ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Research Article

https://doi.org/10.1631/ENG.ITEE.2025.0187

Dual-stream fusion with dynamic grid interaction for Chinese medical named entity recognition

Bing LI[1], Zhanming GONG[1], Zhiqiang ZHANG[1], Haiyu SONG[1], Yuankang SUN[2]✉

[1]*School of Information Technology and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou 310018, China*
[2]*School of Computer Science and Engineering, Southeast University, Nanjing 211189, China*

Abstract: Chinese medical named entity recognition (CMNER) is a fundamental task in medical information extraction. It is crucial for building downstream applications, such as clinical knowledge graphs, and enabling intelligent clinical decision-making. However, existing mainstream approaches, including lexicon-enhanced, span-based, and grid-based tagging methods, struggle with the absence of natural boundaries, complex nested entity structures, and long-range contextual dependencies inherent in clinical texts. To address these challenges, we propose a novel dual-stream fusion with dynamic grid interaction (D2GI) model, which performs deep semantic mining by integrating complementary feature streams and adaptive grid interactions to accurately capture complex entity boundaries and inter-character relations. Specifically, our dual-stream fusion architecture leverages RoFormer to extract long-range dependencies and incorporates Word2Vec to provide stable prior semantics, thereby enhancing the representation of rare medical terms. Furthermore, to overcome the limitations of static refinement, the dynamic grid interaction module employs a gated attention mechanism to adaptively fuse local and global contexts, facilitating accurate recognition of nested entities. Multiple experiments on three public datasets demonstrate that D2GI is superior to state-of-the-art baselines, achieving F1-score improvements of 1.48 percentage points (PPs) on CMeEE-V2, 1.61 PPs on DiaKG, and 3.08 PPs on CCKS2020.

Key words: Dual-stream fusion; Dynamic grid interaction; Medical semantic mining; Medical information extraction

1 Introduction

Chinese medical named entity recognition (CMNER) is a fundamental information extraction technology for downstream applications like clinical decision support (Xu X and Ma, 2025; Yan Y et al., 2025). However, extracting entities from unstructured Chinese medical texts presents multiple challenges. First, unlike English, Chinese lacks natural word delimiters. This continuous character sequence causes boundary ambiguity (Li XN et al., 2020; Liu P et al., 2022). Second, clinical texts are replete with complex entity structures, as illustrated in Fig. 1. Finally, accurate categorization requires capturing long-range dependencies (Xiao et al., 2024), which is challenging in lengthy medical records (Zhang HY et al., 2025). These compounded issues limit the clinical applicability of existing methods.

Type	Example	Entity type
Flat	在当地再行MR检查仍不能排除肝癌可能	Image diseases
Shared boundary	胸片可见不同程度的梗阻性肺气肿	Pro dis sym
Nested	伴有虚弱、厌食、说话和眼球震颤、抽搐、瘫痪及复合感觉障碍	Sym bod dis

Fig. 1 An example of complex entity structures in Chinese medical text. Pro, dis, sym, and bod indicate medical procedures, diseases, symptoms, and the body, respectively

Early CMNER relied on sequence labeling frameworks like bidirectional encoder representations from Transformers (BERT)-bidirectional long short-term memory (BiLSTM)-conditional random field (CRF) (Dai et al., 2019; Xu L et al., 2021), which assign a single label per character. Subsequent

✉ Yuankang SUN, syk@seu.edu.cn
Yuankang SUN, https://orcid.org/0000-0003-3217-020X
CLC number: TP391

Received: Dec. 24, 2025; Revision accepted: May 22, 2026;
Crosschecked: June 13, 2026; Published online: July 16, 2026

work explored two main directions. First, span-based models reformulate the task as dependency parsing (Yu et al., 2020). They naturally support nested structures by classifying all possible spans. Second, grid-based frameworks, such as W^2NER (Li JY et al., 2022), reformulate NER as character–pair relation classification. By modeling two-dimensional spatial relations and internal "next-neighboring-word (NNW)" connections, they effectively address nested entities.

While effective for flat entities, sequence labeling struggles with nested structures (Wang J et al., 2020; Shen et al., 2023). Furthermore, current span- and grid-based approaches retain critical limitations. Span-based models weakly capture internal semantic continuity and incur high computational costs from span enumeration. Meanwhile, many grid-based models use BERT-like encoders, which struggle to capture relative positional information for long-range dependencies (Jin et al., 2022; Su et al., 2024). Additionally, most grid-based models use static feature refinement, like fixed convolutional layers. These lack the adaptivity to dynamically aggregate context based on character–pair semantics.

To address these issues, we propose the dual-stream fusion with dynamic grid interaction (D2GI) model for grid-based relation classification. Specifically, D2GI features a dual-stream fusion architecture combining RoFormer (Su et al., 2024) and Word2Vec (Zhu et al., 2019). RoFormer captures long-range dependencies via relative positional encoding, while Word2Vec provides stable semantic priors for rare terminology. To replace static convolutions, we introduce a dynamic grid interaction module with a gated attention mechanism. This adaptively fuses local and global contexts based on character–pair semantics, refining representations for overlapping boundaries. The key contributions of this article are as follows:

1. We propose a novel dual-stream fusion architecture to address the challenges of complex context understanding and specialized terminology representation. By integrating RoFormer and Word2Vec, it captures long-range dependencies in lengthy texts while providing reliable semantic priors for rare medical terms, ensuring comprehensive initial representations.

2. We design a dynamic grid interaction module to replace static convolutional extractors. Using gated attention, it adaptively fuses local and global contexts based on character–pair semantics, resolving complex and overlapping entity boundaries.

3. Multiple experiments conducted on three public medical datasets with varying nesting ratios and domain characteristics demonstrate that our proposed D2GI model achieves superior performance over various state-of-the-art methods. Besides, comprehensive ablation studies further validate the architectural efficacy and individual contribution of each proposed module.

2 Related works

In this section, we review the three mainstream methods for CMNER: dictionary and feature fusion-based methods, span-based classification methods, and grid-based tagging methods. These approaches have collectively driven advancements in extracting information from complex medical texts and form the foundation for our proposed model.

2.1 Dictionary and feature fusion-based CMNER

To address the challenges of ambiguous word boundaries and difficult terminology recognition in Chinese NER, a major line of research focuses on enhancing model representations by fusing external dictionaries or multi-source features. For instance, the flat-lattice Transformer (FLAT) model (Li XN et al., 2020) converts lexicon-based lattice structures into flattened span sequences, enabling standard Transformers to directly model character–word interactions. To address multi-granularity entity definitions, multi-metadata embedding based cross-Transformer (MECT) (Wu et al., 2021) employs a cross-Transformer network to integrate character-, word-, and radical-level information. Similarly, Wang Y et al. (2025) designed a cross-attention mechanism to explicitly capture character–radical dependencies, thereby reinforcing technical term comprehension. However, while effective, these methods either depend on the quality and coverage of external dictionaries, which are costly to build and maintain in specialized domains, or significantly increase model complexity while enhancing representations. Furthermore, terminology standardization has emerged as a crucial factor in enhancing semantic robustness. Weigang and Brom (2025) introduced a back-translation framework to achieve dynamic semantic embedding, offering a robust baseline for resolving domain-specific ambiguity. Motivated by these insights, our D2GI model adopts a dual-stream fusion architecture. By integrating RoFormer and Word2Vec, we construct a robust, self-contained initial semantic representation that establishes stable terminology priors without over-relying on external discrete lexicons.

2.2 Span-based CMNER

To handle nested entities, span-based methods like the global pointer (GP) model proposed by Su et al. (2022) have ingeniously redefined the nested NER task as a problem of identifying pairs of "head" and "tail" pointers in a two-dimensional (2D) matrix, allowing for the indiscriminate recognition of both nested and non-nested entities. Building on this, researchers have continued to achieve improvements. The regularity-inspired recognition network (RICON) model, proposed by Gu et al. (2022), enhances boundary information by deeply analyzing the internal composition and naming rules of entities, achieving excellent performance on several Chinese NER datasets. Yan H et al. (2023) proposed a multi-head biaffine decoder that treats the feature matrix of all possible spans as a multi-channel image and applies convolutional modules to capture spatial features, achieving strong recognition performance. Despite their ability to recognize nested entities, span-based methods are constrained by exhaustive span enumeration. This results in high computational overhead and requires a maximum span length, which limits the extraction of long medical entities. Moreover, by focusing on boundary matching, these models have limited capacity to capture internal semantic continuity.

2.3 Grid-based tagging CMNER

To uniformly handle flat, nested, and even discontinuous entities within a single framework, Li JY et al. (2022)

introduced the pioneering grid-based relation classification model, W^{2}NER. This paradigm reformulates the NER task as a relation classification problem on a 2D grid, with the goal of identifying the relation between any two characters in a sentence. This effectively reduces all types of entity recognition problems to the judgment of inter-character relations. The character–pair-based method with multi-feature representation and attention mechanism (CPMFA) model, proposed by Ji et al. (2025), similarly adopts the character–pair relation classification approach and introduces a pyramid squeeze attention module to refine grid features. The HiNER model, proposed by Hou et al. (2025), builds upon W^{2}NER by introducing multi-source information and replacing the static convolutional network with a hierarchical attention fusion module to enhance the dynamism of feature extraction. While these grid-based architectures demonstrate robust structural adaptivity, they frequently struggle with long-range dependency modeling and rely heavily on rigid feature extraction pipelines. Inheriting the core grid classification mechanism, our D2GI model systematically addresses these limitations by introducing relative positional parameterization and dynamic grid interactions, optimizing the framework for complex clinical records. However, the current version does not explicitly model the internal structure of Chinese characters, which will be a key focus in future work.

3 Methodology

In this section, we present the D2GI model, designed for the CMNER task. The overall architecture of the model, as illustrated in Fig. 2, comprises four main components: a dual-stream fusion layer, a dynamic grid interaction layer, a co-predictor layer, and a decoding layer. Initially, the dual-stream fusion layer leverages the pre-trained RoFormer and Word2Vec models, along with a BiLSTM network, to mine deep medical semantics. Subsequently, the dynamic grid interaction layer constructs and refines a character–pair grid representation by fusing medical semantics, entity distance, entity position, and attention-based medical information. Finally, joint semantic inference is performed using biaffine and multilayer perceptron (MLP) classifiers to determine the relations between all medical character pairs, from which entities are ultimately decoded.

3.1 Problem definition

Following the work of Li JY et al. (2022), we define the CMNER task as a problem of modeling relations between character pairs within medical texts. This character-level granularity allows the model to capture entity boundaries directly through atomic interactions, eliminating the dependency on external word segmentation. Consequently, the model circumvents potential error propagation caused by ambiguous word boundaries in specialized Chinese medical domains. This is then framed as a grid-based tagging task, where all possible entity pairs are labeled with pre-defined tags. For an input sentence with N characters, represented as $S = \{x_1, x_2, \ldots, x_N\}$, where x_i is the i^{th} character, our goal is to identify the relation R between each pair of characters (x_i, x_j). The pre-defined labels among R include three types: none, next neighboring char (NNC), and tail-head char-* (THC-*). These labels are defined as follows:

1. None: It indicates the absence of any defined relation between (x_i, x_j).

2. NNC: It indicates that the character pair (x_i, x_j) belongs to the same entity and that x_j immediately follows x_i.

3. THC-*: It indicates that in the grid, character x_i is the tail of an entity and x_j is the head of the same entity, where $*$ denotes the entity type.

For better understanding, we provide an example as shown in Fig. 3. Through this character–pair relation modeling approach, we can simultaneously recognize flat and nested entities, which is particularly crucial for the CMNER task, as medical texts contain a high proportion of nested entities.

3.2 Dual-stream fusion layer

3.2.1 Dual-stream feature representation

This architecture consists of a dynamic stream and a static stream. The dynamic stream utilizes the RoFormer pre-trained language model proposed by Su et al. (2024). Unlike the absolute position encoding used in BERT (Devlin et al., 2019), RoFormer features a unique rotary position embedding (RoPE) mechanism. In the absence of explicit word delimiters, RoPE enables the model to parameterize the distance-dependent semantic interactions between tokens. This mechanism facilitates the extraction of structural dependencies within continuous character sequences, effectively capturing the internal semantic structure of lengthy medical entities. By effectively modeling relative positional information, this mechanism robustly processes medical texts of varying lengths (Albahli, 2025). Consequently, it directly addresses the W^{2}NER baseline's deficiency in capturing long-range dependencies within extended clinical records (Jia et al., 2024). For an input sentence $S = \{x_1, x_2, \ldots, x_N\}$ composed of N characters, RoFormer generates a context-dependent dynamic feature vector $\boldsymbol{e}_i^{\mathrm{r}}$ for each character x_i, as shown in the equation below:

$$\boldsymbol{e}_i^{\mathrm{r}} = \mathrm{RoFormer}(x_i), \tag{1}$$

where $\boldsymbol{e}_i^{\mathrm{r}} \in \mathbb{R}^{d_{\mathrm{c}}}$, and d_{c} is the dimension of character embedding.

While RoFormer generates powerful dynamic representations, these representations usually overlook the general core meaning of medical terms in specific contexts. To address this issue and introduce richer medical prior knowledge, the static stream incorporates static word vectors (Word2Vec). By fusing pre-trained word-level representations into the character-level sequence, the model naturally acquires implicit boundary cues, which effectively mitigates semantic ambiguity in unsegmented Chinese texts. Word2Vec provides a context-independent medical semantic benchmark for each term, which has been proven to be an effective supplement in CMNER (Liu T et al., 2023). We introduce a pre-trained Chinese biomedical Word2Vec model constructed from a vocabulary of 5000 medical terms that appear more than 200 times in a large medical corpus (https://github.com/WENGSYX/Chinese-Word2vec-Medicine). For each character x_i in the input sentence, we obtain its static medical representation $\boldsymbol{e}_i^{\mathrm{w}}$ in the Word2Vec

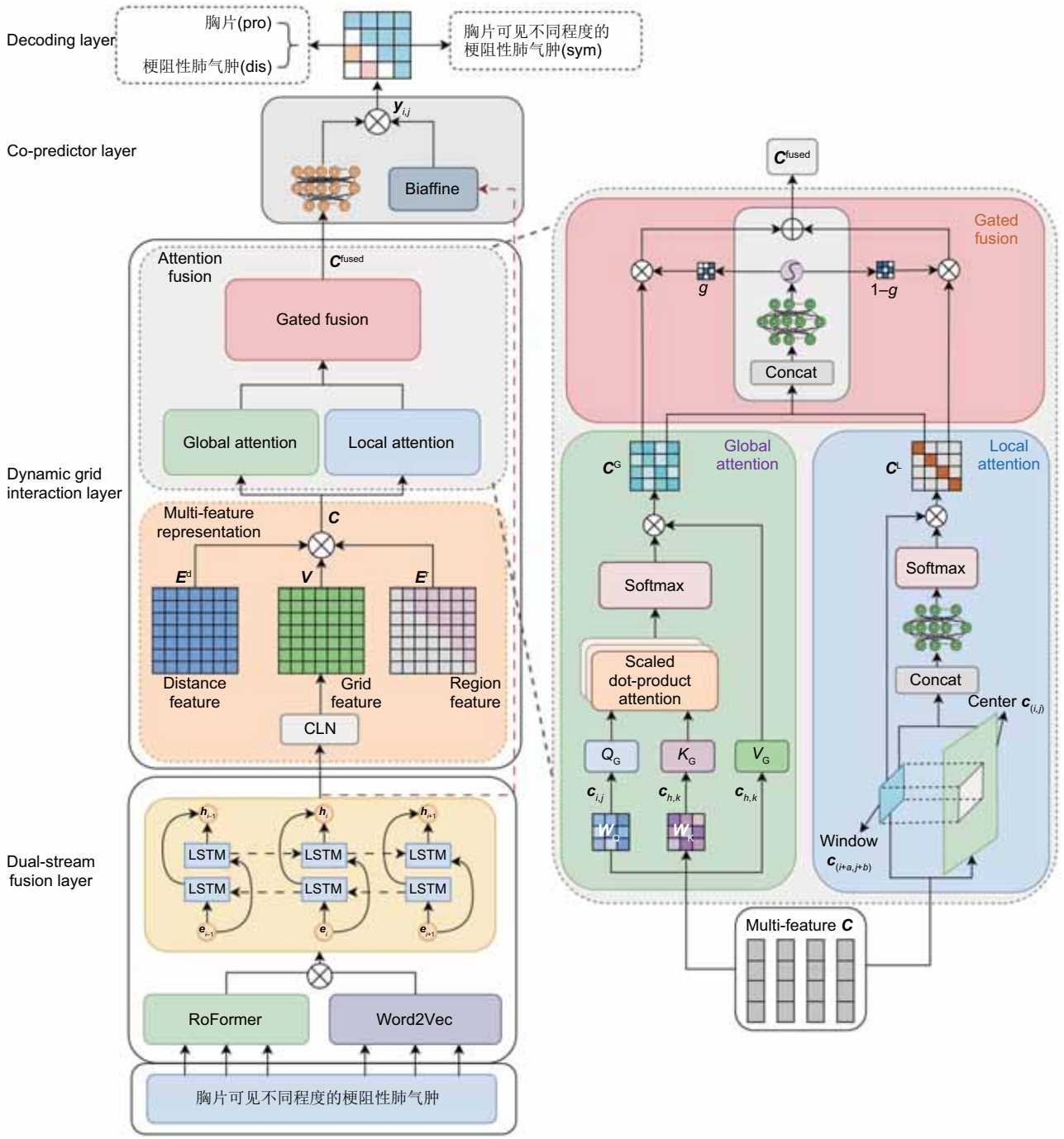

Fig. 2 Overview of our dual-stream fusion with dynamic grid interaction (D2GI) model. ⊗ represents the splicing operation of the elements. CLN: conditional layer normalization

vector space through a lookup table, which is defined as

$$\boldsymbol{e}_i^{\mathrm{w}} = \mathrm{Word2Vec}(x_i), \tag{2}$$

where $\boldsymbol{e}_i^{\mathrm{w}} \in \mathbb{R}^{d_{\mathrm{w}}}$, and d_{w} is the dimension of word vector.

3.2.2 Dual-stream feature fusion

To simultaneously leverage the dynamic contextual features and static semantic priors, we concatenate $\boldsymbol{e}_i^{\mathrm{r}}$ and $\boldsymbol{e}_i^{\mathrm{w}}$:

$$\boldsymbol{e}_i^{\mathrm{fused}} = \mathrm{Concat}(\boldsymbol{e}_i^{\mathrm{r}}, \boldsymbol{e}_i^{\mathrm{w}}). \tag{3}$$

The resulting fused vector sequence is $\boldsymbol{E}^{\mathrm{fused}} = \{\boldsymbol{e}_1^{\mathrm{fused}}, \boldsymbol{e}_2^{\mathrm{fused}}, \ldots, \boldsymbol{e}_N^{\mathrm{fused}}\} \in \mathbb{R}^{N\times(d_{\mathrm{c}}+d_{\mathrm{w}})}$, an approach that losslessly preserves both types of information. Subsequently, we utilize a BiLSTM network (Méndez et al., 2023) as the sequence encoder to further aggregate contextual dependencies. By processing the sequence bidirectionally, BiLSTM effectively integrates the full left and right contexts for each character, mapping the fused features into a comprehensive semantic representation matrix $\boldsymbol{H}$:

$$\boldsymbol{H} = \mathrm{BiLSTM}(\boldsymbol{E}^{\mathrm{fused}}), \tag{4}$$

where $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N\} \in \mathbb{R}^{N\times d_{\mathrm{l}}}$, and d_{l} is the dimension of word representation.

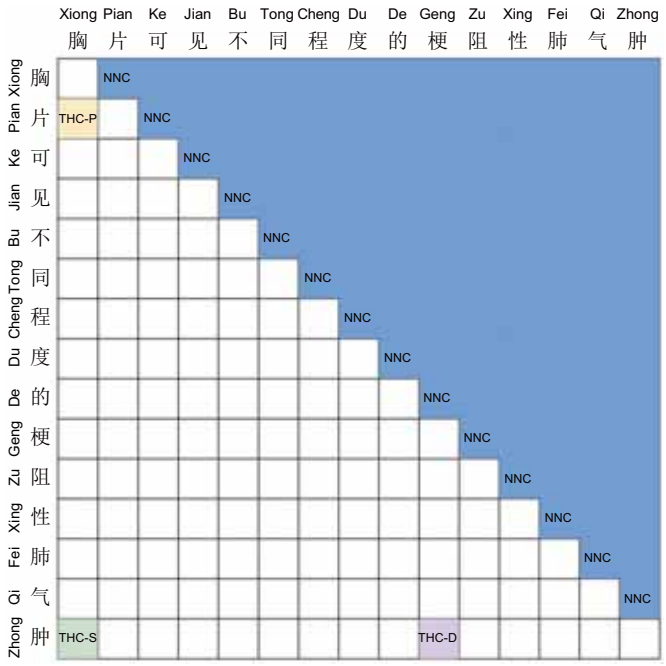

Fig. 3 Schematic diagram of word–word relation. The grid contains three predefined relations, including None, NNC, and THC-*. THC-P indicates that the entity type is medical procedure, THC-S indicates that the entity type is symptom, and THC-D indicates that the entity type is disease

3.3 Dynamic grid interaction layer

After obtaining the deep sequential medical representation $\boldsymbol{H}$ from the dual-stream fusion layer, we model the complex pairwise relations by transforming the one-dimensional (1D) sequence into a 2D grid representation. This transformation is essential for a grid-based framework to identify nested and overlapping entity boundaries. Therefore, we design the dynamic grid interaction layer to construct and subsequently refine a character–pair interaction grid, enabling it to capture features at both local and global granularities.

3.3.1 Conditional layer normalization

To initiate this grid construction, we first employ conditional layer normalization (CLN) (Liu RB et al., 2021) to model the interaction between any two character representations $(\boldsymbol{h}_i, \boldsymbol{h}_j)$ from the sequence $\boldsymbol{H}$. This process constructs a semantic interaction grid $\boldsymbol{V} \in \mathbb{R}^{N\times N\times d_l}$. Each element $\boldsymbol{V}_{i,j}$ represents the interaction representation of the character pair, formulated as

$$\boldsymbol{V}_{i,j} = \mathrm{CLN}(\boldsymbol{h}_i, \boldsymbol{h}_j), \tag{5}$$

where $\boldsymbol{h}_i$ serves as the conditioning vector, and $\boldsymbol{h}_j$ is the input vector to be normalized.

3.3.2 Multi-feature representation

While the semantic interaction grid $\boldsymbol{V}$ implicitly inherits relative positional information from RoFormer, these spatial cues are deeply entangled with the semantic content. To provide explicit structural priors for precise boundary recognition, we introduce two independent geometric features to construct a multi-feature grid representation. Specifically, this representation integrates three components: (1) semantic interaction grid $\boldsymbol{V}$; (2) a feature matrix $\boldsymbol{E}^{\mathrm{d}}$ representing the relative distance information between character pairs (x_i, x_j) (This learnable embedding is generated by passing a pre-computed grid of integer distance indices through an embedding layer); (3) a feature matrix $\boldsymbol{E}^{\mathrm{r}}$ containing information about whether a character is in the upper triangular region $(i < j)$ or lower triangular region $(i > j)$. This learnable feature is generated by first assigning an integer index to each grid cell based on its location and then passing these indices through a separate embedding layer. This provides the model with structural information about entity directionality, which is crucial for distinguishing the head and tail of nested medical entities. Finally, we concatenate these three feature representations along the feature dimension to obtain the final multi-feature representation grid $\boldsymbol{C}$:

$$\boldsymbol{C} = \mathrm{Concat}(\boldsymbol{V}, \boldsymbol{E}^{\mathrm{d}}, \boldsymbol{E}^{\mathrm{r}}), \tag{6}$$

where $\boldsymbol{C} \in \mathbb{R}^{N\times N\times d_{\mathrm{grid}}}$ $(d_{\mathrm{grid}} = d_{\mathrm{l}} + d_{\mathrm{d}} + d_{\mathrm{r}})$ is the complete grid representation. Each element $\boldsymbol{c}_{i,j}$ in this grid represents the fused feature vector for the character pair (x_i, x_j).

3.3.3 Attention fusion

The input grid $\boldsymbol{C}$ fuses semantic and geometric features, but its representation remains restricted to local interactions. This is insufficient for Chinese medical texts, where entity identification requires combining contextual information across different granularities. Unlike traditional grid-based models that rely on static convolutional layers with fixed receptive fields, we design a dynamic attention fusion module, inspired by Hou et al. (2025), to capture both dependency types. This module concurrently computes local and global attention, employing a dynamic gating mechanism (Zhang GS et al., 2024) to adaptively assign fusion weights based on specific character–pair semantics. For example, a local pair like "肺–肿" (lung–swelling) relies more on local attention, while a distant pair like "胸片–肺" (chest X-ray–lung) requires global attention.

Global attention is specifically employed to capture long-range dependencies. This mechanism evaluates the semantic relevance between a specific character pair $\boldsymbol{c}_{i,j}$ and all other elements in the grid, effectively bypassing the spatial constraints of standard convolutions. By introducing learnable projection matrices, $\boldsymbol{W}_{\mathrm{Q}}$ and $\boldsymbol{W}_{\mathrm{K}}$, the global context representation $\boldsymbol{C}^{\mathrm{G}}_{i,j}$ is obtained via a single scaled dot-product attention operation:

$$\boldsymbol{C}^{\mathrm{G}}_{i,j} = \sum_{h=1}^{N}\sum_{k=1}^{N} \mathrm{Softmax}\left(\frac{(\boldsymbol{W}_{\mathrm{Q}}\boldsymbol{c}_{i,j})(\boldsymbol{W}_{\mathrm{K}}\boldsymbol{c}_{h,k})^{\mathrm{T}}}{\sqrt{d_{\mathrm{grid}}}}\right)\cdot \boldsymbol{c}_{h,k}, \tag{7}$$

where d_{grid} denotes the feature dimension of the grid and $\boldsymbol{C}^{\mathrm{G}} \in \mathbb{R}^{N\times N\times d_{\mathrm{grid}}}$.

Complementing the global stream, local attention enhances region-specific features. For each element $\boldsymbol{c}_{i,j}$ in the grid, we define a $k \times k$ sliding window Ω centered at it. While static convolutional kernels aggregate features using fixed weights, our local mechanism dynamically computes attention scores based on the semantic similarity between the central element $\boldsymbol{c}_{i,j}$ and its neighbors $\boldsymbol{c}_{i+a,j+b}$. By integrating the scoring and normalization steps, the local context $\boldsymbol{C}^{\mathrm{L}}_{i,j}$ is

aggregated by a single weighted sum as follows:

$$\boldsymbol{C}^{\mathrm{L}}_{i,j} = \sum_{\boldsymbol{c}_{i+a,j+b}\in\Omega} \mathrm{Softmax}\big(\boldsymbol{W}_{\mathrm{l}}[\boldsymbol{c}_{i,j};\boldsymbol{c}_{i+a,j+b}]+\boldsymbol{b}_{\mathrm{l}}\big)\cdot \boldsymbol{c}_{i+a,j+b}, \quad (8)$$

where the window Ω comprises all neighbors with spatial offsets a and b falling within $[-\lfloor\frac{k}{2}\rfloor, \lfloor\frac{k}{2}\rfloor]$. Here, $\boldsymbol{W}_{\mathrm{l}}$ and $\boldsymbol{b}_{\mathrm{l}}$ denote the learnable parameters of the linear layer, and $[\cdot;\cdot]$ represents concatenation. This process gains the local context $\boldsymbol{C}^{\mathrm{L}} \in \mathbb{R}^{N\times N\times d_{\mathrm{grid}}}$.

Finally, rather than relying on static fusion strategies, we employ a dynamic gating mechanism for instance-aware integration. This mechanism evaluates the concatenated representations to generate an adaptive weight vector $\boldsymbol{g} = \sigma\big(\boldsymbol{W}_{\mathrm{g}}[\boldsymbol{C}^{\mathrm{G}};\boldsymbol{C}^{\mathrm{L}}]+\boldsymbol{b}_{\mathrm{g}}\big)$. Guided by this learned gate, the model adaptively fuses the two context streams:

$$\boldsymbol{C}^{\mathrm{fused}} = \boldsymbol{g}\odot\boldsymbol{C}^{\mathrm{G}} + (1-\boldsymbol{g})\odot\boldsymbol{C}^{\mathrm{L}}. \quad (9)$$

Herein $\boldsymbol{W}_{\mathrm{g}}$ and $\boldsymbol{b}_{\mathrm{g}}$ are learnable projection parameters, and $\odot$ represents element-wise multiplication. The final output of this layer is the multi-grained semantic grid $\boldsymbol{C}^{\mathrm{fused}} \in \mathbb{R}^{N\times N\times d_{\mathrm{grid}}}$.

3.4 Co-predictor layer

Subsequently, we use a co-predictor layer composed of an MLP predictor and a biaffine predictor (Ye and Teufel, 2021) to jointly infer the relations between all medical character pairs. This layer utilizes two parallel branches to leverage features from different stages of the model.

The first branch, an MLP predictor, operates on the final fused grid representation $\boldsymbol{C}^{\mathrm{fused}}$. Its purpose is to capture the complex, high-dimensional medical semantic interactions that are refined by the dynamic grid interaction layer:

$$\boldsymbol{y}'_{i,j} = \mathrm{MLP}(\boldsymbol{C}^{\mathrm{fused}}_{i,j}), \quad (10)$$

where $\boldsymbol{C}^{\mathrm{fused}}_{i,j}$ is the feature vector for the character pair (x_i, x_j), and $\boldsymbol{y}'_{i,j}$ is the relation score vector outputted by this predictor.

The second branch, a biaffine predictor, operates directly on the sequential representation $\boldsymbol{H}$ generated by the dual-stream fusion layer. This establishes a semantic shortcut to the initial contextual representations, effectively complementing the spatially refined features of the grid. Specifically, two distinct MLPs are employed to project the sequence hidden states $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ into dedicated head and tail boundary representations, $\boldsymbol{s}_i$ and $\boldsymbol{o}_j$:

$$\boldsymbol{s}_i = \mathrm{MLP}_{\mathrm{head}}(\boldsymbol{h}_i), \quad (11)$$

$$\boldsymbol{o}_j = \mathrm{MLP}_{\mathrm{tail}}(\boldsymbol{h}_j). \quad (12)$$

The biaffine scoring function is then applied to compute the sequence-level relation score vector $\boldsymbol{y}''_{i,j}$:

$$\boldsymbol{y}''_{i,j} = \boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{U}\boldsymbol{o}_j + \boldsymbol{W}[\boldsymbol{s}_i;\boldsymbol{o}_j] + \boldsymbol{b}, \quad (13)$$

where $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{b}$ are learnable parameters.

Finally, the predictive scores from both branches are aggregated and normalized via a Softmax function to compute the ultimate probability distribution vector $\boldsymbol{y}_{i,j}$ over the predefined relation R:

$$\boldsymbol{y}_{i,j} = \mathrm{Softmax}(\boldsymbol{y}'_{i,j} + \boldsymbol{y}''_{i,j}). \quad (14)$$

3.5 Decoding layer

After the co-predictor layer calculates the relation probability distribution $\boldsymbol{y}_{i,j}$ for all character pairs, the final predicted label grid $\boldsymbol{Y} = \{\boldsymbol{Y}_{i,j}\}_{1\le i,j\le N}$ is obtained by selecting the most likely relation r from the set of all relations R:

$$\boldsymbol{Y}_{i,j} = \underset{r\in R}{\mathrm{argmax}}(\boldsymbol{y}^{r}_{i,j}). \quad (15)$$

The formal procedure for the D2GI model is detailed in Algorithm 1. Given a sequence S of N characters, the framework outputs the final set of extracted medical entities E. The model operation can be conceptually divided into two stages. First, the encoding layer generates the final predicted label grid $\boldsymbol{Y}$ by processing the input sentence through the dual-stream fusion, dynamic grid interaction, and co-predictor layers (lines 3−20). Second, the decoding layer algorithm iterates through all possible character pairs (i, j) to find a THC-* relation trigger (lines 21 and 22). When a trigger is found (line 23), the algorithm verifies the entity's internal continuity (lines 26−31). This is done by checking if all adjacent character pairs $(k, k+1)$ within the span $[j, i]$ are predicted as NNC. If this NNC path is complete and valid, the entity (j, i, T) is added to the set E (line 33). To enhance readability and intuitively illustrate this mechanism, we also provide a simplified flowchart of the decoding process in Fig. 4.

As illustrated in Fig. 5, to decode the medical entity

Algorithm 1 D2GI model

1: **Input:** sentence $S = \{x_1, x_2, ..., x_N\}$
2: **Output:** the set of decoded entities E
// Encoding layer
3: $E \leftarrow \varnothing$ // Initialize an empty set of entities
4: $\boldsymbol{e}^{\mathrm{r}} \leftarrow \mathrm{RoFormer}(x_i)$
5: $\boldsymbol{e}^{\mathrm{w}} \leftarrow \mathrm{Word2Vec}(x_i)$
6: $\boldsymbol{E}^{\mathrm{fused}} \leftarrow \mathrm{Concat}(\boldsymbol{e}^{\mathrm{r}}, \boldsymbol{e}^{\mathrm{w}})$
7: $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_N\} \leftarrow \mathrm{BiLSTM}(\boldsymbol{E}^{\mathrm{fused}})$
8: **for** $1 \le i, j \le N$ **do**
9: $\quad \boldsymbol{V}_{i,j} \leftarrow \mathrm{CLN}(\boldsymbol{h}_i, \boldsymbol{h}_j)$
10: **end for**
11: $\boldsymbol{C} \leftarrow \mathrm{Concat}(\boldsymbol{V}, \boldsymbol{E}^{\mathrm{d}}, \boldsymbol{E}^{\mathrm{r}})$
12: $\boldsymbol{C}^{\mathrm{L}} \leftarrow \mathrm{LocalAttention}(\boldsymbol{C})$
13: $\boldsymbol{C}^{\mathrm{G}} \leftarrow \mathrm{GlobalAttention}(\boldsymbol{C})$
14: $\boldsymbol{C}^{\mathrm{fused}} \leftarrow \mathrm{GatedFusion}(\boldsymbol{C}^{\mathrm{L}}, \boldsymbol{C}^{\mathrm{G}})$
15: **for** $1 \le i, j \le N$ **do**
16: $\quad \boldsymbol{y}'_{i,j} \leftarrow \mathrm{MLP}(\boldsymbol{C}^{\mathrm{fused}}_{i,j})$
17: $\quad \boldsymbol{y}''_{i,j} \leftarrow \mathrm{Biaffine}(\boldsymbol{h}_i, \boldsymbol{h}_j)$
18: $\quad \boldsymbol{y}_{i,j} \leftarrow \mathrm{Softmax}(\boldsymbol{y}'_{i,j} + \boldsymbol{y}''_{i,j})$
19: $\quad \boldsymbol{Y}_{i,j} \leftarrow \underset{r\in R}{\mathrm{argmax}}(\boldsymbol{y}^{r}_{i,j})$
20: **end for**
// Decoding layer
21: **for** $i = 1$ **to** N **do**
22: $\quad$ **for** $j = 1$ **to** i **do**
23: $\quad\quad$ **if** $\boldsymbol{Y}_{i,j}$ is a THC-* relation **then**
24: $\quad\quad\quad T \leftarrow \mathrm{GetType}(\boldsymbol{Y}_{i,j})$
25: $\quad\quad\quad$ path_valid $\leftarrow$ True
26: $\quad\quad\quad$ **for** $k = j$ **to** $i - 1$ **do**
27: $\quad\quad\quad\quad$ **if** $\boldsymbol{Y}_{k,k+1}$ is not NNC **then**
28: $\quad\quad\quad\quad\quad$ path_valid $\leftarrow$ False
29: $\quad\quad\quad\quad\quad$ **break**
30: $\quad\quad\quad\quad$ **end if**
31: $\quad\quad\quad$ **end for**
32: $\quad\quad\quad$ **if** path_valid is true **then**
33: $\quad\quad\quad\quad$ add entity (j, i, T) to E
34: $\quad\quad\quad$ **end if**
35: $\quad\quad$ **end if**
36: $\quad$ **end for**
37: **end for**
38: **Return** E

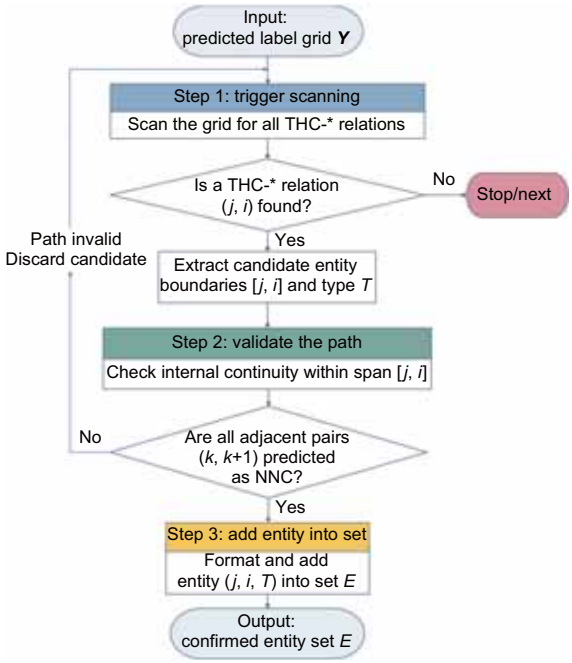

Fig. 4 The simplified flowchart of the decoding process

"胸片" (chest X-ray), the algorithm first identifies a THC-P relation in $\boldsymbol{Y}_{i,j}$ at ($i =$ "片", $j =$ "胸") (line 25). This triggers the NNC path check (line 27). The algorithm checks the relation for ($k =$ "胸", $k+1 =$ "片"), finds that it is NNC, and thus validates the path. This two-relation mechanism is decisive for identifying entities, as it effectively prevents fragmented or overlapping texts from being erroneously merged into a single entity.

3.6 Training

Our training objective is to minimize the negative log-likelihood loss between the model's predicted probability vector $\boldsymbol{y}_{i,j}$ and the corresponding gold-standard labels $\hat{\boldsymbol{y}}_{i,j}$, as formulated below:

$$\mathcal{L} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{r=1}^{R}\hat{\boldsymbol{y}}_{i,j}^{r}\log \boldsymbol{y}_{i,j}^{r}, \tag{16}$$

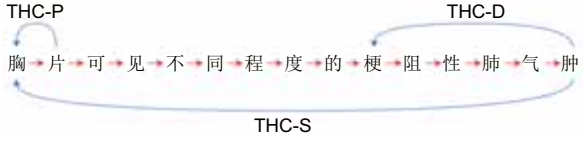

Fig. 5 An example of the entity decoding process based on relations. The red and blue arrows indicate NNC and THC-* relations, respectively

where $\hat{\boldsymbol{y}}_{i,j}^{r}$ is a binary indicator for the true relation label of the character pair (x_i, x_j).

4 Experiments

4.1 Datasets

To comprehensively and rigorously evaluate the effectiveness and generalization capability of our proposed D2GI model on the CMNER task, we conduct experiments on three representative public medical datasets. These datasets, each with a distinct focus, form a complementary evaluation suite. The details of each dataset are elaborated in Table 1. These statistics detail the number of entity categories and number of total entities, along with the average sentence length, the average entity length, and the nested ratio, which collectively profile the scale and complexity of each dataset.

CMeEE-V2 (Zhang NY et al., 2022): The corpus for this dataset is sourced from real electronic health records and includes nine fine-grained medical entity types, such as disease, symptom, drug, and procedure. We select this dataset primarily to evaluate the model's performance on fine-grained entity recognition in standard clinical texts.

DiaKG (Chang et al., 2021): Derived from authoritative Chinese diabetes literature, this dataset contains 18 coarse-grained entity categories. Its defining characteristic is a high density of complex nested structures. Consequently, DiaKG is specifically selected to validate the model's capability in resolving overlapping and nested entity boundaries.

CCKS2020 (Ma and Huang, 2022): The corpus for this dataset also originates from real-world Chinese electronic health records and contains six coarse-grained entity types—anatomical parts, disease diagnosis, drugs, laboratory tests, imaging examinations, and surgery. We choose this dataset to specifically assess our model's baseline performance on the traditional task of flat entity recognition. This allows us to verify whether the modules designed for complex structures have any adverse effects on simpler tasks, thereby examining the model's generalization capability and robustness.

4.2 Evaluation

In this study, we adopt three standard evaluation metrics commonly used in the NER domain: precision, recall, and F1-score. These metrics are calculated based on the principle of exact matching, meaning that a predicted medical entity is considered correct only if its boundaries and entity type are identical to the ground-truth medical label.

Precision (P): It refers to the proportion of medical

Table 1 Detailed statistics and composition of the three benchmark datasets

Dataset	Cate	Ent	Avg Sent Len	Avg Entity Len	Nested ratio (%)	Number of sentences		
						Training	Dev	Test
CMeEE-V2	9	108 730	54.14	5.17	36.90	12 000	4000	4000
CCKS2020	6	18 310	55.31	4.36	0.00	6214	776	778
DiaKG	18	22 050	100.31	4.45	39.99	1639	546	548

Cate: number of entity categories; Ent: number of total entities; Avg Sent Len: average sentence length; Avg Entity Len: average entity length; Dev: development

entities predicted as correct by the model that are actually correct.

Recall (R): It refers to the proportion of all true medical entities in the dataset that are successfully predicted by the model.

F1-score (F1): It refers to the harmonic mean of precision and recall, serving as the core metric for a comprehensive evaluation of the model's performance on this clinical recognition task.

4.3 Baselines

To comprehensively evaluate the performance of our proposed D2GI model, we select a series of representative baseline models, including several recently proposed state-of-the-art methods, for comparative experiments. These baselines span four mainstream technical paradigms in the NER domain:

1. Sequence labeling models: framing the NER task as a sequence labeling problem, assigning a tag to each character.

BiLSTM+CRF (Yang XM et al., 2018): It captures contextual dependencies via BiLSTM and applies a CRF layer to constrain global label transitions.

BERT+BiLSTM (Dai et al., 2019): It builds upon BiLSTM-CRF by introducing BERT as a pre-trained encoder to generate more contextually rich character representations.

2. Lexicon-enhanced models: fusing external lexicon information to enhance character representations using word boundaries.

FLAT (Li XN et al., 2020): It transforms lexicon-based lattice structures into flattened span sequences, enabling standard Transformers to directly model character–lexicon interactions.

MECT (Wu et al., 2021): It employs a cross-Transformer network to fuse multi-granularity features, capturing structural interactions across the character, word, and radical levels.

3. Span-based models: identifying entities by enumerating and classifying all text spans, naturally supporting nested structures.

GP (Su et al., 2022): It reformulates NER as a boundary identification task, utilizing a 2D matrix to parameterize the pairing of entity head and tail pointers.

Efficient GP (Zhang L et al., 2024): It optimizes the standard GP by simplifying the scoring function, thereby reducing parameter overhead and enhancing matrix scalability.

MISS (Yang LY et al., 2025): It fuses multi-granularity information to enrich span representations, integrating character- and word-level cues to improve boundary classification accuracy.

4. Grid-based tagging models: transforming NER into a 2D grid relation classification, uniformly handling all entity types.

W^{2}NER (Li JY et al., 2022): It unifies the NER task by modeling it as a word–word relation classification problem, identifying entities by predicting two types of relations: NNW and THC-*.

CPMFA (Ji et al., 2025): It is a character–pair relation classification method for Chinese nested NER, which is distinguished by its use of the LERT pre-trained model and a pyramid compression attention module to refine grid features.

CK-CMCNER (Chen et al., 2024): It is a domain-specific model that fuses multi-level features and employs cross-stream attention mechanisms to adapt the grid framework to complex clinical contexts.

4.4 Implementation

All experiments in this study are conducted on a single NVIDIA RTX 3090 GPU and developed based on the PyTorch framework. We use the AdamW optimizer for model optimization. For the embedding layer, we use roformer_v2_chinese_char_base to obtain 768-dimensional character embeddings (d_c). These are fused with 512-dimensional pre-trained Chinese biomedical word vectors (d_w). The fused vectors then pass through a BiLSTM network with a hidden dimension of 512 to obtain the final 1024-dimensional word representations (d_l). For other model components, the MLP hidden dimension is set to 288, and both the distance and region embedding sizes are set to 20. For hyper-parameter settings, the batch size is set to 16. We employ a differential learning rate, setting the rate for the RoFormer parameters to 5×10^{-6} and for all other parameters to 1×10^{-3}. Dropout rates of 0.33 and 0.5 are applied to different components of the model. To ensure robust results, we employ three different numbers of random seeds (42, 256, and 512) and repeat the experiment five times for each seed. The final results are the average of these 15 runs.

4.5 Experimental results

We evaluate D2GI on three CMNER datasets. This subsection presents the overall performance and a fine-grained category comparison against the W^{2}NER baseline.

4.5.1 Overall performance comparison

Table 2 summarizes the experimental results across all datasets. Some baseline results are temporarily missing because the original code is not publicly available. Based on these results, we make several key observations.

1. Comparison with sequence labeling models: Sequence labeling cannot natively handle nested entities. By adopting a grid-based paradigm, D2GI overcomes this, achieving a 14.91 percentage points (PPs) absolute F1-score improvement over BERT-BiLSTM-CRF on the highly nested DiaKG dataset.

2. Comparison with lexicon-enhanced models: Handling specialized terminology requires deep semantic integration. Unlike shallow lexicon adapters, D2GI's dual-stream architecture fuses dynamic contexts with static Word2Vec priors, achieving consistent improvements across all datasets (peaking at a 6.68 PPs gain on DiaKG).

3. Comparison with span-based models: Span-based approaches efficiently capture boundaries but struggle with internal semantic continuity. D2GI addresses this limitation, improving the F1-score by 4.47 PPs on CCKS2020 over EGP.

4. Comparison with grid-based tagging models: D2GI consistently outperforms the W^{2}NER baseline (1.91 PPs on CMeEE-V2, 4.15 PPs on CCKS2020, and 2.18 PPs on DiaKG). This demonstrates the overall effectiveness of our design, indicating that dual-stream fusion and dynamic grid interaction successfully complement each other to overcome the

Table 2 Performance of D2GI with representative baselines on three datasets

Method	Model	P (%)			R (%)			F1 (%)		
		CMeEE-V2	CCKS2020	DiaKG	CMeEE-V2	CCKS2020	DiaKG	CMeEE-V2	CCKS2020	DiaKG
Sequence labeling	BiLSTM+CRF	65.39	75.46	71.66	53.77	72.26	55.06	59.01	73.83	62.27
	BERT+BiLSTM	63.76	82.43*	68.23	64.92	82.79*	72.91	64.03	82.61*	70.49
Lexicon-enhanced	FLAT	61.83*	83.10	75.42*	66.42*	81.20	82.32*	64.03*	82.14	78.72*
	MECT	70.54*	83.01	72.75*	73.31*	85.41	70.91*	71.80*	84.19	71.83*
Span-based	GP	73.10*	82.02	82.43	72.25*	83.64	81.26	72.54*	82.91	81.84
	EGP	75.42	84.03	**86.98***	74.16	85.66	79.28*	74.78	84.84	82.95*
	MISS	75.78*	–	–	74.83*	–	–	75.30*	–	–
Grid tagging	W^{2}NER	74.58	83.33	81.50	75.19	87.06	85.01	74.87	85.16	83.22
	CPMFA	–	–	85.40*	–	–	82.23*	–	–	83.79*
	CK-CMCNER	–	85.92*	–	–	86.62*	–	–	86.23*	–
Ours	D2GI	**76.92**	**87.27**	85.01	**76.44**	**91.46**	**85.80**	**76.78**	**89.31**	**85.40**

The best results are highlighted in bold, and the underlined values represent the metrics of the optimal baseline model for each dataset. The results marked with * are from the original papers. The symbol "–" indicates that the results are omitted due to the unavailability of the source code and the absence of corresponding experiments in the original publications

representation bottlenecks of W^{2}NER. Notably, on DiaKG, D2GI is superior to the state-of-the-art CPMFA model, trading a 0.39 PPs precision drop for a 3.57 PPs recall increase, resulting in a 1.61 PPs overall gain in the F1-score. We attribute this to CPMFA's PolyLoss, which promotes conservative high-precision predictions. In contrast, D2GI comprehensively captures potential entities in complex clinical contexts, driving a superior F1-score.

4.5.2 Per-category performance comparison

To understand the origins of the D2GI model's performance improvements, we conduct a granular category comparison against W^{2}NER across the three datasets.

As detailed in the heatmap in Fig. 6, D2GI outperforms most medical categories in CMeEE-V2. Specifically, F1-scores in the core "DIS" and "SYM" categories increase by 1.46 PPs and 1.21 PPs, respectively. The "SYM" improvement is driven by a 3.36 PPs precision gain, indicating that our attention fusion surpasses W^{2}NER's static convolutions in distinguishing ambiguous boundaries. Furthermore, a 5.56 PPs F1-score gain in the low-resource "DEP" category highlights our dual-stream advantage. Here, pre-trained Word2Vec embeddings provide stable prior semantics, preventing misidentification of rare departmental terminology despite scarce training instances.

For the flat-entity CCKS2020 dataset (Fig. 7), D2GI surpasses W^{2}NER across all categories. F1-scores in "LABORATORY" and "IMAGE" increase by 14.36 PPs and 7.22 PPs, respectively. The "LABORATORY" enhancement stems from simultaneous precision (11.02 PPs) and recall (7.14 PPs) gains. This demonstrates that RoFormer's relative positional encoding better captures varied clinical examination phrasings than static encoders.

DiaKG's heatmap (Fig. 8) reveals the most significant module impact. Specifically, D2GI achieves a 20.69 PPs F1-score gain in the "REASON" category, where W^{2}NER obtains an F1-score of 0.00. Because "REASON" entities typically involve lengthy descriptions, this improvement proves that

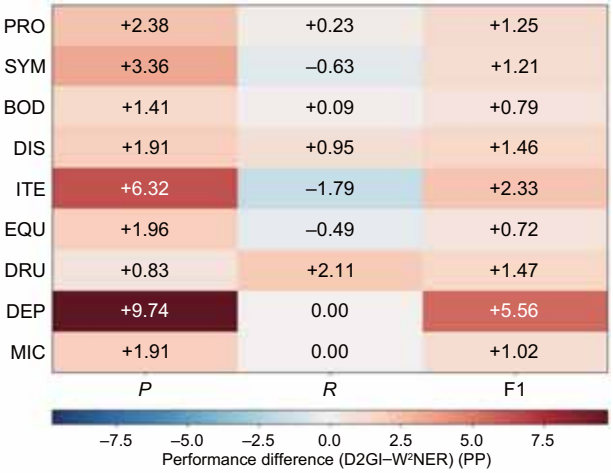

Fig. 6 Per-category performance comparison on CMeEE-V2. Category labels are dataset identifiers without inherent semantic interpretation. PP: percentage point

RoFormer's rotary position embeddings effectively capture long-range dependencies. In addition, D2GI shows modest F1-score drops in high-frequency categories like "TEST" (5.03 PPs) and "LEVEL" (7.47 PPs). This indicates that while static convolutions remain stable for structurally simple entities, our dynamic architecture specifically excels at capturing complex, rare patterns.

4.6 Ablation study

To validate the individual contributions of key architectural components, we conduct ablation studies across the three datasets under identical hyperparameter settings. The results are presented in Table 3, with the analysis as follows:

1. Dual-stream fusion. Removing the RoFormer encoder results in the most severe degradation (an 8.30 PPs F1 drop on DiaKG), confirming its role as the core semantic encoder for complex medical texts. Ablating the Word2Vec representation also leads to a 2.57 PPs decrease on DiaKG, indicating that static prior knowledge is a crucial supplement to dynamic

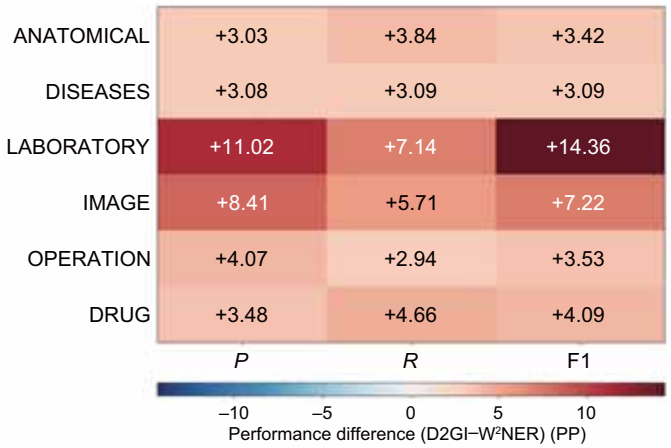

Fig. 7 Per-category performance comparison on CCKS2020. Category labels are dataset identifiers without inherent semantic interpretation

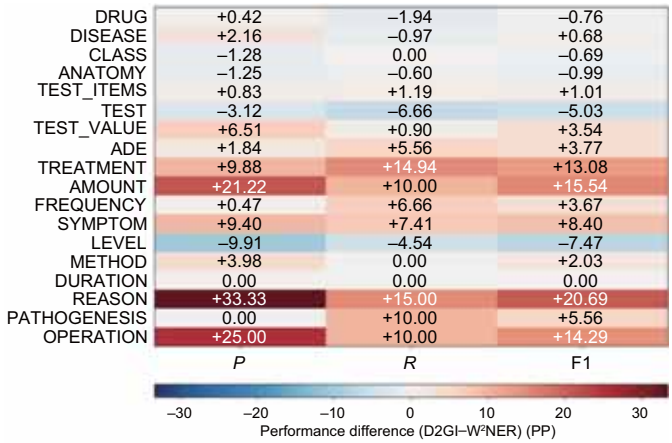

Fig. 8 Per-category performance comparison on DiaKG. The F1-score of W^{2}NER is 0.00 for the "REASON" category. Both models score 0.00 for "DURATION." Category labels are dataset identifiers without inherent semantic interpretation

encoding. Furthermore, the consistent decline after removing BiLSTM fusion (up to 1.65 PPs) highlights that its sequential refinement is indispensable for subsequent grid construction.

2. Multi-feature representation. Removing the distance and region embeddings independently results in consistent performance declines. Specifically, removing distance leads to a 0.56 PPs drop on CCKS2020, while removing region results in a 0.46 PPs decrease on DiaKG. This indicates that although RoFormer implicitly encodes relative positions, explicitly providing distance and region as independent feature channels supplies crucial inductive biases, easing the parameterization of entity boundaries and directionality.

3. Attention fusion. Ablating this entire module leads to a 0.97 PPs decrease on the structurally simpler CCKS2020 dataset, revealing its importance for adaptive feature aggregation during precise classification. Sub-component analysis confirms the necessity of the parallel design: removing local attention causes a 1.23 PPs loss on DiaKG. While removing global attention leads to a slight 0.11 PPs increase on CMeEE-V2, which suggests it may act as "noise" in that context. However, its removal causes performance drops on the other two datasets, notably a 0.56 PPs loss on DiaKG. This confirms that it remains necessary for handling long-range dependencies.

4. Co-predictor. Removing the MLP predictor causes a substantial 6.50 PPs F1-score drop on DiaKG, whereas removing the biaffine predictor results in a 1.81 PPs decrease. This disparity reveals their asymmetric mechanisms—the MLP serves as the primary classifier operating on the spatially refined grid, while the biaffine branch operates directly on the sequential representations to preserve initial contextual priors.

4.7 Effect of gate in attention fusion

To validate the dynamic gated fusion mechanism, we conduct a comparative experiment against a static simple concatenation baseline (followed by a linear layer). As illustrated in Fig. 9, gated fusion consistently improves the F1-score across all datasets. On CCKS2020, it achieves a 0.60 PPs gain, with increases in both precision (+0.26 PPs) and recall (+0.98 PPs), suggesting more effective feature integration.

However, focusing solely on the F1-score improvements underestimates the true contribution of the gated mechanism. The model's demand for local versus global information dynamically changes: adjacent pairs like "高–血" (hypertension) require local attention, while distant semantic associations like "糖尿病" (diabetes) and "血糖" (blood sugar) demand global attention. Simple concatenation is static, but our dynamic gated fusion mechanism allocates the fusion ratio for each character pair. This enables the model to form a more flexible, context-aware strategy tailored for complex medical text.

4.8 Case study and visual analysis

To further illustrate the capabilities of our D2GI model in handling representative challenges within Chinese medical texts, we select three distinct cases for closer examination, as presented in Fig. 10. The first case, "不同磺脲类药物的分子结构" (molecular structures of different sulfonylurea drugs), might be rare in general corpora. Accurate recognition of such terms benefits from the stable, prior medical semantics provided by the Word2Vec stream. The second case demonstrates RoFormer's capability to capture long-range dependencies within extended medical entities, as it correctly identifies the entire phrase "右上肺见约1.5 cm大小结节影" (a 1.5 cm nodule shadow in the right upper lung) as a single entity, rather than truncating it to the shorter fragment "结节影" (nodule shadow). Finally, the third case presents a classic nested structure, "胸片可见不同程度的梗阻性肺气肿" (obstructive emphysema visible on chest X-ray). Our model addresses this challenge by inheriting the grid-based tagging paradigm, which naturally handles nested medical entities by reformulating the task as relation classification, and further enhances accuracy through our dynamic grid interaction module, which adaptively fuses features to better distinguish complex boundaries within the grid.

To examine the behavior of the gated fusion mechanism, we visualize the gating weight λ in Fig. 11. Higher values indicate greater reliance on the local context, while lower values indicate greater reliance on the global context.

The heatmap indicates a predominance of the global context, as most λ values are below 0.5. This suggests that

Table 3 Ablation study results on three datasets, grouped by the model module

Module	Model	F1 (%)		
		CMeEE-V2	CCKS2020	DiaKG
	D2GI	76.78	89.31	85.40
Dual-stream fusion	w/o Word2Vec	76.09 (−0.69)	89.41 (+0.10)	82.83 (−2.57)
	w/o RoFormer	70.46 (−6.32)	84.92 (−4.39)	77.10 (−8.30)
	w/o BiLSTM fusion	75.13 (−1.65)	88.86 (−0.45)	84.11 (−1.29)
Multi-feature representation	w/o region	76.58 (−0.20)	88.90 (−0.41)	84.94 (−0.46)
	w/o distance	76.56 (−0.22)	88.75 (−0.56)	85.03 (−0.37)
Attention fusion	w/o attention fusion	76.70 (−0.08)	88.34 (−0.97)	84.76 (−0.64)
	w/o local attention	76.50 (−0.28)	88.85 (−0.46)	84.17 (−1.23)
	w/o global attention	76.89 (+0.11)	88.96 (−0.35)	84.84 (−0.56)
Co-predictor	w/o MLP	75.96 (−0.82)	87.69 (−1.62)	78.90 (−6.50)
	w/o biaffine	76.50 (−0.28)	89.20 (−0.11)	83.59 (−1.81)

"w/o" indicates without. Values in parentheses denote the F1 score differences (in PPs) relative to D2GI

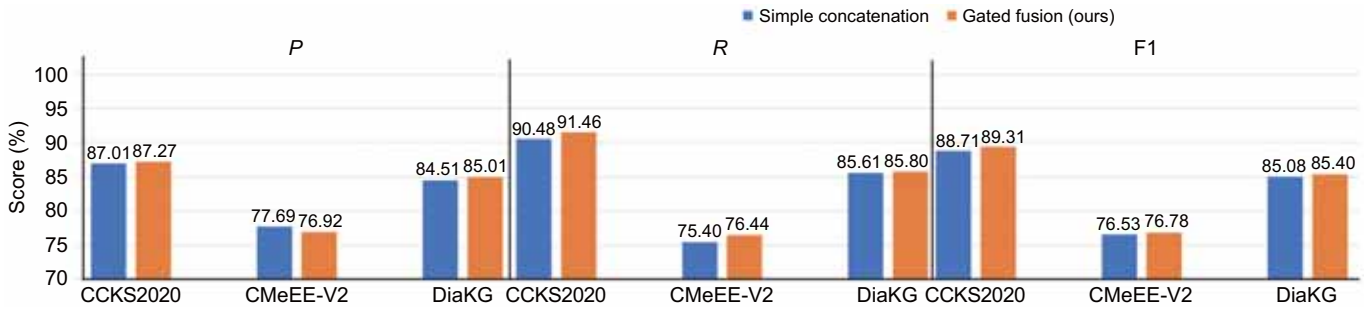

Fig. 9 Performance comparison between simple concatenation and gated fusion on *P*, *R*, and F1 across three datasets

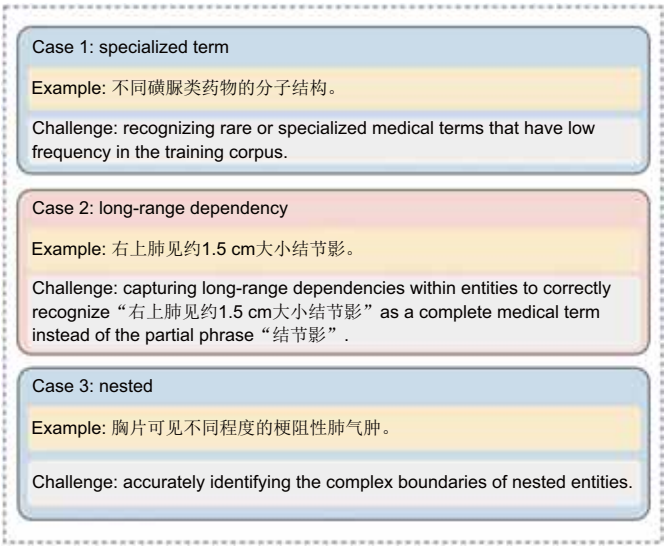

Fig. 10 Examples illustrating CMNER challenges addressed by D2GI

modeling complex and hierarchical medical relations relies more heavily on the global contextual information. At the same time, values along the diagonal are slightly higher, indicating increased reliance on the local context when evaluating adjacent character pairs. In addition to this global–local balance, the heatmap reveals fine-grained, linguistically meaningful patterns. In particular, the matrix exhibits a clear asymmetry: the lower triangular region is consistently darker than the upper triangular region. This suggests that, even under a globally dominant setting, the model assigns higher local weights when incorporating the preceding context. Such asymmetric weighting reflects the directional nature of Chinese medical text, where semantic interpretation often depends on the preceding tokens. Overall, this adaptive balance between global and local contexts, together with directional sensitivity, contributes to improved modeling of complex medical entities.

5 Conclusions

In this paper, we propose the D2GI model for CMNER. The dual-stream fusion architecture integrates RoFormer's long-range dependency modeling with Word2Vec's static semantic priors, establishing robust initial representations. Additionally, the dynamic grid interaction module uses gated attention to adaptively fuse local and global contexts, effectively resolving overlapping nested boundaries. Multiple experiments on three public datasets demonstrate that D2GI is superior to state-of-the-art baselines, confirming its effectiveness and robustness.

Future work will focus on Chinese-specific characteristics to enhance domain adaptability. This includes incorporating radical and glyph embeddings for sub-character semantics and exploring explicit boundary modeling (e.g., latent segmentation and lexicon-aware attention) to address missing word delimiters. We also plan to develop linguistically informed decoding and investigate continued pretraining on large-scale

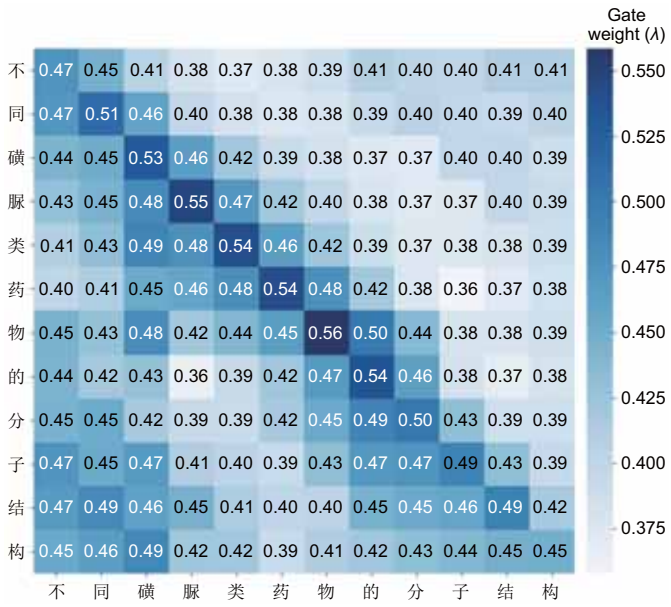

Fig. 11 Heatmap of the dynamic gate weights (λ)

medical corpora to better capture domain-specific patterns.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 61972336 and 62573379) and the Zhejiang Provincial Natural Science Foundation (No. LY23F020001).

Author contributions

Bing LI proposed the methodology and reviewed, revised, and finalized the paper. Zhanming GONG conducted the experiments, analyzed the results, and drafted the paper. Zhiqiang ZHANG and Haiyu SONG provided software support and resources. Yuankang SUN is responsible for project administration.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The CMeEE-V2, DiaKG, and CCKS2020 datasets are publicly online. The other data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration on the use of generative AI tools

During the preparation of this paper, the authors used ChatGPT for language polishing and improving readability. All scientific content, including the research design, theoretical analysis, experimental results, and conclusions, was developed and verified by the authors. The authors reviewed and edited the paper as necessary and take full responsibility for the content of the published article.

References

Albahli S, 2025. An advanced natural language processing framework for Arabic named entity recognition: a novel approach to handling morphological richness and nested entities. *Appl Sci*, 15(6):3073. https://doi.org/10.3390/app15063073

Chang DJ, Chen MS, Liu CZ, et al., 2021. DiaKG: an annotated diabetes dataset for medical knowledge graph construction. Proc 6th China Conf on Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction, p.308-314. https://doi.org/10.1007/978-981-16-6471-7_26

Chen YX, Dan ZP, Gao Z, et al., 2024. Chinese medical continual named entity recognition based on continual learning and knowledge distillation. Proc IEEE Int Conf on High Performance Computing and Communications, p.885-893. https://doi.org/10.1109/HPCC64274.2024.00121

Dai ZJ, Wang XT, Ni P, et al., 2019. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. Proc 12th Int Congress on Image and Signal Processing BioMedical Engineering and Informatics, p.1-5. https://doi.org/10.1109/CISP-BMEI48845.2019.8965823

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional Transformers for language understanding. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), p.4171-4186. https://doi.org/10.18653/v1/N19-1423

Gu YJ, Qu XY, Wang ZF, et al., 2022. Delving deep into regularity: a simple but effective method for Chinese named entity recognition. Findings of the Association for Computational Linguistics, p.1863-1873. https://doi.org/10.18653/v1/2022.findings-naacl.143

Hou SX, Qian YR, Chen JY, et al., 2025. HiNER: hierarchical feature fusion for Chinese named entity recognition. *Neurocomputing*, 611:128667. https://doi.org/10.1016/j.neucom.2024.128667

Ji XY, Chen LN, Gao H, et al., 2025. A high-precision generality method for Chinese nested named entity recognition. Proc 18th Int Conf on Wireless Artificial Intelligent Computing Systems and Applications, p.290-301. https://doi.org/10.1007/978-3-031-71470-2_24

Jia HT, Huang J, Zhao K, et al., 2024. Demonstration-based and attention-enhanced grid-tagging network for mention recognition. *Electronics*, 13(2):261. https://doi.org/10.3390/electronics13020261

Jin ZG, He XY, Wu XD, et al., 2022. A hybrid Transformer approach for Chinese NER with features augmentation. *Expert Syst Appl*, 209:118385. https://doi.org/10.1016/j.eswa.2022.118385

Li JY, Fei H, Liu J, et al., 2022. Unified named entity recognition as word–word relation classification. Proc 36th AAAI Conf on Artificial Intelligence, p.10965-10973. https://doi.org/10.1609/aaai.v36i10.21344

Li XN, Yan H, Qiu XP, et al., 2020. FLAT: Chinese NER using flat-lattice Transformer. Proc 58th Annual Meeting of the Association for Computational Linguistics, p.6836-6842. https://doi.org/10.18653/v1/2020.acl-main.611

Liu P, Guo YM, Wang FL, et al., 2022. Chinese named entity recognition: the state of the art. *Neurocomputing*, 473:37-53. https://doi.org/10.1016/j.neucom.2021.10.101

Liu RB, Wei J, Jia CY, et al., 2021. Modulating language models with emotions. Findings of the Association for Computational Linguistics, p.4332-4339. https://doi.org/10.18653/v1/2021.findings-acl.379

Liu T, Gao J, Ni WJ, et al., 2023. A multi-granularity word fusion method for Chinese NER. *Appl Sci*, 13(5):2789. https://doi.org/10.3390/app13052789

Ma C, Huang WK, 2022. Named entity recognition and event extraction in Chinese electronic medical records. Proc 6th China Conf on Knowledge Graph and Semantic Computing, p.133-138. https://doi.org/10.1007/978-981-19-0713-5_15

Méndez M, Merayo MG, Núñez M, 2023. Long-term traffic flow forecasting using a hybrid CNN-BiLSTM model. *Eng Appl Artif Intell*, 121:106041. https://doi.org/10.1016/j.engappai.2023.106041

Shen YL, Song KT, Tan X, et al., 2023. DiffusionNER: boundary diffusion for named entity recognition. Proc 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), p.3875-3890. https://doi.org/10.18653/v1/2023.acl-long.215

Su JL, Murtadha A, Pan SF, et al., 2022. Global pointer: novel efficient span-based approach for named entity recognition. https://doi.org/10.48550/arXiv.2208.03054

Su JL, Ahmed M, Lu Y, et al., 2024. RoFormer: enhanced Transformer with rotary position embedding. *Neurocomputing*, 568:127063. https://doi.org/10.1016/j.neucom.2023.127063

Wang J, Shou L, Chen K, et al., 2020. Pyramid: a layered model for nested named entity recognition. Proc 58th Annual Meeting of the Association for Computational Linguistics, p.5918-5928. https://doi.org/10.18653/v1/2020.acl-main.525

Wang Y, Wei Q, Yu H, et al., 2025. Cross-interaction of Chinese characters structures and boundary features for improving clinical named entity recognition. *IEEE J Biomed Health Inform*, 29(11):8508-8521. https://doi.org/10.1109/JBHI.2025.3589381

Weigang L, Brom PC, 2025. LLM-BT-Terms: back-translation as a framework for terminology standardization and dynamic semantic embedding. https://doi.org/10.48550/arXiv.2506.08174

Wu S, Song XN, Feng ZH, 2021. MECT: multi-metadata embedding based cross-Transformer for Chinese named entity recognition. Proc 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (Volume 1: Long Papers), p.1529-1539. https://doi.org/10.18653/v1/2021.acl-long.121

Xiao YL, Ji ZC, Li JQ, et al., 2024. MVT: Chinese NER using multi-view Transformer. *IEEE/ACM Trans Audio Speech Lang Process*, 32:3656-3668. https://doi.org/10.1109/TASLP.2024.3426287

Xu L, Li S, Wang YC, et al., 2021. Named entity recognition of BERT-BiLSTM-CRF combined with self-attention. Proc 18th Int Conf on Web Information Systems and Applications, p.556-564. https://doi.org/10.1007/978-3-030-87571-8_48

Xu X, Ma K, 2025. Named entity recognition for Chinese electronic medical records by integrating knowledge graph and Clinical-BERT. *Front Artif Intell*, 8:1634774. https://doi.org/10.3389/frai.2025.1634774

Yan H, Sun Y, Li XN, et al., 2023. An embarrassingly easy but strong baseline for nested named entity recognition. Proc 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), p.1442-1452. https://doi.org/10.18653/v1/2023.acl-short.123

Yan Y, Kang YF, Huang WB, et al., 2025. Chinese medical named entity recognition utilizing entity association and gate context awareness. *PLoS ONE*, 20(2):e0319056. https://doi.org/10.1371/journal.pone.0319056

Yang LY, Xing SL, Mao GJ, 2025. MISS: multiple information span scoring for Chinese named entity recognition. *Comput Speech Lang*, 92:101783. https://doi.org/10.1016/j.csl.2025.101783

Yang XM, Gao ZH, Li YM, et al., 2018. Bidirectional LSTM-CRF for biomedical named entity recognition. Proc 14th Int Conf on Natural Computation, Fuzzy Systems and Knowledge Discovery, p.239-242. https://doi.org/10.1109/FSKD.2018.8687117

Ye YX, Teufel S, 2021. End-to-end argument mining as biaffine dependency parsing. Proc 16th Conf of the European Chapter of the Association for Computational Linguistics (Main Volume), p.669-678. https://doi.org/10.18653/v1/2021.eacl-main.55

Yu JT, Bohnet B, Poesio M, 2020. Named entity recognition as dependency parsing. Proc 58th Annual Meeting of the Association for Computational Linguistics, p.6470-6476. https://doi.org/10.18653/v1/2020.acl-main.577

Zhang GS, Gao ML, Li QL, et al., 2024. Multi-modal generative DeepFake detection via visual-language pretraining with gate fusion for cognitive computation. *Cogn Comput*, 16(6):2953-2966. https://doi.org/10.1007/s12559-024-10316-x

Zhang HY, Lyu L, Chang WF, et al., 2025. A Chinese medical named entity recognition method considering length diversity of entities. *Eng Appl Artif Intell*, 150:110649. https://doi.org/10.1016/j.engappai.2025.110649

Zhang L, Xia PF, Ma XX, et al., 2024. Enhanced Chinese named entity recognition with multi-granularity BERT adapter and efficient global pointer. *Compl Intell Syst*, 10(3):4473-4491. https://doi.org/10.1007/s40747-024-01383-6

Zhang NY, Chen MS, Bi Z, et al., 2022. CBLUE: a Chinese biomedical language understanding evaluation benchmark. Proc 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), p.7888-7915. https://doi.org/10.18653/v1/2022.acl-long.544

Zhu JY, Ni P, Li YM, et al., 2019. An Word2Vec based on Chinese medical knowledge. Proc IEEE Int Conf on Big Data, p.6263-6265. https://doi.org/10.1109/BigData47090.2019.9005510