Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.050229

ARTICLE

A U-Shaped Network-Based Grid Tagging Model for Chinese Named Entity Recognition

Yan Xiang[1,2], Xuedong Zhao[1,2], Junjun Guo[1,2,*], Zhiliang Shi[3], Enbang Chen[3] and Xiaobo Zhang[3]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650504, China

[2]Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, 650500, China

[3]Kunming Enersun Technology Co., Ltd., Kunming, 650217, China

*Corresponding Author: Junjun Guo. Email: guojjgb@163.com

Received: 31 January 2024 Accepted: 01 April 2024 Published: 20 June 2024

ABSTRACT

Chinese named entity recognition (CNER) has received widespread attention as an important task of Chinese information extraction. Most previous research has focused on individually studying flat CNER, overlapped CNER, or discontinuous CNER. However, a unified CNER is often needed in real-world scenarios. Recent studies have shown that grid tagging-based methods based on character-pair relationship classification hold great potential for achieving unified NER. Nevertheless, how to enrich Chinese character-pair grid representations and capture deeper dependencies between character pairs to improve entity recognition performance remains an unresolved challenge. In this study, we enhance the character-pair grid representation by incorporating both local and global information. Significantly, we introduce a new approach by considering the character-pair grid representation matrix as a specialized image, converting the classification of character-pair relationships into a pixel-level semantic segmentation task. We devise a U-shaped network to extract multi-scale and deeper semantic information from the grid image, allowing for a more comprehensive understanding of associative features between character pairs. This approach leads to improved accuracy in predicting their relationships, ultimately enhancing entity recognition performance. We conducted experiments on two public CNER datasets in the biomedical domain, namely CMeEE-V2 and Diakg. The results demonstrate the effectiveness of our approach, which achieves F1-score improvements of 7.29 percentage points and 1.64 percentage points compared to the current state-of-the-art (SOTA) models, respectively.

KEYWORDS

Chinese named entity recognition; character-pair relation classification; grid tagging; U-shaped segmentation network

1 Introduction

Chinese named entity recognition (CNER) aims at locate entity mentions from unstructured Chinese natural language text and classify them into predefined entity categories, which is the foundation of many downstream tasks such as knowledge graph construction, entity linking, and question answering system [1–5]. As a fundamental technology, it has widespread application scenarios

across general-purpose technology [6–11], cybersecurity [12], industrial production [13], clinical records (biomedical) [14,15], and many other domains. However, CNER faces unique challenges not present in English named entity recognition (NER). Chinese word lacks explicit boundaries, and an entity may consist of multiple Chinese characters with no clear boundaries [3–5], or have complex structures and combinations [6,7]. In real-world applications, especially in specialized domains, the task of entity recognition becomes more challenging due to the technical nature of the text and the complexity of semantics. As shown in Fig. 1, with Chinese medical texts, a single sentence densely includes both flat and overlapped entities, requiring models to uniformly extract entities of different structures. Most previous works focused on flat CNER and were unable to recognize entities with other structures [4–9], leading to a significant amount of useful entity information not being extracted, which affects the performance of downstream tasks. Therefore, researching a unified CNER has significant implications.

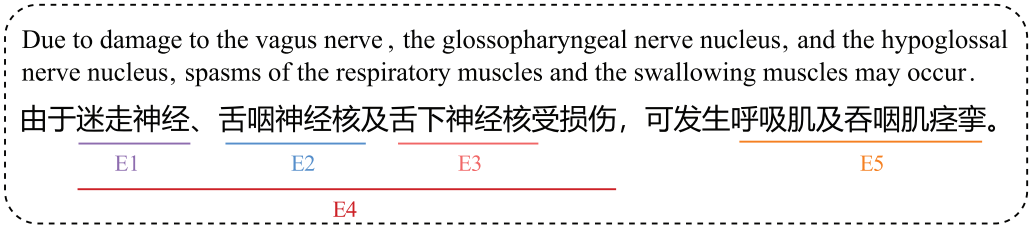

Figure 1: Example of complex entity structures in Chinese medical text (from CMeEE-V2 dataset)

Early studies have demonstrated that Chinese Word Segmentation (CWS) based CNER methods may propagate the word segmentation errors to the subsequent named entity recognition. Therefore, the current mainstream CNER is character-based methods [4–6]. Consistent with NER, CNER can be divided into three subtasks: Flat CNER, overlapped CNER and discontinuous CNER [9,16]. Previous research has mainly focused on flat CNER, which is usually regarded as a sequence labeling problem. Neural sequence tagging models represented by the Bi-LSTM-CRF classic architecture have become a general solution for flat CNER [17,18]. Later, researchers attained higher performance by integrating more Chinese character prototype features. Most representative among these is the technique of incorporating lexical information to enhance CNER [4–6,19], since Chinese words convey more independent and abundant semantic content closely associated with entity boundaries. However, due to the inherent complexity of Chinese and the multi-granularity of semantics, overlapping and discontinuous entities are also commonly present [6,18,19]. Therefore, researchers have focused attention on overlapped NER, and span-based methods have gained most prominence [18–22]. Yu et al. [21] explored all possible entity spans through a biaffine attention mechanism to achieve more accurate named entity recognition. Yuan et al. [18] and Zhu et al. [22] improved the performance of span-based NER methods by enhancing span feature representations. Su et al. [10] proposed a novel span-based CNER method called Global Pointer, which realizes the unified recognition of nested and non-nested entities. The latest research trend focuses on the study of unified NER framework, with the most impactful being the W2NER [9] method based on grid tagging. W2NER transforms the NER task into a word-pair relation classification problem, modeling the relationships between word pairs based on the Grid Tagging Scheme (GTS). This approach achieves a unified extraction of flat entities, nested entities, and discontinuous entities, attaining the latest SOTA results on multiple NER benchmark datasets.

The general idea of the grid tagging-based method is: The input sequence is converted into a character-pair grid vector matrix, and the character-pair relation tag matrix is obtained by feature

learning of the character-pair grid vector matrix [4,9,16]. Each element in this relation tag matrix represents a predefined relation category, and the entities and their categories in the sentence can be obtained by decoding the relation tag matrix. As shown in Fig. 2 in the relation tag matrix, "NONE" indicates that there is no predefined relation between character pairs; "Next-Neighboring-Character (NNC)" indicates that two characters are adjacent to each other in an entity; "Tail-Head-Character-∗ (THC-∗)" represents the tail and head boundaries of the "∗" entity. By decoding the relation matrix, the three entities of "唾液 (Saliva)", "汗 (Sweat)", "唾液分泌和出汗增加 (Increased salivary secretion and perspiration)" and their corresponding categories can be obtained.

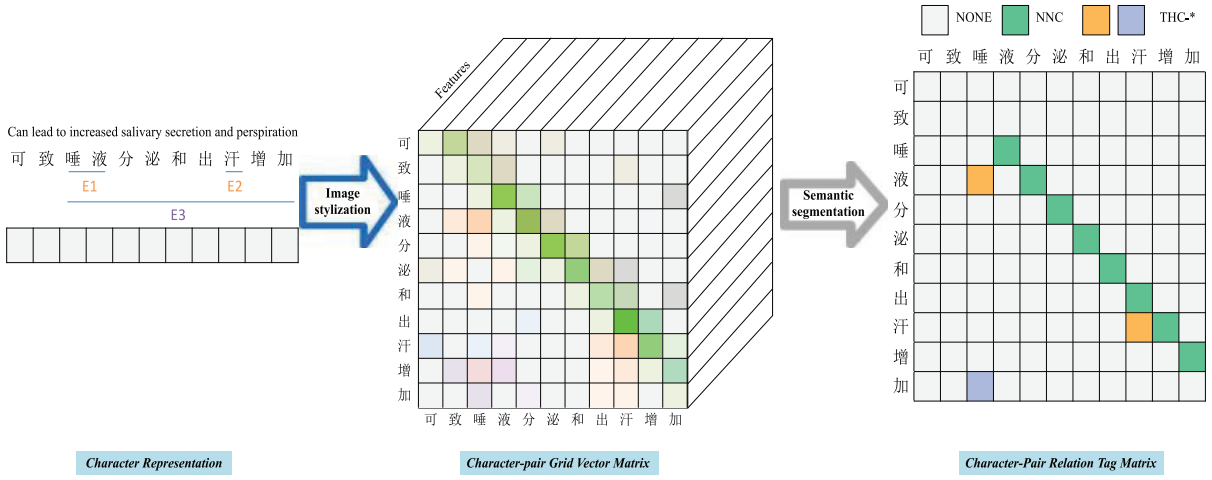

Figure 2: Overall idea of the grid-tagging-based entity recognition method

Although the grid tagging-based method has achieved great success on multiple general benchmark datasets, it still has deficiencies for CNER. First, existing methods do not fully consider the semantic information of Chinese characters when constructing character-pair grid representations. To address this problem, we enrich the character-pair information by fusing local and global information between characters to obtain a more informative character-pair grid representation. In addition, in terms of refining the grid representations, drawing inspiration from Computer Vision [23,24] and other related research [25,26], we conceptualize the grid vector matrix as an image, its grids as pixels, and the relation tag matrix as analogous to the pixel-level mask used in semantic segmentation. In this analogy, the process of obtaining the relation tag matrix resembles that of image semantic segmentation. To facilitate this semantic segmentation process, we have designed a U-shaped semantic segmentation module, which enables the capture of comprehensive and profound relationships between different character pairs, ultimately yielding a more accurate tag matrix.

The main contributions of this study are as follows:

(1) We propose a U-shaped network-based gird tagging model for CNER. To the best of our knowledge, it is the first attempt to effectively convert the grid tagging-based CNER task into an image semantic segmentation task. Specifically, we represent a sentence as a vector matrix of Chinese character pairs, and employ the proposed U-shaped semantic segmentation network to generate a high-quality character-pair relation tag matrix. This transformation enhances the accuracy of entity recognition.

(2) We design a grid feature extractor called the U-shaped segmentation network, characterized by its symmetric structure and skip connections, allowing for the extraction of hierarchically fused features at different scales. Additionally, we incorporate a spatial attention mechanism, enabling the U-shaped network to focus on grid regions relevant to entities. Through the U-shaped segmentation network, we extract comprehensive features of character-pair relationships.
(3) We conduct experiments on two Chinese medical named entity recognition benchmark datasets containing both flat and overlapped entities. Our method outperforms the baseline models obviously.

2 Related Work

According to the different tagging schemes, the existing mainstream methods for CNER can be broadly classified into the following categories: Sequence tagging-based methods, span-based methods and grid tagging-based methods.

Sequence tagging-based methods. Most of the past CNER tasks have mainly addressed flat entities, and most of these works viewed the CNER task as a sequence-labeling problem, where a specific label is assigned to each tag in the sequence [2–4]. End-to-end Bi-LSTM-CRF models are the most representative architecture [17]. Due to the natural differences between Chinese and English, word granularity-based sequence-tagging models are susceptible to the influence of Chinese word-splitting errors, leading to boundary errors. Therefore, most CNER are character-based approaches. However, the character-based approaches cannot utilize word information that plays a very important role in determining entity boundaries [4–6,19]. Therefore, researchers have proposed methods to enhance the performance of CNER by incorporating external dictionaries. Zhang et al. [5] proposed Lattice-LSTM, which incorporates lexical information into character-based CNER methods through the Lattice structure. Li et al. [19] proposed FLAT model, which models the information interaction between characters and lexicon while capturing long-distance dependencies, and solves the problems of fuzzy word boundaries and missing word meanings. Wu et al. [20] improved CNER performance by incorporating Chinese character structure information. Ma et al. [27] proposed a simple and effective method to introduce lexical information with a simple adjustment in the character representation layer, and this method can be easily migrated to other sequence tagging architectures. Sequence tagging-based methods are simple and universal, and have become the de facto standard solution for flat NER. However, entities with different structures generally require different representation or tagging schemes, and it is difficult to design a unified sequence tagging scheme for all NER tasks [3,4,9,16], which difficult to solve nested NER.

Span-based methods. The span-based approaches identify entity boundaries and assign entity category labels by enumerating all possible entity spans and performing span classification. Su et al. [10] devised the Global Pointer model, which utilizes the idea of global normalization, and can indiscriminately identify nested and non-nested entities. Huawei Cloud [8] proposed the RICON model, which utilizes entity-specific naming laws to enhance boundary information and further mitigates the effect of naming laws on boundary recognition through contextual information, achieving superior performance on multiple CNER datasets. Yan et al. [11] proposed the CNN-NER model to recognize nested entities, and designed a multi-head dual affine decoder to represent all possible span-corresponding features, achieving good performance. Zhu et al. [22] improved the model's recognition performance for long-span entities and nested structures through deep and span-specific representations. Span-based methods are one of the most mainstream approaches for nested NER. However, it is limited by

the enumeration property, which makes it more difficult to handle longer sequences and entities, and the complexity of the model is particularly large [3,9,16].

Grid tagging-based methods. This idea first appeared in Relation Extraction (RE), Aspect-oriented Fine-grained Opinion Extraction (AFOE). Wang et al. [28] proposed the TPLinker model, which transforms the entity-relation joint extraction task into token-pair linking, which exhibits superior performance on overlapping and multiple relation extraction tasks. Wu et al. [29] proposed a novel GTS to solve the complete AFOE tasks by means of unified tags. GTS approaches show strong performance in information extraction tasks. Recently, researchers have also proposed some GTS approaches to handle the NER task. The current most advanced methods transform the NER into word-word relation classification by modeling the relationship between entity boundary words and internal component words, achieving a unified NER framework [9]. Liu et al. [16] built on W2NER by enhancing the interrelations between tags, words to better recognize discontinuous entities. Li et al. [30] designed a word pair relation tagging scheme for Joint Multimodal Entity-Relation Extraction (JMERE) task, which can fully exploit the bidirectional interaction information between entity recognition and relation extraction, and avoids error propagation due to the pipeline framework problems. In real-world usage scenarios, it is often necessary to recognize entities of different structures simultaneously, and the Grid tagging-based methods provides a new idea for handling all NER tasks in a unified way [4,9,16]. However, for CNER, existing methods have somewhat overlooked global semantic information between Chinese characters, as well as Chinese lexical information. When dealing with domain-specific corpora, the performance has not been satisfactory.

In particular, the innovative combination of concepts and methods from Computer Vision into the field of Natural Language Processing (NLP) helps us to understand the task itself from more perspectives while improving the performance and diversity of NLP models. Liu et al. [25] investigated Incomplete Utterance Rewriting as a semantic segmentation task, and achieved SOTA performance on multiple datasets. Zhang et al. [26] considered the correlation features between entity pairs as images, and for the first time innovatively transformed the document-level relation extraction problem into a semantic segmentation problem, achieving SOTA performance on multiple document-level relation extraction benchmark datasets. These approaches also inspire us to study the NER task from the perspective of semantic segmentation in computer vision.

3 Our Model

3.1 Preliminary

The unified NER framework based on grid tagging can be formalized as follows: Given an input sentence composing of N Chinese characters, the goal is to extract the relation $r_{i,j} \in R$ between every character-pair(x_i, x_j), where $R = \{NONE, NNC, THC - *\}$ is a predefined relation set. Our goal is to obtain an $N * N$ character-pair relationship matrix, where each element in the matrix represents the relationship between the corresponding character pairs. The character-pair relationship matrix is similar to the pixel-level mask in semantic segmentation, establishing a connection between character-pair relation classification and semantic segmentation.

Our model framework, as depicted in Fig. 3, comprises four key components: 1) Character Representation: Initially, we process the input sentence to obtain character representations that encompass both contextual information and word-level information. 2) Character-Pair Grid Vector Matrix Acquisition: We sent the character representation of the sentence to Biaffine Attention and Conditional Layer Normalization (CLN), for the acquisition of the character-pair grid feature matrix. 3) Character-Pair Relation Tag Matrix Extraction: We utilize a U-shaped segmentation network

to extract both local and global features from the Character-Pair Grid Vector Matrix, leading to the generation of a high-quality character-pair relation matrix. 4) Entity Prediction: We decode the character-pair relation tag matrix to identify the potential entities and their categories.

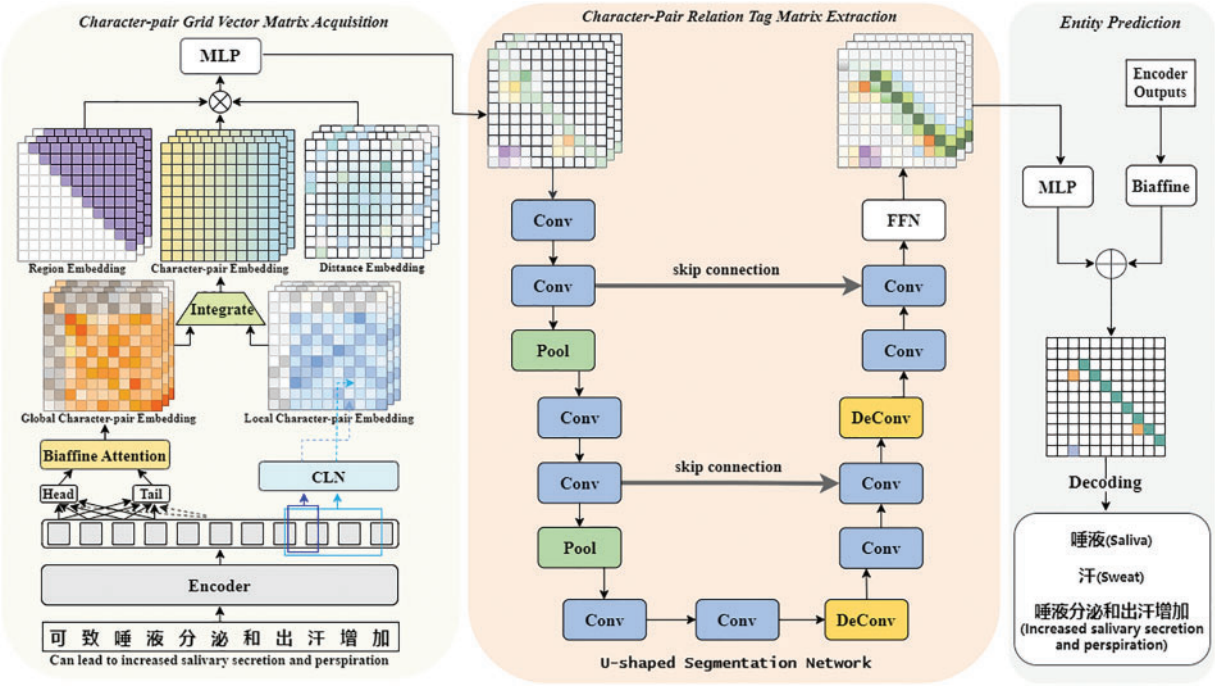

Figure 3: Overall architecture

3.2 Character Representation

An input sentence $X = \{x_1, x_2, \ldots, x_N\}$ is fed into the pre-trained language model BERT [31] to derive embedding for each Chinese character. Subsequently, the character embedding is further refined through the application of a Bi-LSTM network [32]. After this encoding process, the input sentence X can be represented as $e_c \in \mathbb{R}^{N \times d_h}$:

$$e_c = Encoder\{x_1, x_2, \ldots, x_N\} = \{e_{c1}, e_{c2}, \ldots, e_{cN}\} \tag{1}$$

where x_i represents the i-th Chinese character in the sentence, e_{ci} represents the character embedding of x_i, d_h represents the embedding dimension, and N represents the length of the sentence.

In contrast to Chinese characters, Chinese words contain independent semantic information. For this reason, we incorporate word information into character embedding based on the FLAT structure [19]. The word w_j matched by the dictionary is mapped to the word embedding $e_{w_j} \in \mathbb{R}^{d_h}$.

$$e_{w_j} = E_w(w_j) \tag{2}$$

where $E_w(\cdot)$ represents the pre-trained word embedding lookup table.

The character and word embedding are concatenated to form the composite embedding, denoted as e_x:

$$e_x = Concat(e_c, e_w) = \{e_{c1}, e_{c2}, \ldots, e_{cN}, e_{w1}, e_{w2}, \ldots, e_{wj}\} \tag{3}$$

where $e_x \in \mathbb{R}^{(N+N_w)\times d_h}$, N_w represents the number of the matched words, $Concat()$ is concatenation operation.

Subsequently, we enable full interaction between character embedding and word embeddings using the same approach as in Transformer-XL [33] and FLAT [19], which leverages relative position encoding for spans, combines the outcomes of the attention heads within the Transformer encoder, and feeds them into the feed-forward neural network layer [34]. This process yields the final hybrid character representation, denoted as: $H = \{h_1, h_2, \ldots, h_N\} \in \mathbb{R}^{N\times d_h}$.

3.3 Character-Pair Grid Vector Matrix Acquisition

In our study, to capture the dependency relationships between character pairs, a crucial step is obtaining a high-quality character-pair grid vector matrix representing character-pairs.

Specifically, we first mine the local and global character-pair relation representations. we employ Conditional Layer Normalization [9,16,35–37], where the representation h_i of the character x_i and the representation h_j of character x_j serve as the condition vector and input vector, respectively, to generate the local character-pair relation vector $V^{(pair)} \in \mathbb{R}^{N\times N\times d_h}$.

$$V_{ij}^{(pair)} = CLN(h_i, h_j) = \gamma_i \odot \left(\frac{h_j - \mu}{\sigma}\right) + \lambda_i \tag{4}$$

where $\gamma_i = W_\alpha h_i + b_\alpha$ and $\lambda_i = W_\beta h_i + b_\beta$, all W and b are learnable parameters, μ and σ are the mean and standard deviation across the elements of h_j.

In addition, we use Biaffine Attention [8,21,38,39] to obtain the global character-pair relation vector $V_{ij}^{(span)} \in \mathbb{R}^{N\times N\times d_h}$ that implies the correlation of the span between x_i and x_j:

$$V_{ij}^{(span)} = Biaffine\ Attention(h_i^{(head)}, h_j^{(tail)}) \tag{5}$$

$$h_i^{(head)} = LeakyReLU\left(h_i W_s\right), h_j^{(tail)} = LeakyReLU(h_j W_e) \tag{6}$$

where $LeakyReLU()$ is the activation function, W_s and W_e are learnable parameters.

Subsequently, we fuse the local and global relation vectors through a gating mechanism [8], to obtain the final character-pair representation $V \in \mathbb{R}^{N\times N\times d_h}$:

$$g_{s_{i,j}} = \sigma(W_z[V_{ij}^{(pair)}, V_{ij}^{(span)}] + b_z) \tag{7}$$

$$V_{ij} = g_{s_{i,j}} \cdot V_{ij}^{(pair)} + (1 - g_{s_{i,j}}) \cdot V_{ij}^{(span)} \tag{8}$$

where W_z and b_z are learnable parameters, $\sigma()$ is the sigmoid function.

Finally, we obtain the relative position vector $E^d \in \mathbb{R}^{N\times N\times d_d}$ and the position vector $E^t \in \mathbb{R}^{N\times N\times d_t}$ that distinguishes the upper and lower triangular areas in the matrix, following the approach outlined in [9,16,40]. We then combine these two position vectors with the previously mentioned character-pair relation vector through the multi-layer perceptron (MLP) to obtain the ultimate character-pair grid vector matrix $G \in \mathbb{R}^{N\times N\times D}$:

$$G = MLP([V; E^d; E^t]) \tag{9}$$

3.4 Character-Pair Relation Tag Matrix Extraction

Considering the character-pair grid vector matrix G obtained in the previous module as a D-channel image, we can formalize the character-pair relation prediction as a pixel-level mask within this image. Inspired by the seminal U-Net architecture [23–26], we utilize the U-shaped segmentation

network to achieve this goal. As shown in Fig. 3, it consists of two down-sampling and up-sampling blocks, connected in the middle by skip connection, with a U-shaped overall architecture [23–26].

In detail, each down-sampling block consists of two convolutional layers and a maximum pooling layer. The number of channels in each down-sampling stage is doubled compared to the previous stage. By enlarging the receptive field of character pairs (x_i, x_j), we can capture dependency information between character pairs at different distances, extract more diverse associative features. These features, in turn, provide richer information for the subsequent up-sampling stage. Let Q_l^i represent the output of the i-th convolutional layer in the l-th down-sampling block, and M_l represents the output of the l-th down-sampling block. The input of each convolutional layer comes from the previous module or the output of the previous convolutional layer, and the calculation is as follows:

$$Q_l^i = ReLU\left(Conv_{3\times3}\left(Input_{Conv}\right)\right) \tag{10}$$

$$Q_l^{i+1} = ReLU(Conv_{3\times3}(Q_l^i)) \tag{11}$$

$$M_l = MaxPool(Q_l^{i+1}) \tag{12}$$

where $Conv_{3\times3}$ represents a convolution operation with the filter size of 3×3, $ReLU()$ represents the activation function, $MaxPool()$ represents max pooling, $Input_{Conv}$ represents the input vector of the convolution layer.

Conversely, each up-sampling block comprises a deconvolution layer followed by two convolutional layers. In these up-sampling stages, the number of channels is reduced by a factor of two compared to the previous stage. This process facilitates the distribution of aggregated information to each pixel-level character-pair. Let Z_l represent the output stemming from the deconvolutional layer within the l-th up-sampling block, and U_l^i represents the output of the i-th convolutional layer in the l-th up-sampling block. The input of the first deconvolution layer comes from the deepest layer of down-sampling, and the input of the subsequent deconvolution layer comes from the output of the previous part. The input of each convolutional layer comes from the output of the deconvolution layer or the previous convolutional layer, and the calculation is as follows:

$$Z_l = DeConv_{2\times2}(Input_{DeConv}) \tag{13}$$

$$U_l^i = ReLU(Conv_{3\times3}(Z_l)) \tag{14}$$

$$U_l^{i+1} = ReLU(Conv_{3\times3}(U_l^i)) \tag{15}$$

where $DeConv_{2\times2}$ represents the deconvolution operation with a filter size of 2×2, $Input_{DeConv}$ represents the input vector of the deconvolution layer.

Next, by employing skip connections, we integrate the up-sampling features with the down-sampling features, thereby achieving multi-scale feature fusion.

$$f_l = ReLU(Conv_{3\times3}(Concat(M_{l+1}, U_l)) \tag{16}$$

where f_l represents the feature map after skip connections.

Additionally, we employ Spatial Attention mechanisms [41,42] to selectively focus on regions and positions within the feature map. Specifically, we begin by aggregating channel information within the feature map through average pooling and max pooling operations. Subsequently, we adopt convolutions to generate spatial attention maps. Finally, we fuse these maps with the original feature map to obtain the ultimate output.

$$M_s(f_l) = \sigma(Conv_{7\times7}([AvgPool(f_l); MaxPool(f_l)])) \tag{17}$$

$$F_l = M_s(f_l) \otimes f_l \tag{18}$$

where $Conv_{7\times7}$ represents the convolution operation with a filter size of 7×7, $AvgPool()$ represents average pooling, $\otimes$ represents element-wise multiplication, and F_l represents the final feature output after using Spatial Attention.

Finally, we obtain the character-pair relation tag matrix $F \in \mathbb{R}^{N\times N\times d_u}$, where d_u represents the embedding dimension of the matrix.

3.5 Entity Prediction

Previous work has shown that co-prediction allows the network to better utilize shallow and deep features for joint inference and enhance the classification of the model [9,16,43,44]. We apply the same approach to improve the performance of NER models. After obtaining the character-pair relation tag matrix F, we use the MLP predictor and the Biaffine classifier to compute two independent relationship distributions y'_{ij} and y''_{ij} for the character pair (x_i, x_j), respectively, and combine them to enhance the final prediction of the model.

Specifically, the character-pair relation tag matrix F is sent to the MLP predictor, and the relationship prediction score for each character pair (x_i, x_j) is:

$$y'_{ij} = MLP\left(F_{ij}\right) \tag{19}$$

The character representation is sent to the Biaffine classifier [39], and the relationship score for each character pair (x_i, x_j) is computed as in the following equation:

$$y''_{ij} = s_i^{\top} U o_j + W[s_i; o_j] + b \tag{20}$$

$$s_i = MLP\left(h_i\right), o_j = MLP\left(h_j\right) \tag{21}$$

where U, W and b are learnable parameters.

Finally, the scores of the two predictors are combined as the final score:

$$y_{ij} = Softmax\left(y' + y''\right) \tag{22}$$

We decode the associated entities and their respective categories using a depth-first path search algorithm [9]. For entities comprised of a single character, they are directly recognized based on the THC-$*$ tag. In the case of entities composed of multiple characters, we create a character graph where the constituent characters serve as nodes, and the NNC tags between characters serve as edges. We apply a depth-first search algorithm to identify all paths within the graph, each path corresponding the character index sequence of an entity. We determine the entity's category based on the THC-$*$ tag between the head node and the tail node.

3.6 Training Objective

For a given input sentence $X = \{x_1, x_2, \ldots, x_N\}$, the objective function of model training is to minimize the negative log-likelihood loss between the predicted label probability $\boldsymbol{y}_{ij}$ and the gold label $\hat{\boldsymbol{y}}_{ij}$ for the character pair (x_i, x_j):

$$\mathcal{L} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{r=1}^{|R|}\hat{\boldsymbol{y}}_{ij}^{r} log \boldsymbol{y}_{ij}^{r} \tag{23}$$

where r denotes the r-th relation in the set R of predefined relations.

4 Experimental Settings

4.1 Datasets

To verify the effectiveness of our proposed method, we conducted evaluations on the CMeEE-V2 dataset and Diakg dataset. The CMeEE-V2 dataset is sourced from the Chinese Biomedical Language Understanding Evaluation (CBLUE) [14], while the Diakg dataset is a high-quality Chinese dataset for Diabetes knowledge graph [15].

Statistical details of these datasets are presented in Table 1. It is worth noting that both datasets include not only flat entities but also a certain proportion of overlapping entities. CMeEE-V2 encompasses nine entity categories, including diseases, clinical manifestations, and medical procedures, whereas Diakg comprises 18 entity categories, such as diseases, pathogenesis, and drug names.

Table 1: Detailed statistics of each dataset

	CMeEE-V2			Diakg		
	Train	Dev	Test	Train	Dev	Test
Sentence	14881	2477	2478	1899	407	408
Entity	80166	12556	12669	14865	3177	3087
Overlapped entity	12192	1781	1795	1810	377	360
Overlapped entity (%)	15.21	14.18	14.17	12.18	11.87	11.66

4.2 Parameter Setting and Evaluation Metrics

To get the word embedding lookup table $E_w(\cdot)$ in formula (2), we used the open source word segmentation toolkit pkuseg[1] to segment the collected corpus[2,3], and conducted pre-training on the segmented corpus using the skip-gram model [45]. We employed the Adam optimizer for the parameter optimization, and the primary hyperparameter settings used in our experiments are shown in Table 2. If there are multiple values, they represent the values on the CMeEE-V2 and Diakg data sets, respectively. Our model is implemented with PyTorch and trained on NVIDIA RTX 3090 GPU, and all hyperparameter values are obtained after tuning on the development set.

Table 2: Main experimental parameter settings

Parameters	Values	Parameters	Values
Bert hidden size	768	d_d	20
Bi-LSTM hidden size	256, 512	d_t	20
Biaffine hidden size	512	d_u	288
Learning rate (Bert)	5e-6	Weight decay	0.1, 0.001
Learning rate (others)	1e-3, 1e-4	Word dropout	0.01, 0.5

[1] https://github.com/lancopku/pkuseg-python.git
[2] https://github.com/GanjinZero/awesome_Chinese_medical_NLP.git
[3] https://github.com/Toyhom/Chinese-medical-dialogue-data.git

Consistent with previous research in named entity recognition, we use Precision (P), Recall (R), and F1-score (F1) as evaluation metric for our model. In line with the conventions of previous Span-based methods [8,10,11] and Grid-tagging-based methods [9,16], a predicted entity is considered to be true positive when its characters and corresponding category exactly match the true entity.

4.3 Baseline Models

To evaluate the performance of our proposed method, we compared it with several baseline methods employing different tagging schemes:

Simple-Lexicon [27]: This approach optimizes the integration of lexical information into Lattice by obtaining a collection of vocabulary for each character's corresponding BMES (Beginning, Middle, End, Single) positions. This simple method seamlessly incorporates word information into character representations, avoiding the complexity of model structures and facilitating easy adaptation to other sequence labeling frameworks.

FLAT [19]: The model introduces a Flat-Lattice Transformer structure that represents the Lattice structure using a set of spans. It employs four different positional encodings to interact with both character and word-level information. FLAT also addresses the issue of non-parallelizable operations in Lattice-LSTM, achieving state-of-the-art results on multiple CNER datasets at that time.

MECT [20]: By utilizing multivariate data embedding, MECT fuses the features of Chinese characters with sub-radical-level embedding through a cross-transformer architecture. This enhancement allows the model to better capture the semantic information of Chinese characters and further improve performance with random attention.

DSpERT [22]: A method of deep span representations that aggregates token information into span representations progressively from bottom layers to top layers. This allows effectively decoupling representations of overlapping spans and making representations of different categories more separable in the feature space, improving the model's recognition of long entities and nested entities.

CNN-Nested-NER [11]: This approach utilizes a multi-head Biaffine decoder to obtain prediction score matrices, treating them as images. It employs CNN to model the spatial relationships between adjacent spans in the score matrix, enhancing the model's ability to recognize nested entities.

Global Pointer [10]: Using a global normalization approach and a multiplicative attention mechanism, Global Pointer incorporates relative positional information. By considering the start and end positions of entities, it achieves a global view for the unified recognition of both non-nested and nested entities.

W2NER [9]: This approach employs a GTS method that transforms the NER task into word-pair relation classification. It effectively models the adjacency and head-tail relationships between word pairs, converting sentences into two-dimensional tables. W2NER achieves state-of-the-art performance on multiple Chinese and English NER benchmark datasets by using multi-granularity dilated convolutions to capture relationships between words at different distances.

TOE [16]: The enhanced version of W2NER incorporates two additional tags, focusing not only on the relationships between words but also on the relationships between words and tags. The prediction of the relationship between words and tags is strengthened through a more fine-grained tag system.

5 Result Analysis

5.1 Overall Results

We compare our proposed model with the existing CNER methods on two datasets to validate the superiority of the proposed method, as shown in Table 3.

Table 3: Experimental results of different models

Tagging scheme	Model	CMeEE-V2			Diakg		
		P	R	F1	P	R	F1
BIO/BMES	Simple-Lexicon	61.00	60.31	60.64	72.39	67.48	69.85
	Simple-Lexicon + BERT	65.60	66.80	66.19	72.84	73.40	73.11
	FLAT	61.83	66.42	64.03	75.42	82.32	78.72
	FLAT + BERT	65.28	68.53	66.86	79.13	84.25	81.61
	MECT	70.54	73.11	71.80	72.75	70.91	71.83
Span	DSpERT	73.74	68.15	70.83	83.15	83.28	83.21
	CNN-Nested-NER	72.29	73.69	72.98	78.86	74.55	76.65
	Global Pointer	73.10	72.25	72.54	84.27	83.46	83.79
GTS	W2NER	74.58	75.19	74.87	83.56	84.20	83.87
	TOE	75.92	74.37	75.14	84.69	84.21	84.45
	Ours	**82.20**	**82.11**	**82.16**	**86.40**	**84.64**	**85.51**

The following observations can be made: (1) Our model demonstrates superior performance on both datasets. When compared to the state-of-the-art model W2NER in the baseline, our proposed method exhibits a remarkable improvement of 7.29 percentage points in the F1-score on the CMeEE-V2 dataset and an enhancement of 1.64 percentage points on the Diakg dataset. These improvements can be attributed to two key innovations. Firstly, when constructing character-pair representations, we integrated both local and global information between characters to obtain a more enriched character-pair grid representation. Secondly, we utilized the U-shaped network to extract deeper dependencies between character pairs on the image-style three-dimensional character pair grid, which helps to obtain a more accurate character pair relationship matrix, thereby improving the accuracy of entity recognition. As an improved version of W2NER, TOE achieved performance gains on both datasets, mainly because TOE extended the label system by modeling multi-granular interactions between characters and labels. However, this also led to another problem: TOE itself is only suitable for single-entity-type datasets. To adapt to multi-entity-type datasets, TOE needs to bind the Head-Tail relations and entity types following the method in W2NER. With the expansion of the label system, the number of labels would increase exponentially after binding entity types, and too many labels could easily make prediction more difficult. Therefore, for multi-entity-type datasets, W2NER and our method are more suitable. (2) Overall, across various tagging schemes, the performance trends are as follows: GTS methods exhibit the best performance, followed by span-based methods, while performance is relatively lower for sequence tagging methods. Grid-tagging-based methods show superior performance as this labeling approach can simultaneously represent entity boundaries and the relationships between characters within the entity, providing rich information. This is more advantageous for handling complex CNER tasks, facilitating the achievement of unified CNER, and resulting in better model performance and generalization. (3) On the whole, nearly all methods

demonstrate superior performance on the Diakg dataset compared to the CMeEE-V2 dataset. This observation aligns with the statistical information presented in Section 4.1. The Diakg dataset has a higher average number of entities per sentence and a lower proportion of nested entities. Additionally, entities in the Diakg dataset are generally shorter in length. These factors provide more representation learning opportunities to distinguish entity features.

5.2 Ablation Study

In order to validate the impact of the proposed modules on the performance of our model, we conducted the following ablation experiments:

w/o All U-Net: remove the full U-shaped segmentation network from our model.

w/o global: remove the global character-pair relation representation from our model.

w/o local: remove the local character-pair relation representation from our model.

w/o Spatial Attention: remove the Spatial Attention from the U-shaped segmentation network.

w/o Skip Connections: remove the Skip Connections from the U-shaped segmentation network.

r/w DConv: replace our proposed U-shaped segmentation network with Multi-Granularity Dilated Convolution.

r/w Conv: replace our proposed U-shaped segmentation network with vanilla convolution network.

The experimental results, as shown in Table 4, lead to the following observations: (1) Ablation experiments on two datasets show that removing either global or local character pair information leads to varying degrees of decline in named entity recognition performance. The model achieves optimal performance when integrating both types of information, consistent with our analysis in Section 1. Indicating that local and global character pair information can promote each other synergistically and provide richer semantic information. (2) There is a significant decrease in F1-scores on both datasets when removing the U-shaped segmentation network. This indicates that this module plays a crucial role in refining character-pair grid representations. By fusing shallow and deep multi-scale features, this module enhances relationship modeling and boundary delineation between character pairs. This leads to gains in overall entity recognition. (3) Removing the Spatial Attention or Skip Connections has various impacts on the model's performance. This can be attributed to that the Spatial Attention module assists the network in selectively attending to regions more relevant to entities and Skip Connections effectively merges low-level positional information with high-level semantic information. These elements may contribute to entity boundaries and category recognition, thus affecting the model's overall performance. (4) When the U-shaped segmentation network is replaced with Multi-Granularity Dilated Convolution, there is a significant decrease in performance. We speculate that this is because Multi-Granularity Dilated Convolution might not integrate multi-scale contextual information as effectively as the U-shaped network. The U-shaped network, with its unique structure that includes both downsampling and upsampling paths as well as skip connections, obtains hierarchically fused features at different scales. This aids the model in understanding the relationships between Chinese characters, capturing more intricate detail features and associative information. In contrast, although Multi-Granularity Dilated Convolution offers varied receptive fields, it lacks the necessary hierarchical integration of features, which is crucial for comprehensive entity recognition in complex datasets. (5) When replacing the proposed U-shaped segmentation network with the vanilla convolution, there is a significant drop in the model performance, especially on the CMeEE-V2 dataset. Combining the analysis based on the statistical information of the datasets

analyzed earlier, we attribute this phenomenon primarily to the generally longer entity lengths present in the CMeEE-V2 dataset. As a result, the vanilla convolution struggles to capture long-distance dependency relationships between characters in the CMeEE-V2 dataset, thereby affecting the model's performance.

Table 4: Results of ablation experiments

Model	CMeEE-V2			Diakg		
	P	R	F1	P	R	F1
Ours	**82.20**	82.11	**82.16**	**86.40**	84.64	**85.51**
w/o global	78.18	80.81	79.47	85.28	85.06	85.17
w/o local	79.78	80.74	80.26	84.90	84.41	84.65
w/o All U-Net	81.40	80.18	80.79	84.70	85.03	84.86
w/o Spatial Attention	81.24	**82.57**	81.90	83.34	85.90	84.60
w/o Skip Connections	**80.68**	82.49	81.58	82.38	**86.03**	84.17
r/w DConv	79.70	82.70	81.17	83.76	85.74	84.74
r/w Conv	73.70	76.00	74.63	84.33	85.26	84.79

5.3 Effect of the U-Shaped Segmentation Network Depth

We further investigated the influence of the U-shaped segmentation network depth. Specifically, we measured the network depth by the skip connections number in the U-shaped segmentation network. The experimental results are shown in Fig. 4.

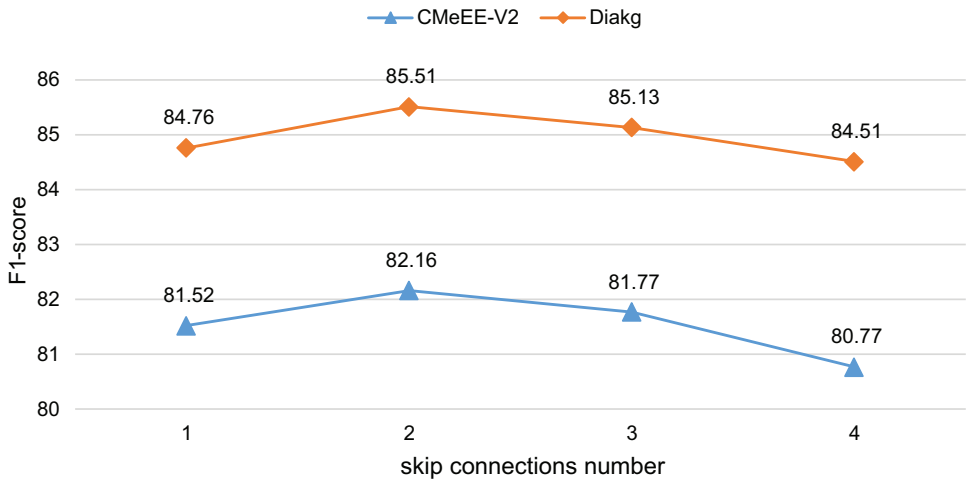

Figure 4: Experimental results for different U-shaped network depths

We can observe that the network depth also affects the model performance. When the skip connections number is 2, the model performs best. We believe the main reasons are as follows: Shallow networks have limited feature representation capabilities due to insufficient layers, making them unable to learn the complex dependency relationships between character pairs. As the number of layers increases, the network's expressive power is continuously enhanced, enabling learning increasingly rich

feature dependencies. When the network reaches a certain depth, its feature representation capabilities are sufficient to learn the dependencies between character-pair grid features. Further increasing the depth leads to marginally decreasing performance gains. Adding too many layers significantly increases the number of network parameters, reducing the model's generalization ability and easily leading to overfitting, resulting in degraded performance on the test set. There exists an optimal network depth that strikes a balance between feature representation power and overfitting to ensure best model performance.

5.4 Effect of the Character-Pair Representation Dimension

We explored the impact of the character-pair representation dimension d_h on model performance, and the experimental results are shown in Fig. 5.

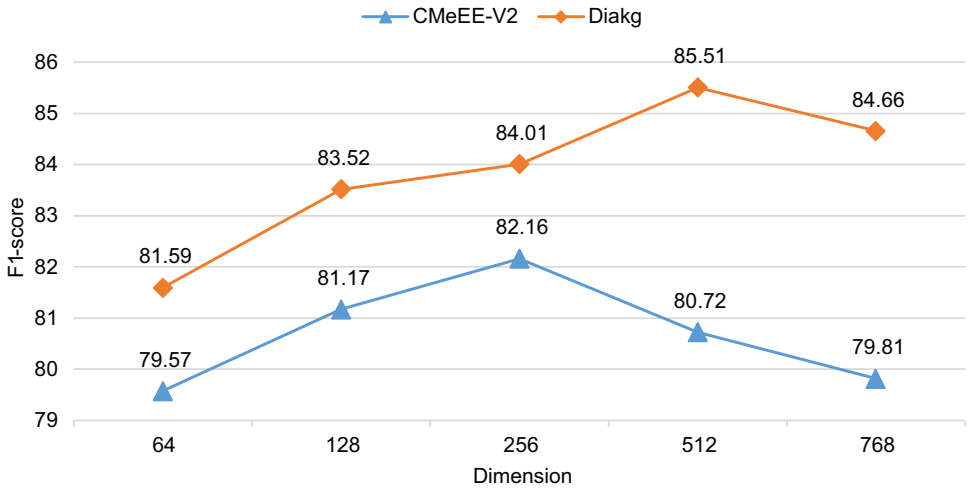

Figure 5: Experimental results of different character-pair grid feature channel numbers

We observe that the model performs best when d_h is around 256 and 512 on CMeEE-V2 and Diakg, respectively. Too low or too high dimensions can adversely affect the performance. We speculate that the embedding dimension of the character-pair determines the model's ability to express the features of the character-pair. A shallow dimension prevents the U-shaped network from fully capturing the dependency information in the character-pair grid vector matrix, setting the dimension too high will significantly increase the number of model parameters, thereby reducing its generalization ability and making it prone to overfitting. An appropriate character pair representation dimension can balance information richness and model complexity, and thus achieve optimal performance.

5.5 Case Study

We selected several samples from CMeEE-V2 and Diakg for conducting case studies on different models. The results are shown in Fig. 6.

We have the following analysis: (1) For the first case, we can observe that the Simple-Lexicon method cannot recognize the second entity in a sequence of consecutive flat entities, while the other models can successfully recognize all entities. Demonstrating that sequence tag-based methods lack the ability to identify multiple contiguous entities. (2) For the second case, the challenge for the model arises from the short length of the sentences densely containing two semantically similar but distinct

entities of the same category. It requires distinguishing very closely related and similar entities. Both sequence tagging and span-based methods struggle with this task as they typically rely on broader context information or struggle to handle high-density entity recognition efficiently. On the contrary, grid-based methods successfully identify the correct entity sequences. Our approach further accurately identifies entity categories, possibly because the grid-based character pair relationship classification method can more finely analyze the relationships between characters. This approach considers the potential relationships between each character pair, allowing the model to recognize and differentiate neighboring entities in more detailedly. Therefore, in this scenario, it successfully identifies the correct entity sequence, and our method integrates and captures richer Chinese character-pair information, which may contribute to the accurate identification of entity categories. (3) For the third case, this sentence contains both nested and flat entities, with a high density of entities, making the situation more complex. Simple-Lexicon failed to identify nested entities, mainly due to the limitations of the BIO/BMES tagging framework. The model cannot learn relevant information under this framework. The other tagging schemes accurately recognized the overlapped entities. However, Global Pointer failed to recognize subsequent flat entities. This may be because the nested entity "肾" (kidney) within the nested entity "糖尿病肾病" (diabetic nephropathy) and the flat entity "肾功能" (renal function) have overlapping semantic relationships. The Global Pointer method focuses only on the start and end positions of spans without explicitly modeling information within the span. Therefore, it performs poorly in complex scenarios with dense and semantically overlapping entities. Both W2NER and our model correctly identified entities and their categories, indicating that understanding the relationships between Chinese characters can address more complex scenarios effectively.

Sentence (Truncated)	Golden Entity {boundary, entity, type}	Prediction Model/Tagging Scheme			
		Simple-Lexicon/BIO	Global Pointer/Span	W2NER/GTS	Ours/GTS
Chinese:	{[83, 86], 乙肝疫苗, dru}	{[83, 86], 乙肝疫苗, dru}	{[83, 86], 乙肝疫苗, dru}	{[83,86],乙肝疫苗,dru}	{[83, 86], 乙肝疫苗, dru}
乙肝疫苗接种3剂次；脊灰疫苗口服4剂次，……	{[87, 88], 接种, pro}	{[93, 96], 脊灰疫苗, dru}	{[87, 88], 接种, pro}	{[87,88],接种,pro}	{[87, 88], 接种, pro}
English:	{[93, 96], 脊灰疫苗, dru}	*{[87, 88], 接种, pro}*	{[93, 96], 脊灰疫苗, dru}	{[93,96],脊灰疫苗,dru}	{[93, 96], 脊灰疫苗, dru}
3 doses of hepatitis B vaccine; 4 doses of polio vaccine orally, …	{[97, 98], 口服, pro}	*{[97, 98],口服, pro}*	{[97, 98], 口服, pro}	{[97,98],口服,pro}	{[97, 98], 口服, pro}
Chinese:					
(2)尼龙单丝：深感觉评估	{[3, 6], 尼龙单丝, Test}	*N/A*	*N/A*	{[3, 6],尼龙单丝, *ADE* }	{[3, 6],尼龙单丝,Test}
English:	{[8, 12], 深感觉评估, Test}			{[8, 12],深感觉评估, *Class* }	{[8, 12],深感觉评估,Test}
(2) Nylon monofilament deep sensory assessment					
Chinese:					
(3)糖尿病肾病的筛查和肾功能评价 ……	{[3, 7], 糖尿病肾病, Disease}	*{[3, 7],糖尿病肾病, Disease}*	{[3, 7], 糖尿病肾病, Disease}	{[3,7], 糖尿病肾病, Disease}	{[3,7], 糖尿病肾病, Disease}
	{[6, 6], 肾, Anatomy}	*{[6, 6],肾, Anatomy}*	{[6, 6], 肾, Anatomy}	{[6, 6], 肾, Anatomy}	{[6,6], 肾, Anatomy}
English:	{[12, 14], 肾功能, Test}	{[12, 14], 肾功能, Test}	*{[12,14],肾功能, Test}*	{[12, 14], 肾功能, Test}	{[12,14], 肾功能, Test}
(3) Screening of diabetic nephropathy and evaluation of renal function …					

Figure 6: Case analysis. Note: *Red italics* indicates that the model failed to recognize the corresponding entity in the sequence. *N/A* indicates that the model failed to recognize any entities in the current sequence. *Orange italics* indicates that the model failed to correctly identify the entity category

6 Conclusion

In this paper, we propose a new unified CNER method based on the grid tagging framework. The key innovation lies in our approach to handling CNER from the perspective of semantic segmentation in Computer Vision. We integrate local and global inter-character information via Conditional Layer

Normalization and Biaffine Attention to obtain semantically richer and more comprehensive grid representations. Moreover, we propose a new U-shaped network to distill deep dependencies between character pairs, which leads to a more accurate character-pair relationship matrix and improved entity recognition performance. Extensive experimental results on two Chinese medical NER datasets demonstrate that the proposed method significantly outperforms competing approaches. In-depth analysis is also conducted to validate the effectiveness of each proposed module. Although the texts in our datasets are derived from real medical natural language texts and contain flat and overlapped entities due to the professional and complex nature of medical texts, they have been well-chosen and annotated. Real industrial application scenarios may present additional challenges, such as non-standard expressions, redundant information in the texts, and the scarcity of annotated data, which increase the difficulty of the CNER task. In future work, we will continue to explore how to improve the model's performance in real-world, low-resource application scenarios.

Acknowledgement: The authors sincerely thank all the scholars who have contributed to this field of study. It is by standing on the shoulders of those who have gone before us that we have been able to achieve our research results. We are also grateful to the reviewers and editors whose suggestions and comments have helped refine and improve this paper.

Funding Statement: This work is supported by Yunnan Provincial Major Science and Technology Special Plan Projects (Grant Nos. 202202AD080003, 202202AE090008, 202202AD080004, 202302AD080003) and National Natural Science Foundation of China (Grant Nos. U21B2027, 62266027, 62266028, 62266025), Yunnan Province Young and Middle-Aged Academic and Technical Leaders Reserve Talent Program (Grant No. 202305AC160063).

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Yan Xiang, Xuedong Zhao; data collection: Zhiliang Shi, Enbang Chen, Xiaobo Zhang; analysis and interpretation of results: Yan Xiang, Xuedong Zhao, Junjun Guo; draft manuscript preparation: Yan Xiang, Xuedong Zhao. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: CMeEE-V2 (https://tianchi.aliyun.com/dataset/95414. Accessed: Mar. 27, 2024), Diakg (https://tianchi.aliyun.com/dataset/88836. Accessed: Mar. 27, 2024).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2020. doi: 10.1109/TKDE.2020.2981314.

[2] N. Alsaaran, N. Alrabiah, and M. Alrabiah, "Arabic named entity recognition: A BERT-BGRU approach," *Comput. Mater. Contin.*, vol. 68, no. 1, pp. 471–485, 2021. doi: 10.32604/cmc.2021.016054.

[3] P. Liu, Y. Guo, F. Wang, and G. Li, "Chinese named entity recognition: The state of the art," *Neurocomputing*, vol. 473, no. 1, pp. 37–53, 2022. doi: 10.1016/j.neucom.2021.10.101.

[4] W. Wu, C. Zhang, S. Niu, and L. Shi, "Unify the usage of lexicon in Chinese named entity recognition," in *Int. Conf. Database Syst. Adv. Appl.*, Tianjin, China, 2023, pp. 665–681.

[5] Y. Zhang and J. Yang, "Chinese NER using lattice LSTM," arXiv preprint arXiv:1805.02023, 2018.

[6] S. Wu, X. Song, Z. Feng, and X. J. Wu, "NFLAT: Non-flat-lattice transformer for Chinese named entity recognition," arXiv preprint arXiv:2205.05832, 2022.

[7] J. Liu, C. Liu, N. Li, S. Gao, M. Liu and D. Zhu, "LADA-Trans-NER: Adaptive efficient transformer for Chinese named entity recognition using lexicon-attention and data-augmentation," in *Proc. AAAI Conf. Artif. Intell.*, Washington DC, USA, Jun. 2023, vol. 37, no. 11, pp. 13236–13245. doi: 10.1609/aaai.v37i11.26554.
[8] Y. Gu *et al.*, "Delving deep into regularity: A simple but effective method for Chinese named entity recognition," arXiv preprint arXiv:2204.05544, 2022.
[9] J. Li *et al.*, "Unified named entity recognition as word-word relation classification," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 10, pp. 10965–10973.
[10] J. Su *et al.*, "Global pointer: Novel efficient span-based approach for named entity recognition," arXiv preprint arXiv:2208.03054, 2022.
[11] H. Yan, Y. Sun, X. Li, and X. Qiu, "An embarrassingly easy but strong baseline for nested named entity recognition," arXiv preprint arXiv:2208.04534, 2022.
[12] Z. Zhen and J. Gao, "Chinese cyber threat intelligence named entity recognition via RoBERTa-wwm-RDCNN-CRF," *Comput. Mater. Contin.*, vol. 77, no. 1, pp. 299–323, 2023. doi: 10.32604/cmc.2023.042090.
[13] J. Luo *et al.*, "A federated named entity recognition model with explicit relation for power grid," *Comput. Mater. Contin.*, vol. 75, no. 2, pp. 4207–4216, 2023. doi: 10.32604/cmc.2023.034439.
[14] H. Zhang *et al.*, "Building a pediatric medical corpus: Word segmentation and named entity annotation," in *Chinese Lexical Semant.*, Hong Kong, China, 2021, vol. 21, pp. 652–664.
[15] D. Chang *et al.*, "Diakg: An annotated diabetes dataset for medical knowledge graph construction," in *Knowledge Graph and Semantic Computing*, Guangzhou, China: Springer, 2021, vol. 6, pp. 308–314.
[16] J. Liu *et al.*, "TOE: A grid-tagging discontinuous NER model enhanced by embedding tag/word relations and more fine-grained tags," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 177–187, 2022. doi: 10.1109/TASLP.2022.3221009.
[17] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint arXiv:1508.01991, 2015.
[18] Z. Yuan *et al.*, "Fusing heterogeneous factors with triaffine mechanism for nested named entity recognition," arXiv preprint arXiv:2110.07480, 2021.
[19] X. Li *et al.*, "FLAT: Chinese NER using flat-lattice transformer," arXiv preprint arXiv:2004.11795, 2020.
[20] S. Wu, X. Song, and Z. Feng, "MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition," arXiv preprint arXiv:2107.05418, 2021.
[21] J. Yu, B. Bohnet, and M. Poesio, "Named entity recognition as dependency parsing," arXiv preprint arXiv:2005.07150, 2020.
[22] E. Zhu, Y. Liu, and J. Li, "Deep span representations for named entity recognition," arXiv preprint arXiv:2210.04182, 2022.
[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Med. Image Comput. Comput.-Assist. Interven.*, Munich, Germany, Oct. 2015, pp. 234–241.
[24] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021. doi: 10.1109/ACCESS.2021.3086020.
[25] Q. Liu *et al.*, "Incomplete utterance rewriting as semantic segmentation," arXiv preprint arXiv:2009.13166, 2020.
[26] N. Zhang *et al.*, "Document-level relation extraction as semantic segmentation," arXiv preprint arXiv:2106.03618, 2021.
[27] R. Ma *et al.*, "Simplify the usage of lexicon in Chinese NER," arXiv preprint arXiv:1908.05969, 2019.
[28] Y. Wang *et al.*, "TPLinker: Single-stage joint extraction of entities and relations through token pair linking," arXiv preprint arXiv:2010.13415, 2020.
[29] Z. Wu *et al.*, "Grid tagging scheme for aspect-oriented fine-grained opinion extraction," arXiv preprint arXiv:2010.04640, 2020.

[30] L. Yuan, Y. Cai, J. Wang, and Q. Li, "Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging," in *Proc. AAAI Conf. Artif. Intell.*, Washington DC, USA, Jun. 2023, vol. 37, no. 9, pp. 11051–11059. doi: 10.1609/aaai.v37i9.26309.
[31] J. Devlin *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[32] G. Lample *et al.*, "Neural architectures for named entity recognition," arXiv preprint arXiv:1603.01360, 2016.
[33] Z. Dai *et al.*, "Transformer-XL: Attentive language models beyond a fixed-length context," arXiv preprint arXiv:1901.02860, 2019.
[34] A. Vaswani *et al.*, "Attention is all you need," in *Adv. Neural Inform. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.
[35] J. Su, "Conditional text generation based on conditional layer normalization," (in Chinese), 2019. Accessed: Mar. 27, 2024. [Online]. Available: https://spaces.ac.cn/archives/7124
[36] B. Yu *et al.*, "Semi-open information extraction," in *Proc. Web Conf. 2021*, Ljubljana, Slovenia, Apr. 2021, pp. 1661–1672.
[37] Y. Wang *et al.*, "Discontinuous named entity recognition as maximal clique discovery," arXiv preprint arXiv:2106.00218, 2021.
[38] Y. Nongming *et al.*, "A joint model for entity boundary detection and entity span recognition," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 10, pp. 8362–8369, 2022.
[39] T. Dozat and C. D. Manning, "Deep biaffine attention for neural dependency parsing," arXiv preprint arXiv:1611.01734, 2016.
[40] Y. Yao *et al.*, "DocRED: A large-scale document-level relation extraction dataset," arXiv preprint arXiv:1906.06127, 2019.
[41] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 3–19.
[42] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," arXiv preprint arXiv:1612.03928, 2016.
[43] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.
[44] J. Li *et al.*, "MRN: A locally and globally mention-based reasoning network for document-level relation extraction," in *Find. Assoc. Comput. Linguist.: ACL-IJCNLP 2021*, Aug. 2021, pp. 1359–1370.
[45] T. Mikolov *et al.*, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.