the introduction of external resources can provide boundary information for the Chinese named entity recognition model which is not completely based on words. Thus, the model performance is improved, which is regarded as one of the auxiliary tools to improve the model performance. Therefore, the method of determining entity boundary can be roughly summarized into two aspects: word segmentation and Chinese word feature fusion. Gate Recurrent Unit (GRU) is a special kind of Recurrent Neural Network (RNN) [23,24]. GRU is similar to LSTM in performance but simpler and more efficient, and the training time is greatly shortened while retaining important timing features. Yu et al. [25] used GRU to capture the long-distance features of Chinese characters and applied them to Chinese named entity recognition, and achieved good results. Setiawan et al. [26] used BiGRU to extract character-level word vectors as part of the input in fine-grained sentiment analysis, and the results showed that character-level features extracted by BiGRU had a positive impact on model performance.

Relevant studies have shown that character-level features are effective in biomedical named entity recognition tasks, but different extractors have different character-level features, and the effect of using a single character-level feature extractor is limited. Therefore, this paper uses CNN and BiLSTM to extract different types of character-level features, adaptively integrates the two character-level features in different contexts, and proposes an adaptive named entity recognition model based on character-level features.

3. Proposed Adaptive Entity Recognition Method

3.1. Character Level Feature Extraction

Character-level features have been proved to be effective in various NLP tasks and can improve the performance of such tasks. References [27,28] shows that character-level features can significantly improve the performance of machine translation. Tran et al. [29] had applied character-level features to text classification to improve the performance to some extent. The advantage of using character-level features was that they could be extracted directly from the original text without designing additional manual features, and there was no need for complex preprocessing of the original corpus. In this paper, CNN, BiLSTM and Bidirectional Gate Recurrent Unit (BiGRU) are used to extract character-level features of words.

A. Character level CNN model

CNN is suitable for feature extraction of presuffixes and composition of words. The structure of CNN is shown in Figure 1, and the specific process is as follows. Each word in the original text is disassembled and filled to the maximum word length, so that the dimension size of the character vector matrix is consistent. The local features of the current word are extracted from the matrix formed by the characters of each word through convolution operation. The size of convolution kernel determines the range of local features

that can be extracted by CNN. The key information in the feature is extracted through the pooling process. The result is a 30-dimensional character vector.

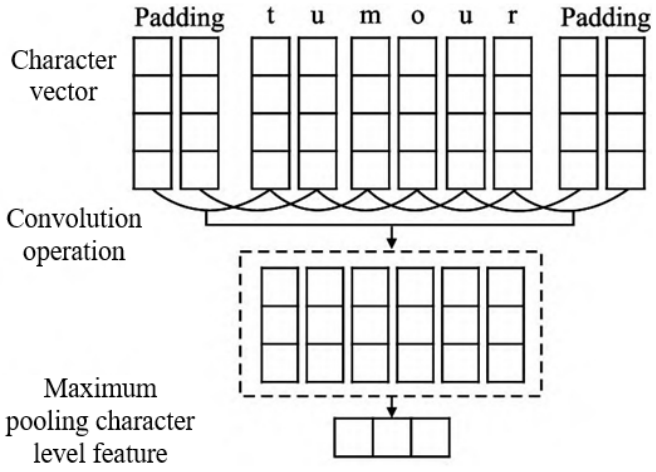

Fig. 1. Schematic diagram of character level feature extraction by CNN

B. Character level BiLSTM model

BiLSTM is good for spelling information about words. The structure of character features extracted by BiLSTM is shown in Figure 2. First, each character of the word is read from front to back to calculate the forward character vector V_f, and then the backward character vector is calculated by reading the characters of the word from back to front. The forward vector and the backward vector are connected in series to obtain the vector V_T of the word character level. V_T retains both forward and backward information and is ultimately represented as a 30-dimensional character vector.

C. Character-level BiGRU model

Different character-level feature extraction models have different characteristics. CNN is suitable for extracting local features and representing information such as presuffixes and compositions of words, while it lacks attention to long-distance dependent information. BiLSTM and BiGRU pay more attention to timing features, which are suitable for representing the character spelling information of words, and pay less attention to local features. Deng et al. [30] directly spliced the character-level features extracted by CNN and BiLSTM to form combined character features and took them as part of the input in biomedical named entity recognition. The experimental results showed that the effect of combined features was better than that of single features. Xu et al. [31] extracted character-level features from BiGRU and then extracted local features in the convolution process for Chinese named entity recognition, and achieved good results.

Different character-level feature extractors focus on different points. When using different extractors to extract features, if different weights can be assigned to different features, so that the model can make full use of information in character granularity, it will have a positive impact on the performance of the model. For a sentence

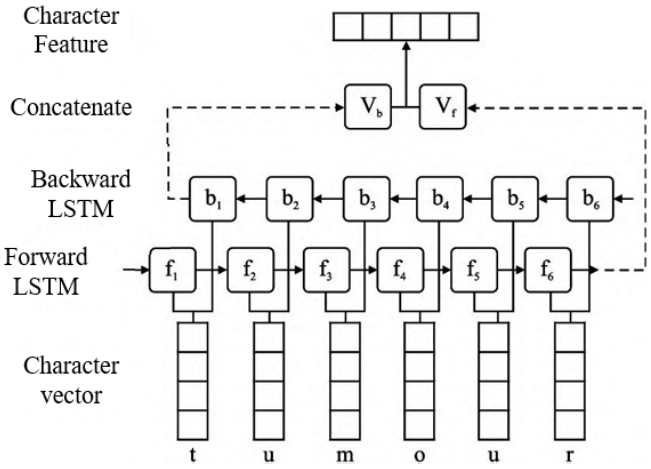

Fig. 2. Schematic diagram of character level feature extracted by BiLSTM

$S = W_1, W_2, \cdots, W_n$, CNN is used to extract the local feature V_P of the word character level in S, and one of BiLSTM or BiGRU is used to extract the sequence feature V_T of the character level. V_P is consistent with the dimension of V_T, and $\tilde{V}$ is used to represent the character-level combination feature.

The first strategy. According to the gradient method, the character features extracted by different extractors are given different weights on the macro level to test the degree of influence on the performance of the model. For example, the weight assigned to V_P is α_0, the weight assigned to V_T is β_0, and $\oplus$ represents the concatenation operation. The combination process satisfies formula (1), and $0.1 \leq \alpha_0 \leq 1.9, 0.1 \leq \beta_0 \leq 1.9$. When α_0 and β_0 are equal, it is equivalent to V_T, and $\tilde{V}$ and is directly concatenated. The process of generating character-level combination feature $\tilde{V_0}$ under this strategy can be expressed as:

$$\alpha_0 + \beta_0 = 2. \tag{1}$$

$$\tilde{V_0} = \alpha_0 V_P \oplus \beta_0 V_T. \tag{2}$$

Artificially controlling the weights of different types of character features may have positive or negative effects on model performance, which is highly uncertain. Therefore, the second strategy is proposed. To calculate dynamic weights in the process of model training, different types of extractors are selected to extract character level features of words. In the training process, the weight of each feature is dynamically calculated, so that the weight of important features becomes larger, the weight of unimportant features becomes smaller, the important part is strengthened, and the unimportant part is weakened. The process of generating character-level combination feature $\tilde{V_1}$ under the adaptive strategy can be expressed as:

$$Mish(x) = x \times (\tanh(\ln(1 + e^x))). \tag{3}$$

$$z_1 = \sigma(Mish(V_P)). \tag{4}$$

$$z_2 = \sigma(Mish(V_T)). \tag{5}$$

$$\alpha_1 = 2 \times \frac{z_1}{z_1 + z_2}. \tag{6}$$

$$\beta_1 = 2 - \alpha_1. \tag{7}$$

$$V'_P = \alpha_1 \times V_P. \tag{8}$$

$$V'_T = \beta_1 \times V_T. \tag{9}$$

$$\tilde{V_1} = V'_P \oplus V'_T. \tag{10}$$

In the above formula, $Mish$ represents the Mish activation function, which allows better information to penetrate into the neural network. σ is the sigmoid activation function. z_1 and z_2 are transition matrices after a series of nonlinear transformations. Each number in z_1 and z_2 ranges from 0 to 1. α_1 represents the weight matrix of the local eigenmatrix. β_1 represents the weight matrix of the temporal eigenmatrix. For each word in the input text, the model selectively strengthens or weakens the character-level local features corresponding to the word, and weakens or strengthens the character-level temporal features of the word. The operation trend of the two features is opposite. Finally, the two transformed character level features are concatenated to obtain character level combination features.

3.2. BERT

In the whole pre-training process, BERT adopts Transformer Encoder model to realize joint context constraint optimization among all layers, realize left-to-right bidirectional training, and can also carry out cross-information transfer between upper and lower layers, instead of just focusing on the text information of the current input words.

The Transformer Encoder consists of a multi-head self-attention mechanism layer and a feed-forward neural network layer. The multi-head attention mechanism layer is mainly composed of multiple attention mechanism units, and a single attention mechanism unit adopts the principle of scaled dot-product attention (SDPA) [32]. A fully bidirectional attention value is calculated for each input. After adding the input embedded word vector I to the positional encoding matrix representing each word position, the Q (Query) matrix, K (Key) matrix, and V (Value) matrix are generated by multiplying the matrix weights W_Q, W_K, and W_V. The Q matrix $[Q_1, Q_2, Q_3, \cdots]$ is generated by the current word. The K matrix $[K_1, K_2, K_3, \cdots]$ and V matrix $[V_1, V_2, V_3, \cdots]$ are generated by context words. As shown in formula (11), in a single attention mechanism unit, the input Q matrix of the current word is multiplied by $[K_1, K_2, K_3, \cdots]$ and then divided by the dimension d_k of the K matrix. After normalization by Softmax, the weight matrix is obtained and multiplied by $[V_1, V_2, V_3, \cdots]$ to obtain the self-attention unit Z_i. As shown in equation

(12), all self-attention units Z_i are splicing and linear mapping is performed through W to obtain multi-head attention matrix Z, which is used to reflect the relationship between current words and context words. Finally, Z is added and regularized with the original input embedding word vector I, and then added and regularized with the feedforward layer input matrix again through the feedforward neural network to obtain Transformer encoding vector.

$$Z_i = Softmax(\frac{Q \times K^T}{\sqrt{d_k}})V. \tag{11}$$

$$Z = Concat(Z_1, \cdots, Z_h)W. \tag{12}$$

3.3. Source Domain Pretrained LSTM Layer

LSTM neural network model has been widely used in feature extraction for sequence annotation tasks. LSTM can be used to train traditional news models with rich corpus sets. Reference [33] realized the fast training of the source domain model, as shown in Figure 3.

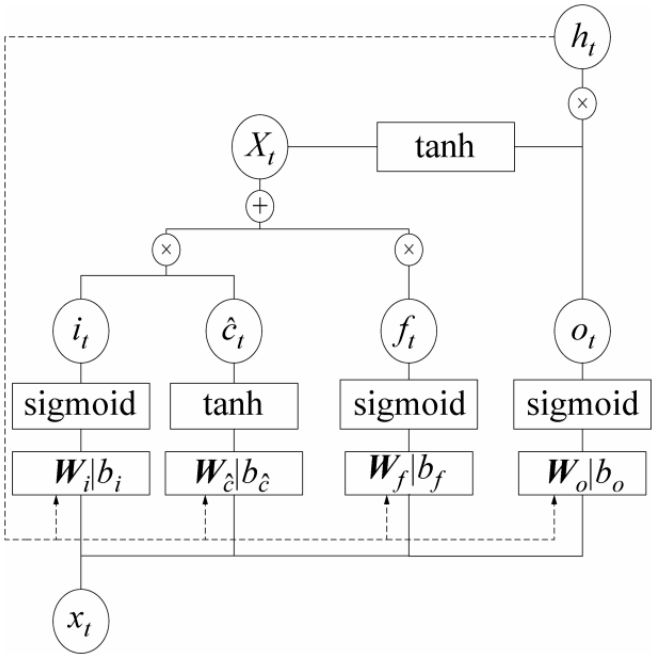

Fig. 3. LSTM structure

In the structure diagram of the LSTM unit, i_t, f_t and o_t respectively represent the states of the input gate, forget gate and output gate in the LSTM unit at time t. h_{t-1} indicates the hidden state at time $t-1$. c_t represents the cell state at time t. σ indicates

Sigmoid. $tanh$ represents the hyperbolic tangent excitation function, as shown in equations (13)-(18).

$$i_t = \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i). \tag{13}$$

$$f_t = \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f). \tag{14}$$

$$\tilde{c}_t = \sigma(W_{ch}h_{t-1} + W_{cx}x_t + b_c). \tag{15}$$

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t. \tag{16}$$

$$o_t = \sigma(W_{oh}h_{t-1} + W_{ox}x_t + b_o). \tag{17}$$

$$h_t = o_t \tanh(c_t). \tag{18}$$

At t time, the word vector of BERT pre-training layer is used as the input of LSTM layer, and the hidden state of LSTM at $t-1$ time is combined to obtain the hidden state of t time. Finally, the output hiding state of BiLSTM can be obtained by combining the forward LSTM hiding state and the reverse LSTM hiding state.

3.4. BiLSTM Output Adaptation Layer Integrating Attention Mechanism

By introducing the attention mechanism to extract the feature vector in the context, the key words in the text will get more efficient feature extraction, so that the feature vector will integrate more contextual semantics. For the input sentence $L = [x_1, x_2, x_3, \cdots]$, it inputs the current word x_t of the BERT embedding word vector at time t. Firstly, the attention weight α_{tj} is calculated, where the greater the value of α_{tj} indicates that the current word x_t is more influenced by x_j, as shown in equation (19).

$$\alpha_{tj} = \frac{exp(score(x_t, x_j))}{\sum_{j=1}^{n} exp(score(x_t, x_j))}. \tag{19}$$

This paper uses the following four alignment algorithms to calculate $score(x_t, x_j)$ to measure the degree of association between the current word and the context word including Manhattan distance; Euclidean distance; Cosine distance; Perceptron mechanism. Here, the greater Manhattan distance and Euclidean distance denote the higher degree of correlation. The larger cosine distance and the perceptron denote the smaller correlation degree. The final result is $score(x_t, x_j) = a + b - c - d$, as shown in equation (20).

$$score(x_t, x_j) = \begin{cases} W_a|x_t - x_j| \\ W_a(x_t - x_j)^T(x_t - x_j) \\ \dfrac{W_a(x_t \cdot x_j)}{|x_t| \cdot |x_j|} \\ \tanh(W_a[x_t; x_j]) \end{cases} \tag{20}$$

Then, the output hidden state h_j and the attention weight α_{tj} at each moment of the BiLSTM frozen layer of the pre-trained source domain are weighted and summed to obtain the global variable g_t. After concatenating with the output hidden state h_t at the current moment, the linear mapping is carried out through matrix W_g, and the output M_t of the attention layer is obtained through $tanh$ activation function. Further feature extraction is carried out by randomly initialized BiLSTM, as shown in equations (21)-(22).

$$g_t = \sum_{j=1}^{N} \alpha_{tj} h_j. \tag{21}$$

$$M_t = \tanh(W_g[g_t; h_t]). \tag{22}$$

3.5. CRF Labeling Layer

From the context features extracted from the BiLSTM output adaptation layer that incorporates the attention mechanism, the CRF layer can further consider the dependencies between sequence labels. The emission matrix P_{x_t,y_t} is the x_t output probability given for the BiLSTM output adaptation layer. The transition matrix T_{y_{t-1},y_t} represents the probability of transferring the output y_{t-1} to the output y_t from time $t-1$ to time t. All preset labels are normalized to obtain the probability of the current output label. Given the input x_t, the score $s(x_t, y_t)$ of the output y_t is obtained. The probability of the current output label is obtained by normalization, as shown in equations (23) and (24).

$$s(x_t, y_t) = \sum_{i=1}^{n} (T_{y_{t-1},y_t} + P_{x_t,y_t}). \tag{23}$$

$$P(y_t|x_t) = \frac{exp(s(x_t, y_t))}{\sum_{y_t \in y} exp(s(x_t, y_t))}. \tag{24}$$

In the training process, Viterbi [34] algorithm is used for maximum likelihood estimation to maximize the probability of the model predicting the calibration label for the input text, and the expression of Loss is obtained, as shown in equation (25).

$$Loss = \underset{\tilde{y}}{argmax}\, s(x, \tilde{y}). \tag{25}$$

3.6. Training Process

In the process of transfer learning training, when pre-training the language model source domain BERT, it is inevitable to encounter some unknown words, which make the model have to generate new vocabularies and train them. In order to make the best use of the rich semantic features in the pre-trained word vector model, this paper only trains the high frequency unknown words and integrates them with the source language model. After initialization, freeze BERT language model parameters to avoid overfitting.

The LSTM model parameter θ_S trained by the source domain data set D_S is taken as the LSTM parameter θ_T of the target domain model. Because of the difference between

the feature semantic space of source domain and target domain, the radical training strategy will cause the whole model to lose the feature information of source domain language which is needed for the optimization of target domain. However, if the training source domain neural network is updated too cautiously, the iterative efficiency of the whole model will decrease, and overfitting will easily occur. In the transfer learning of this paper, the hyperparameter ξ is set as the ratio of the learning rate of the LSTM layer and the adaptive neural network layer that integrates the attention mechanism. The specific definitions are shown in equations (26) and (27).

$$lr_S = \sigma \times lr_T. \tag{26}$$

$$\sigma = \frac{Count_T}{Count_S} \times \xi. \tag{27}$$

Where $Count_T$ and $Count_T$ are the number of tokens in the source domain and target domain data set respectively, and ξ is the hyperparameter. In this paper, the new adaptive neural network layer is inserted into the last layer. Because this layer generally contains the least amount of knowledge, the source domain model dominates transfer learning.

4. Experiment and Analysis

To verify the validity of the proposed method, experiments are conducted on five English datasets and a specialized cross-domain dataset. The experimental analysis is carried out from four aspects, namely ablation experiment, saliency detection, and fine-grained analysis.

4.1. Experimental Data

The five English datasets are CoNLL-2003 (Conll03), Twitter (T), Broad Twitter(BT), BioNLP13PC (PC), and BioNLP13CG (CG). CoNLL-2003, Twitter, and Broad Twitter datasets are similar domains with roughly similar entity types, including person (PER), location (LOC), and organization (ORG). CoNLL-2003 contains more miscellaneous (MISC) entities than Twitter. The BioNLP13PC dataset and BioNLP13CG dataset belong to the medical and biological fields, and the entity types mainly include simple chemical (CHEM), cellular component (CC), gene and gene product (GGP). BioNLP13CG also includes species (SPE) and cell. The specific data set statistics are shown in Table 1. The cross-domain dataset [35] is a specialized cross-domain NER dataset CrossNER, which contains five domains, namely politics, naturalscience, music, literature, and artificial intelligence intelligence (AI). Each domain contains a specific entity type, and the specific data set statistics are shown in Table 2.

According to the different types of entities in the data set and the differences in related fields, the experiment can be divided into two groups. Group 1: The same experimental group is selected from 5 English data sets. For example, when Twitter and Broad Twitter are used as target domain data sets, CoNLL2003 is used as source domain data. When BioNLP13PC is used as the target domain data set, CoNLL-2003 is selected as the source domain data set, and experiments are conducted between different domains and similar

Table 1. English data set statistics

Data set	Type	Training number	Verification number	Testing number
Conll03	Sentence	15100	3500	3700
Conll03	Entity	23500	5900	5700
Conll03	Word	219600	55000	50300
T	Sentence	4300	1300	1500
T	Entity	7500	2500	2500
T	Word	68700	22900	23100
BT	Sentence	6300	1000	2000
BT	Entity	8800	1700	4300
BT	Word	106300	16000	37400
PC	Sentence	2500	900	1700
PC	Entity	7900	2700	5300
PC	Word	71700	24500	47900
CG	Sentence	3000	1000	1900
CG	Entity	10800	3600	6900
CG	Word	86500	28600	54700

Table 2. CrossNER data set statistics

Data set	Type	Training number	Verification number	Testing number
politics	Sentence	200	500	600
politics	Entity	1300	3400	4200
natural science	Sentence	200	400	500
natural science	Entity	1000	2500	3500
music	Sentence	100	300	400
music	Entity	600	2600	3300
literature	Sentence	100	400	400
literature	Entity	500	2100	2200
AI	Sentence	100	300	400
AI	Entity	500	1500	1800

domains to verify the effect of MSKECDNER's migration between different domains. Group 2: Five different domain-specific data are selected from the same experimental group as CrossNER as the target domain data set, and CONLL-2003 is selected as the source domain data.

4.2. Experimental Setup

For the five publicly available English data sets, it initializes the parameters of the model with reference NCRF++ [36]. As the data sets of source domain and target domain change in different experiments, the parameters of the model also change. For example, when the source domain is BioNLP13PC and the target domain is BioNLP13CG, the optimizer is SGD algorithm, the learning rate is set to 0.005, the learning rate decay is set to 0.01, the batch size is set to 10, the hidden state dimension is 250, and the hidden variable dimension is 200, and the dropout is set to 0.5 to prevent overfitting. When Twitter and Broad Twitter are used as the target domain data set, Glove 100-dim [37] is used for initialization to obtain the feature representation of word vector. When BioNLP13PC and BioNLP13CG are the target fields, PubMed 200-dim is used to initialize the word vector. The char vector adopts the form of random initialization, it extracts the character feature representation through the convolutional neural network, and finally splices the obtained word feature representation and character feature representation to get the final feature representation.

Based on the CrossNER data set, the parameters on five English data sets are initially set. After parameter adjustment, it is found that when the same parameters are used in most fields, the performance of MSKE-CDNER is better, which proves that the model has strong robustness. At the same time, the model performance is further improved when the model parameters in some domains are fine-tuned with changes in the target domain data set. For example, in the Music domain, the optimizer uses the SGD algorithm, the learning rate is set to 0.003, the learning rate decay is set to 0.03, the batch size is 32, the hidden state dimension is 250, and the hidden variable dimension is 200, and the dropout is set to 0.5 to prevent overfitting. In the experiments, Glove100-dim is used for initialization to obtain the feature representation of word vector, and BERT optimizes the feature representation of word vector obtained. The character vector is randomly initialized, the character feature representation is extracted by convolutional neural network, and the obtained word feature representation and character feature representation are spliced as the final feature representation.

In particular, during the experimental training of five English data sets, the end of a batch is marked when the end signal of the target domain is obtained. The read operation of the source domain data is not reset at the end of the batch; it continues to load the data until it reaches the end symbol of the source domain data. When performing experiments on the CrossNER dataset, the end of a batch is replaced by an end symbol that reads into both the source domain and the target domain data.

4.3. Evaluation Index

This paper adopts the evaluation index consistent with reference [38], which holds that the prediction is accurate only when the entity types and boundaries are correctly identified.

Precision (P), Recall (R) and F1 values are used to calculate the final score, which is calculated as follows:

$$P = \frac{TP}{TP + FP}. \tag{28}$$

$$R = \frac{TP}{TP + FN}. \tag{29}$$

$$F1 = \frac{2PR}{P + R}. \tag{30}$$

TP indicates the number of correctly identified entities, FP indicates the number of incorrectly identified entities, and FN indicates the number of unidentified entities.

4.4. Comparison Model

In order to verify the effect of the proposed method on cross-domain NER, comparative experiments with relevant models are conducted on different data sets.

1. BILSTM-CRF. BILSTM-CRF [39] combines two-way LSTM and conditional random fields for named entity recognition, combines source domain data with target domain data, and trains the model together.
2. Coach. Liu et al. [40] proposed a framework Coach with NER domain adaptation, which divided tasks into two stages, first detecting entities and then classifying entities to solve the problem of data scarcity in specific domains.
3. MULTI-TASK+PGN. Jia et al. [41] integrated language model tasks in the source domain and target domain to perform cross-domain knowledge transfer, thus solving the problem that the model could not be trained in an unsupervised environment.
4. MULTI-TASK+GRAD. Zhou et al. [42] proposed a new transmission method to integrate features under high and low resources through adversarial transmission networks, and introduced generalized resource adversarial discriminator to improve the generalization ability of the model.
5. MULTI-CELL+LSTM. Jia et al. [43] proposed a structure based on Bert representation, which was modeled for different entity types respectively and carried out cross-domain knowledge transfer at the entity level to solve the problem of different meanings of entities in different fields.

4.5. Experimental Result

The proposed method is compared with other relevant methods in five English data sets and CrossNER data sets, and the results are shown in Table 3 and Table 4. Overall, the proposed method achieves good results in different data sets.

As shown in Table 3, compared with BILSTM-CRF, F1 value with Coach method on Conll03 data, T, CG increases by 2.37%, 0.86%, 1.82% respectively. It shows that multi-task architecture can improve entity recognition capability in CD-NER. Compared with Coach, the F1 value of MULTI-TASK+PGN, MULTI-TASK+GRAD and MULTI-CELL+LSTM on CG increases by 0.11%, 0.57% and 0.95% respectively. It is suggested

Table 3. F1 value on English dataset/%

Method	Conll03	T	BT	PC	CG
BILSTM-CRF	77.29	73.09	82.14	76.47	79.35
Coach	79.66	73.95	83.20	77.84	81.17
MULTI-TASK+PGN	80.18	73.81	85.65	79.97	81.28
MULTI-TASK+GRAD	79.83	74.23	85.91	80.23	81.74
MULTI-CELL+LSTM	80.97	74.94	86.37	80.85	82.12
Proposed	81.09	74.56	86.72	82.01	83.04

that fully learning and use of domain invariant knowledge between features can alleviate the problem of poor model effect caused by lack of data resources. Compared with the current popular model MULTI-CELL+LSTM, the F1 value on CG is increased by 0.92%, indicating that structured semantic knowledge can promote cross-domain knowledge transfer and alleviate the problem of entity ambiguity. Among them, the Broad Twitter domain model is not effective, considering that Broad Twitter belongs to the news domain, Conl103 also belongs to the news domain, and there is little difference between fields. When cross-domain alignment is carried out, the constraint between graph matching and domain migration is poor, resulting in poor migration effect. In the BT group of experiments, BioNLP13PC belongs to the medical field, and the data between Conll03 and BioNLP13PC are quite different. Graph matching is a good constraint for cross-domain migration. The experimental comparison shows that the greater difference between domains shows the better model transfer effect, which also shows that structured information in semantic features can promote cross-domain knowledge transfer. The greater domain difference denotes the stronger constraining effect of structured knowledge and the better transfer effect. However, the existing research methods lack the mining and utilization of this kind of information. The multi-level structure transfer method in this paper can use structured information to enhance the cross-domain transfer ability of the model.

Table 4. F1 value on CrossNER dataset/%

Method	Politics	Science	Music	Literature	AI	Average
BILSTM-CRF	56.71	50.08	44.90	43.14	43.67	47.70
Coach	61.61	52.20	51.77	48.46	45.26	51.86
MULTI-TASK+PGN	68.55	64.42	63.67	59.70	53.81	62.03
MULTI-CELL+LSTM	70.67	66.53	70.63	67.07	58.39	66.66
Proposed	71.36	67.13	73.18	67.98	61.00	68.13

The results are shown in Table 4. Compared with BILSTM-CRF, the F1 value of MULTI-CELL+LSTM in five different fields has been improved, and the average F1 value has increased by 18.96%. Since BILSTM-CRF is a single-task model, it cannot make good use of the cross-domain invariant knowledge in the source domain, while MULTI-CELL+LSTM builds a network based on the multi-task architecture and can make full use of the cross-domain invariant knowledge in the source domain. Therefore, the multi-task learning paradigm is adopted as the basic framework when constructing the method in

this paper. Compared with MULTI-CELL+LSTM, the F1 value of the proposed method is significantly improved in five different fields. Among them, F1 value has increased by 0.69% in Politics, 0.60% in Science, 2.55% in Music, 0.91% in Literature, 2.61% in AI, and 1.47% in F1 average value. Since MULTI-CELL+LSTM only considers the feature information at the entity level, ignoring the structured knowledge in the feature information, the proposed method can effectively utilize the structured knowledge in the feature information, thus achieving better performance.

4.6. Ablation Experiment

In order to verify the effectiveness of the multi-level structured alignment mechanism, ablation experiments are conducted in the CG group, and the comparative results obtained are shown in Table 5. It can be seen that the three modules in the mechanism are helpful to the improvement of entity recognition performance. X represents the comparison difference of F1 values after different ablation methods.

Table 5. Ablation experiments on CG data set/%

Method	P	R	F1	X
Proposed	83.91	82.19	83.04	none
$\delta 1$	83.85	81.26	82.67	0.37
$\delta 2$	83.90	81.79	82.83	0.21
$\delta 3$	83.36	81.98	82.53	0.51

In Table 5, $\delta 1$ represents the result obtained when the boundary detection module is eliminated, P decreases by 0.06%, R decreases by 0.93%, and F1 value decreases by 0.37%. Among them, R is decreased the most in the three ablation experiments, indicating that learning common boundary information can help the model correctly identify entity types and improve the performance of the model. Adding this module can improve the effect of cross-domain named entity recognition. Similarly, $\delta 2$ represents the experimental results of the elimination of latent layer alignment module, P decreases by 0.01%, R decreased by 0.4%, and F1 value decreases by 0.21%, indicating that the common knowledge in the migration source domain and the target domain can promote the cross-domain migration of entity recognition ability. $\delta 3$ represents the result of removing the structured alignment module, P decreases by 0.55%, R decreases by 0.21%, and F1 value decreases by 0.51%. F1 value is decreased the most in the three ablation experiments, because the structured alignment module acquires structured knowledge while obtaining feature representation. During the migration, structured information can significantly improve the entity recognition performance of the target domain due to its cross-domain stability.

4.7. Significant Difference Test

The significance test is conducted in the CG experiment, and the results are shown in Table 6. The P value in the table is $Prob > F$, and when $P < 0.05$, it indicates that there is a significant difference between PC and CG. In Table 6, $P = 0.0278$ indicates

that there are significant differences between the methods in this paper. F is the statistic of the test; P is the P value used for testing.

Table 6. Analysis of variance on $PC \rightarrow CG$ dataset

Source	Quadratic sum	Degree of freedom	Mean square error	F	P
Group	1.0087	1	1.0087	11.45	0.0278
Error	0.35254	4	0.08814		
Total	1.36115	5			

4.8. Fine-grained Analysis

Table 7 shows the fine-grained experimental results of the proposed method in CG experiments. The P, R and F1 values of the relevant entities are recorded in Table 7. Because there are many types of entities in the relevant data set, for the sake of discussion, the entity types are randomly selected for example. Compared with the current cross-domain models, F1 values of all models are improved, and the overall F1 value is increased by nearly 0.92%, which confirms the effectiveness of the proposed method.

Table 7. Fine-grained analysis on CG data set/%

Entity Type	P	R	F1
Cell component	82.69	81.78	82.23
Multi-organization structure	77.29	76.02	76.65
Biology	87.45	80.03	83.58
Simple chemistry	81.16	74.25	77.55
Organization	65.31	80.54	72.13

In order to clearly compare the method presented in this paper to obtain better results at the entity type level, it is compared with the original model for fine-grained analysis in different types of entities, and the results are shown in Table 8. Under the same entity type, the results of the proposed method are significantly better than the other two methods, which is due to the fact that the structured knowledge inside the entity effectively alleviates the problem of entity ambiguity in different domains when carrying out cross-domain migration.

5. Conclusion

This paper proposes an English text named entity recognition model based on character-level feature adaptive network. The model uses CNN to extract local features of word character sequences, and BiLSTM to extract sequential features of word sequences. In the process of model training, different weights are dynamically assigned to the two

Table 8. Comparison of fine-grained F1 values between different models/%

Entity Type	Baseline	MULTI-CELL+LSTM	Proposed
Cell component	78.81	79.67	82.23
Multi-organization structure	75.45	75.15	76.65
Biology	81.21	81.52	83.58
Simple chemistry	76.23	73.57	77.55
Organization	70.14	70.45	72.13

character-level features, which further strengthens important character-level features, weakens unimportant character-level features, and makes full use of information in character granularity. After the combined character-level features are obtained, the part-of-speech information and block analysis features are used as additional features to assist the model to judge the entity boundary, which further improves the performance of the model. Experiments are carried out on five English datasets and CrossNER datasets, and compared with current cross-domain methods, the results show that the proposed method achieves better results in cross-domain tasks, indicating that learning and using structured knowledge can better promote cross-domain knowledge transfer. In future work, we will better decouple domain invariant knowledge and domain specific knowledge to obtain better feature representation.

Acknowledgments. This work was supported by Henan Provincial Education Reform Project "Reform and Practice of Practical Teaching System for Applied Translation Undergraduate Majors from the Perspective of Technology Hard Trend" in 2024 (Project number: 2024SJGLX0581) and Zhengzhou University of Science and Technology Key Education Reform project "Innovation Research on Practical Teaching of Translation with Digital-Intelligence Technology Enabling Production and Education" in 2024 (Project number: 2024JGZD11) of phased research results.

References

1. Ehrmann M, Hamdi A, Pontes E L, et al. Named entity recognition and classification in historical documents: A survey[J]. ACM Computing Surveys, 2023, 56(2): 1-47.
2. Liu P, Guo Y, Wang F, et al. Chinese named entity recognition: The state of the art[J]. Neurocomputing, 2022, 473: 37-53.
3. Zhao Y, Li H, Yin S. A multi-channel character relationship classification model based on attention mechanism[J]. Int. J. Math. Sci. Comput.(IJMSC), 2022, 8: 28-36.
4. Ravikumar J, Kumar P R. Machine learning model for clinical named entity recognition[J]. International Journal of Electrical and Computer Engineering, 2021, 11(2): 1689-1677.
5. Khan W, Daud A, Shahzad K, et al. Named entity recognition using conditional random fields[J]. Applied Sciences, 2022, 12(13): 6391.
6. Zhang R, Zhao P, Guo W, et al. Medical named entity recognition based on dilated convolutional neural network[J]. Cognitive Robotics, 2022, 2: 13-20.
7. Su S, Qu J, Cao Y, et al. Adversarial training lattice lstm for named entity recognition of rail fault texts[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 21201-21215.
8. Carbonell M, Riba P, Villegas M, et al. Named entity recognition and relation extraction with graph neural networks in semi structured documents[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 9622-9627.

9. Uzair M, Mian A. Blind domain adaptation with augmented extreme learning machine features[J]. IEEE transactions on cybernetics, 2016, 47(3): 651-660.
10. Liang Y, Lv H, Li Y, et al. Tibetan-BERT-wwm: A Tibetan Pretrained Model With Whole Word Masking for Text Classification[J]. IEEE Transactions on Computational Social Systems, 2024.
11. Chen X, Cong P, Lv S. A long-text classification method of Chinese news based on BERT and CNN[J]. IEEE Access, 2022, 10: 34046-34057.
12. Teng L, Qiao Y. BiSeNet-oriented context attention model for image semantic segmentation[J]. Computer Science and Information Systems, 2022, 19(3): 1409-1426.
13. Pokharel S, Sidorov E, Caragea D, et al. NLP-based encoding techniques for prediction of post-translational modification sites and protein functions[M]//Machine Learning in Bioinformatics of Protein Sequences: Algorithms, Databases and Resources for Modern Protein Bioinformatics. 2023: 81-127.
14. van Erkelens A M, Thompson N A, Chalmers D. The dynamic construction of an incubation context: a practice theory perspective[J]. Small Business Economics, 2024, 62(2): 583-605.
15. Lambooy J. The transmission of knowledge, emerging networks, and the role of universities: an evolutionary approach[J]. European Planning Studies, 2004, 12(5): 643-657.
16. Juarrero A. Dynamics in action: Intentional behavior as a complex system[J]. Emergence, 2000, 2(2): 24-57.
17. Wang Y, Sun Y, Ma Z, et al. An ERNIE-based joint model for Chinese named entity recognition[J]. Applied Sciences, 2020, 10(16): 5711.
18. Peng D L, Wang Y R, Liu C, et al. TL-NER: A transfer learning model for Chinese named entity recognition[J]. Information Systems Frontiers, 2020, 22(6): 1291-1304.
19. Xi Q, Ren Y, Yao S, et al. Chinese named entity recognition: applications and challenges[J]. MDATA: A New Knowledge Representation Model: Theory, Methods and Applications, 2021: 51-81.
20. Chen Y, Wu L, Zheng Q, et al. A boundary regression model for nested named entity recognition[J]. Cognitive Computation, 2023, 15(2): 534-551.
21. Vashishth S, Newman-Griffis D, Joshi R, et al. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets[J]. Journal of biomedical informatics, 2021, 121: 103880.
22. Zhang D, Wei S, Li S, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(16): 14347-14355.
23. Yin S, Li H, Laghari A A, et al. An Anomaly Detection Model Based On Deep Auto-Encoder and Capsule Graph Convolution via Sparrow Search Algorithm in 6G Internet-of-Everything[J]. IEEE Internet of Things Journal, 2024.
24. Yin S. Object Detection Based on Deep Learning: A Brief Review[J]. IJLAI Transactions on Science and Engineering, 2023, 1(02): 1-6.
25. Yu C, Wang S, Guo J. Learning Chinese word segmentation based on bidirectional GRU-CRF and CNN network model[J]. International Journal of Technology and Human Interaction (IJTHI), 2019, 15(3): 47-62.
26. Setiawan E I, Ferry F, Santoso J, et al. Bidirectional GRU for Targeted Aspect-Based Sentiment Analysis Based on Character-Enhanced Token-Embedding and Multi-Level Attention[J]. International Journal of Intelligent Engineering & Systems, 2020, 13(5).
27. Wang F, Chen W, Yang Z, et al. Hybrid attention for Chinese character-level neural machine translation[J]. Neurocomputing, 2019, 358: 44-52.
28. Lee J, Cho K, Hofmann T. Fully character-level neural machine translation without explicit segmentation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 365-378.
29. Tran P, Dinh D, Nguyen H T. A character level based and word level based approach for Chinese-Vietnamese machine translation[J]. Computational intelligence and Neuroscience, 2016, 2016.

30. Deng J, Cheng L, Wang Z. Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification[J]. Computer Speech & Language, 2021, 68: 101182.
31. Xu C, Wang F, Han J, et al. Exploiting multiple embeddings for chinese named entity recognition[C]//Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 2269-2272.
32. Leelaluk S, Minematsu T, Taniguchi Y, et al. Scaled-Dot Product Attention for Early Detection of At-risk Students[C]//2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE). IEEE, 2022: 316-322.
33. Lin J C W, Shao Y, Djenouri Y, et al. ASRNN: A recurrent neural network with an attention model for sequence labeling[J]. Knowledge-Based Systems, 2021, 212: 106548.
34. Di Y, Li R, Tian H, et al. A maneuvering target tracking based on fastIMM-extended Viterbi algorithm[J]. Neural Computing and Applications, 2023: 1-10.
35. Chen Y, Zhong H, He X, et al. Real20M: A Large-scale E-commerce Dataset for Cross-domain Retrieval[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 4939-4948.
36. Thu Y K, Aung T, Supnithi T. Neural Sequence Labeling Based Sentence Segmentation for Myanmar Language[C]//Conference on Information Technology and its Applications. Cham: Springer Nature Switzerland, 2023: 285-296.
37. Sarang P. ANN-Based Applications: Text and Image Dataset Processing for ANN Applications[M]//Thinking Data Science: A Data Science Practitioner‘s Guide. Cham: Springer International Publishing, 2023: 289-327.
38. Yin S, Li H, Teng L, et al. Attribute-based multiparty searchable encryption model for privacy protection of text data[J]. Multimedia Tools and Applications, 2023: 1-22.
39. Knight K, Nenkova A, Rambow O. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.
40. Liu J, Yu M, Chen Y, et al. Cross-domain slot filling as machine reading comprehension: A new perspective[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 673-685.
41. Rasmussen N F, Jensen K N, Placenti M, et al. Cross-Domain Sentiment Classification using Vector Embedded Domain Representations[C]//Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing. 2019: 48-57.
42. Zhou J T, Zhang H, Jin D, et al. Dual adversarial neural transfer for low-resource named entity recognition[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3461-3471.
43. Jia C, Zhang Y. Multi-cell compositional LSTM for NER domain adaptation[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 5906-5917.

Liuxin Gao is with the School of Foreign Languages, Zhengzhou University of Science and Technology. Several papers had been published related to the work in this paper. Research direction is English text analysis, English data analysis..

Received: April 18, 2024; Accepted: June 29, 2024.