Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

CLART: A cascaded lattice-and-radical transformer network for Chinese medical named entity recognition

Yinlong Xiao [a,1], Zongcheng Ji [b,1], Jianqiang Li [a], Qing Zhu [a,*]

[a] *Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*
[b] *PAII Inc., CA 94087, United States of America*

ARTICLE INFO

Dataset link: https://tianchi.aliyun.com/dataset/86819

Dataset link: https://www.biendata.xyz/competition/CCKS2017_2/

Keywords:
Chinese medical named entity recognition
Lattice structure
Radical information
Attention mechanism
Transformer

ABSTRACT

Chinese medical named entity recognition (NER) is a fundamental task in Chinese medical natural language processing, aiming to recognize Chinese medical entities within unstructured medical texts. However, it poses significant challenges mainly due to the extensive usage of medical terms in Chinese medical texts. Although previous studies have made attempts to incorporate lexical or radical knowledge in order to improve the comprehension of medical texts, these studies either focus solely on one of these aspects or utilize a basic concatenation operation to combine these features, which fails to fully utilize the potential of lexical and radical knowledge. In this paper, we propose a novel Cascaded LAttice-and-Radical Transformer (CLART) network to exploit both lexical and radical information for Chinese medical NER. Specifically, given a sentence, a medical lexicon, and a radical dictionary, we first construct a flat lattice (*i.e.*, character-word sequence) for the sentence and radical components of each Chinese character through word matching and radical parsing, respectively. We then employ a lattice Transformer module to capture the dense interactions between characters and matched words, facilitating the enhanced utilization of lexical knowledge. Subsequently, we design a radical Transformer module to model the dense interactions between the lattice and radical features, facilitating better fusion of the lexical and radical knowledge. Finally, we feed the updated lattice-and-radical-aware character representations into a Conditional Random Fields (CRF) decoder to obtain the predicted labels. Experimental results conducted on two publicly available Chinese medical NER datasets show the effectiveness of the proposed method.

1. Introduction

Chinese medical NER is a crucial task in Chinese medical text analysis [16], which aims to extract predefined types of entities from unstructured medical texts. These texts may come from different sources, such as Chinese electronic medical records (EMRs) and traditional Chinese medicine (TCM) manuals, each with its own predefined types. For example, predefined types in EMRs may include “疾病 (Disease)”, “治疗 (Treatment)”, and “检查 (Lab Test)”, while those in TCM manuals may include “药品 (Drug)”, “药物成分 (Drug Ingredient)”, and “药物性味 (Drug Taste)”. Due to the broad range of applications associated with the extracted medical

[*] Corresponding author.
E-mail address: ccgszq@bjut.edu.cn (Q. Zhu).

[1] These authors contributed equally to this work. Work was done when Yinlong Xiao was interning at PAII Inc.

https://doi.org/10.1016/j.heliyon.2023.e20692
Received 1 September 2023; Received in revised form 1 October 2023; Accepted 4 October 2023

Available online 10 October 2023

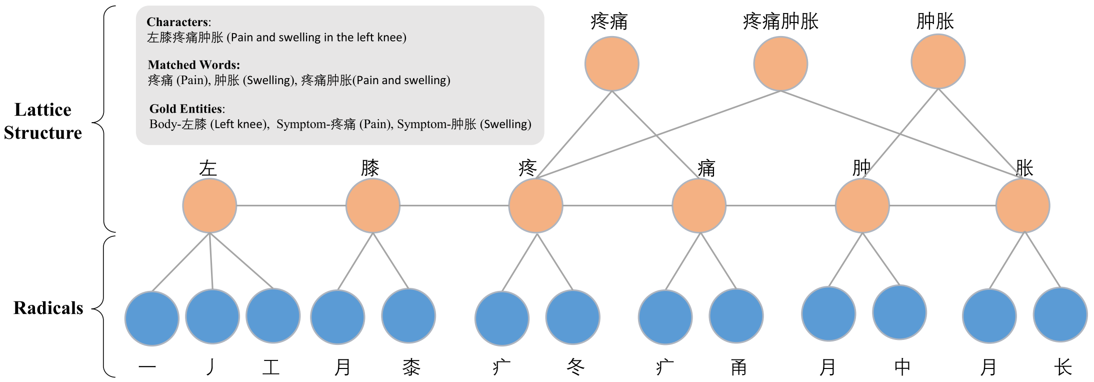

Lattice Structure
Radicals
Characters:
左膝疼痛肿胀 (Pain and swelling in the left knee)
Matched Words:
疼痛 (Pain), 肿胀 (Swelling), 疼痛肿胀(Pain and swelling)
Gold Entities:
Body-左膝 (Left knee), Symptom-疼痛 (Pain), Symptom-肿胀 (Swelling)
疼痛
疼痛肿胀
肿胀
左
膝
疼
痛
肿
胀
一
丿
工
月
桼
疒
冬
疒
甬
月
中
月
长

Fig. 1. An example to illustrate the character-word lattice structure and the radicals of each Chinese character. We do not translate the radicals since some of the radicals only represent the pronunciation of the character (e.g. the radical "中" of the character "肿").

entities, such as medical knowledge graph construction [24], medical relation extraction [38], and medical question answering [15], Chinese medical NER has attracted considerable attention [40,18,27,34]. Although Chinese NER has shown promising results in the general domain [39,28,17,10], it encounters significant challenges when applied in the medical domain. This is primarily due to the extensive range of medical terms involved in Chinese medical NER, which usually requires a deep understanding of medical texts to ensure accurate recognition [40].

A widely adopted approach to enhance the recognition of medical terms from unstructured medical texts is to introduce medical knowledge with a lexicon [20,32,40,34]. Earlier approaches [20,32] employ the bidirectional maximum matching (BDMM) algorithm [5] to obtain matched lexical words in a sentence and then convert these matched words as features for the NER model. However, matched words acquired by the BDMM algorithm may be inaccurate. For instance, in Fig. 1, if the BDMM algorithm is used to match the sentence with the medical lexicon, only the word "疼痛肿胀 (Pain and swelling)" will be matched. However, the correct entities are "疼痛 (Pain)" and "肿胀 (Swelling)". The wrong matched word "疼痛肿胀 (Pain and swelling)" may affect the prediction. To address the issue of incorrect matched words, LatticeLSTM [39,40] is proposed to integrate the lexical information with a character-word lattice structure, which is a graph composed of characters and all matched words, as illustrated in Fig. 1. The lattice structure contains three matched words, including "疼痛 (Pain)", "疼痛肿胀 (Pain and swelling)", and "肿胀 (Swelling)". The model dynamically calculates the weight of each matched word to mitigate issues caused by incorrect matching. Besides, MKRGCN [34] is also proposed to leverage the lexical information from both the lexicon and knowledge graph. It models a relational graph constructed from characters and all matched words using a relational graph convolutional network [26]. To optimize the efficiency, Li et al. [17] design a flat-lattice Transformer (FLAT), which flattens the lattice into a character-word sequence (*i.e.,* token sequence[2]). The self-attention mechanism [30,2] is employed in FLAT to model the lattice features. Although FLAT achieves state-of-the-art performance in terms of both effectiveness and efficiency in the general domain, this method does not take into account the radical information that is essential for medical entity recognition.

Recently, radicals of each Chinese character are also exploited to enhance the recognition of medical terms from unstructured medical texts, especially helping the identification of entity types [18,27]. For instance, in Fig. 1, the character "膝 (Knee)" contains two radicals "月" and "桼". The radical "月" is often associated with entities with "body" type, which helps to identify the "body" type of the entity "左膝 (Left knee)". Therefore, the effective incorporation of radical information is crucial for the task. A straightforward solution is to integrate the radical features in the embedding layer [18,27] through a concatenation operation. However, this only achieves token-level fusion and does not model the dense interactions between radical and lattice features. We argue that the dense interactions between radical and lattice features can capture global information, benefiting the prediction of the entity type. For example, although the character "肿 (Swelling)" contains the radical "月" (typically related to body parts), "肿胀 (Swelling)" refers to a symptom. For "肿 (Swelling)", being able to interact with the radical "疒" (usually associated with symptoms and diseases) contained in the nearby characters "疼 (Pain)" and "痛 (Pain)" would facilitate the accurate identification of "肿胀 (Swelling)" as a symptom entity.

In this paper, we propose a novel Cascaded LAttice-and-Radical Transformer (CLART) network to exploit both lexical and radical information for Chinese medical NER. CLART takes advantage of FLAT [17] in modeling lattice features and further extends FLAT by incorporating radical features, which play a guiding role in medical texts. Unlike the straightforward integration of both lattice and radical features by concatenation operation [18,27], CLART employs a radical Transformer to effectively model the dense interactions between the lattice and radical features, thus facilitating better fusion of the lexical and radical knowledge. Specifically, given a sentence, a medical lexicon, and a radical dictionary, we construct a flat lattice (*i.e.,* character-word sequence or token sequence) for the sentence and radical components of each Chinese character through word matching and radical parsing, respectively. Then, we convert the flat lattice, radical components, and relative positions to the corresponding embeddings using the embedding layer. Next, we feed the lattice embeddings and relative position embeddings into the lattice Transformer module to extract the lattice features.

[2] Following [17], we use the term "token" to represent a character in a sentence or a matched word in a lexicon.

Subsequently, we feed the lattice features, radical-level embeddings, and relative position embeddings into the radical Transformer module to fuse the radical and lattice features. Finally, we feed the updated lattice-and-radical-aware character representations into a conditional random fields (CRF) decoder to predict labels for each Chinese character.

Our contributions can be summarized as follows:

- We propose a novel CLART network that cascades a lattice Transformer module and a radical Transformer module for Chinese medical NER. The network aims to leverage both lexical and radical knowledge effectively in order to achieve a deep understanding of medical texts, thereby ensuring accurate recognition of medical entities.
- We employ a lattice Transformer module to effectively capture the dense interactions between characters and matched words, thereby facilitating the enhanced utilization of lexical knowledge from a medical lexicon.
- We design a radical Transformer module to effectively model the dense interactions between the lattice and radical features, thus facilitating better fusion of the lexical and radical knowledge.
- We conduct extensive experiments on two public Chinese medical NER datasets. The experimental results show that CLART significantly outperforms previous methods, thus highlighting the effectiveness of the proposed approach. We will release our code at https://github.com/starlight0818/clart to facilitate further research in this field.

2. Related work

Chinese medical NER has received widespread attention among researchers, as it can extract valuable structured information from unstructured medical text, assisting doctors in efficiently analyzing medical documents. This task is typically defined as a sequence labeling task, and traditional methods often utilize statistical machine learning methods [37,21,22]. Although these methods exhibit good generalization performance, they require manual feature engineering, which entails high labor costs.

Recently, deep learning methods have experienced rapid development. Unsupervised pre-trained language models [23,25,3] are used to obtain text representations, and different neural networks [9,12,30] are designed to automatically extract text features, reducing manual efforts. Consequently, many studies have applied deep learning methods to the Chinese medical NER task. Among them, the LSTM-based models [14,19,29,33], which use LSTM [9] as the encoder to capture text features and CRF as the decoder to predict labels, are the most widely used methods. However, the recurrent structure of LSTM hinders the full utilization of GPU parallel processing capabilities, thus limiting the speed of its training and inference. To enhance the overall efficiency, Transformer-based models [7,35,31] have been employed for the NER task. Transformer-based models can fully exploit the parallel computing capabilities of GPUs, leading to substantial improvements in efficiency. Moreover, their ability to model long-distance dependencies brings an additional performance boost. In this work, we also leverage the advantages of the Transformer architecture and design a novel Transformer-based network for the Chinese medical NER task.

Although the aforementioned models have demonstrated promising results in the general domain [39,28,17,10], Chinese NER faces significant challenges when applied in the medical domain. This can be attributed primarily to the extensive range of medical terms involved, which requires a profound understanding of medical knowledge for accurate recognition [40]. To this end, researchers have attempted to introduce external medical knowledge to enhance the recognition of medical terms from unstructured medical texts [39,17,4]. A commonly employed strategy is to incorporate medical knowledge through the use of a lexicon [20,32,40,34]. Earlier research efforts [20,32] leverage the BDMM algorithm [5] to extract matched lexical words from a sentence. These matched words are subsequently transformed into features for the NER model. In order to address the issue of incorrect matched words introduced by the BDMM algorithm, AT-LatticeLSTM [40] employs LatticeLSTM [39] encoder to dynamically integrate the lexical information. Additionally, adversarial training [6] is employed to enhance the model's robustness. Besides, MKRGCN [34] is introduced to leverage the lexical information derived from both the lexicon and knowledge graph. It constructs a relational graph consisting of characters and all matched words. This graph is then utilized by employing a relational graph convolutional network [26] to effectively model the relationships within the graph. Recently, there has been a growing focus on leveraging the radicals associated with each Chinese character to enhance the recognition of medical terms from unstructured medical texts, especially helping the identification of entity types [18,27]. Li et al. [18] integrate the radical information in the embedding layer, encode the character sequence with LSTM, and exploit the lexical information in a post-processing way. MLSFN [27] fuses both the radical and lexical information through concatenation in the embedding layer and also encodes the character sequence with LSTM. Besides, MLSFN also considers phonetic and syntactic information. In this work, we consider both the lexical and radical knowledge and design a cascaded Transformer network to better integrate these two types of semantic information.

3. Proposed method

Fig. 2 shows the overall architecture of the proposed method for Chinese medical NER, which consists of four main components, *i.e.,* lattice and radical construction, embedding layer, CLART encoder, and CRF decoder. Given a sentence, a medical lexicon, and a radical dictionary, we construct a flat lattice (*i.e.,* character-word sequence or token sequence) for the sentence and radical components of each Chinese character through word matching and radical parsing. Then, we convert the flat lattice, radical components, and relative positions to the corresponding embeddings using the embedding layer. Next, we encode the sentence by passing the lattice embeddings, radical-level embeddings, and relative position embeddings to the CLART encoder to integrate the lattice and radical features. Finally, we feed the updated lattice-and-radical-aware character representations into the CRF decoder to obtain the prediction results.

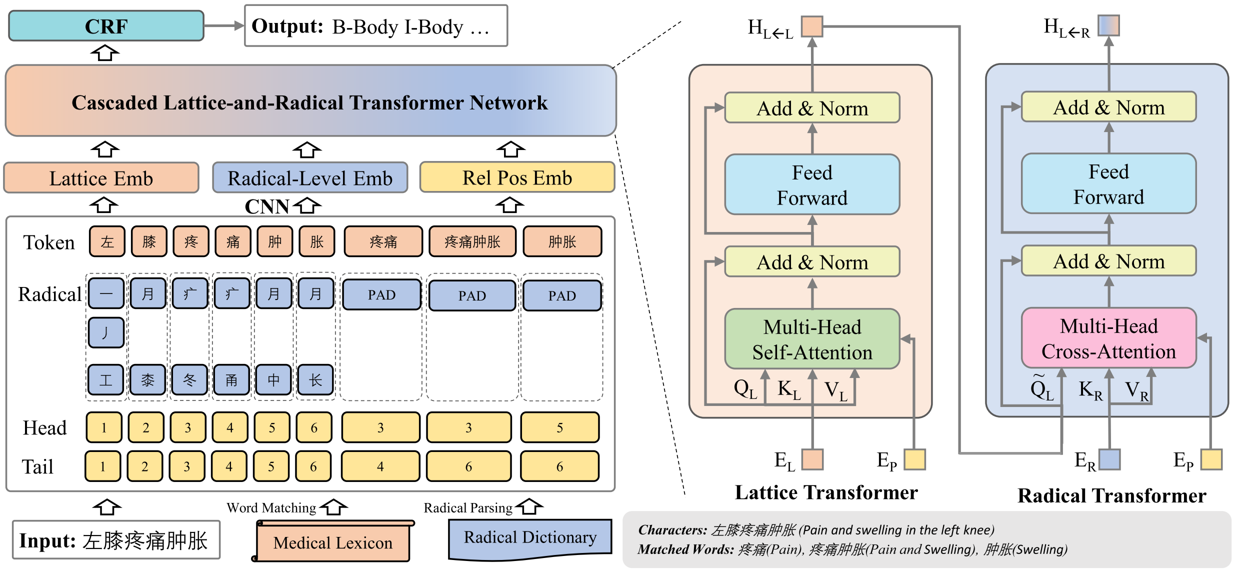


Fig. 2. The overall architecture of the proposed method for Chinese medical NER. The left part illustrates the pipeline of the architecture, and the right part shows the detail of the CLART network. "Rel Pos" and "Emb" are short for relative position and embeddings, respectively.

3.1. Lattice and radical construction

3.1.1. Lattice construction

A medical lexicon usually contains a large number of medical terms, such as "肿胀 (swelling)" and "疼痛肿胀 (Pain and swelling)", which can provide prior knowledge for Chinese medical NER. Traditional methods integrate lexical information via word segmentation, which may suffer the error propagation problem [39], as mentioned in Section 1. To alleviate this issue, following LatticeLSTM [39] and FLAT [17], we introduce the lattice structure, which can integrate the lexical information from all matched words. Specifically, given a sentence (*i.e.,* character sequence) and a medical lexicon, we get a set of matched words by matching the sentence with the lexicon. Then, we can obtain a lattice, which is a graph structure consisting of characters, and all matched words, as shown in Fig. 1.

As the Transformer architecture cannot directly model graph-structured data, following FLAT [17], we convert the graph-structured lattice to a flat lattice, as shown in Fig. 2. We first flatten the lattice by appending the matched words to the end of the character sequence. We use the term "token" to represent a character or a word. For example, we append all matched words, including "疼痛 (Pain)", "疼痛肿胀 (Pain and swelling)", "肿胀 (Swelling)", to the end of the character sequence "左膝疼痛肿胀 (Pain and swelling in the left knee)". We then assign head and tail positions to each token. The head and tail positions are the position of the first and last characters in the token, respectively. If a token comprises only one character, the head and tail positions are the same. For example, since the positions of the first character "疼 (Pain)" and the last character "痛 (Pain)" in the matched word "疼痛 (Pain)" are "3" and "6", we assign head and tail positions to the token "疼痛 (Pain)" as "3" and "6", respectively.

3.1.2. Radical construction

Unlike English, radicals are unique features in Chinese characters and contain rich semantic information. They play an important role in understanding the meaning of Chinese characters [18]. For example, the radical "疒" is usually associated with symptoms and diseases, which can provide guidance for the model. By incorporating radical information, we can leverage this unique linguistic feature to boost the performance on the Chinese medical NER task. Specifically, given a sentence and a radical dictionary, we get radical components of each Chinese character by parsing each Chinese character in the sentence with the radical dictionary. As shown in Fig. 2, we get the radical component "疒" and "冬" of the Chinese character "疼 (Pain)" by parsing the character "疼 (Pain)" with the radical dictionary.

3.2. Embedding layer

3.2.1. Lattice embeddings

Since a lattice consists of characters and matched words, we first initialize character embeddings and word embeddings, and then obtain the lattice embeddings based on the character embeddings and word embeddings. For character embeddings, we map each character to its corresponding character embeddings using pre-trained character vectors. The pre-trained character vectors can be either static vectors, such as Word2Vec [23], or dynamic vectors, such as BERT [3]. For word embeddings, we randomly initialize the embeddings of the matched words. We then employ a linear layer to transform the dimensions of the character and word embeddings into d_{model}. Finally, we concatenate the transformed character and word embeddings to obtain the lattice embeddings E_L.

3.2.2. Radical-level embeddings

We randomly initialize the embeddings for each radical component. We then utilize a CNN [12] network to aggregate the features of multiple radical components contained within the character (*e.g.,* the character "左 (Left)" contains three radical components: "一, 丿, 工"), thereby obtaining the final aggregated radical-level embeddings for the given character. We use E_R to represent the aggregated radical-level embeddings.

3.2.3. Relative position embeddings

Following FLAT [17], we represent the relative position information of tokens by calculating the distance between their head and tail positions. Let $head_i$ and $tail_i$ donate the head and tail position of the token indexed at i. We can calculate the relative distances d_{ij}^{hh} and d_{ij}^{tt} between tokens indexed at i and j in the lattice. The formulas are shown in Eq. (1) and Eq. (2):

$$d_{ij}^{hh} = head_i - head_j \tag{1}$$

$$d_{ij}^{tt} = tail_i - tail_j \tag{2}$$

We use E_P to represent the relative position embeddings of the token sequence, and we use the following Eq. (3) to calculate the relative position embeddings for the tokens indexed at i and j:

$$E_{P_{ij}} = \mathrm{ReLU}\left(W^P\left(p_{d_{ij}^{hh}} \oplus p_{d_{ij}^{tt}}\right)\right) \tag{3}$$

where W^P is a learnable parameter, $\oplus$ denotes concatenation operation, and p_d is the position encoding calculated with the Eq. (4) and Eq. (5) [30]:

$$p_d^{(2k)} = \sin\left(d/10000^{2k/d_{model}}\right) \tag{4}$$

$$p_d^{(2k+1)} = \cos\left(d/10000^{2k/d_{\mathrm{model}}}\right) \tag{5}$$

where d is d_{ij}^{hh} or d_{ij}^{tt}, and k denotes the index of the dimension of p_d.

The corresponding radical components of each token share the same head and tail positions of the token, and thus share the same relative position embeddings.

3.3. CLART encoder

3.3.1. Lattice transformer

The lattice Transformer aims to integrate lexical information into the character representations via the character-word lattice structure. The self-attention (SA) module in the lattice Transformer models the dense interactions between characters and matched words. We adopt a variant of the self-attention mechanism [2] that incorporates relative positions, as the inclusion of directional information from relative positions yields substantial advantages for the NER task [35].

Given the lattice embeddings E_L and the relative position embeddings E_P, we perform a linear mapping of E_L to obtain the matrices Q_L, K_L, and V_L. Then, we calculate the self-attention with the following Eqs. (6), (7), and (8):

$$\mathrm{SA}_{L\leftarrow L}\left(E_L, E_P\right) = \mathrm{softmax}\left(A_{L\leftarrow L}\right)V_L \tag{6}$$

$$A_{L\leftarrow L,ij} = \left(Q_{L,i} + u_L\right)^{\mathrm{T}} K_{L,j} + \left(Q_{L,i} + v_L\right)^{\mathrm{T}} E_{P_{ij}} W_L^P \tag{7}$$

$$Q_L, K_L, V_L = E_L\left[W_L^Q, W_L^K, W_L^V\right] \tag{8}$$

where W_L^Q, W_L^K, W_L^V, and $W_L^P \in \mathbb{R}^{d_{model}\times d_{head}}$, and $u_L, v_L \in \mathbb{R}^{d_h ead}$ are learnable parameters.

Next, we calculate the multi-head self-attention (MSA) by concatenating the outputs from h heads followed by a linear projection, as shown in Eq. (9):

$$\mathrm{MSA}\left(E_L, E_P\right) = [\mathrm{SA}_1(E_L, E_P), \dots, \mathrm{SA}_h(E_L, E_P)]W_L^O \tag{9}$$

where $W_L^O \in \mathbb{R}^{d_{model}\times d_{head}}$ is a learnable parameter.

The output of MSA will be further processed by residual connection [8] and layer normalization (LayerNorm) [1]. Subsequently, we obtain the intermediate output denoted as Z. We then feed Z into a feed-forward network (FFN) for non-linear transformation. Finally, we apply another round of residual connection and layer normalization to generate the output denoted as $H_{L\leftarrow L}$. The calculation formulas are shown in Eq. (10) and Eq. (11):

$$H_{L\leftarrow L} = \text{ LayerNorm(FFN }(Z) + Z) \tag{10}$$

$$Z = \text{ LayerNorm(MSA }\left(E_L, E_P\right) + E_L) \tag{11}$$

3.3.2. Radical transformer

After the lattice Transformer, lexical information has been integrated into the character representations. We design the radical Transformer to further integrate radical information into the character representations. The radical Transformer models dense interactions between the lattice and radical features through a cross-attention module. Following the self-attention used in the lattice Transformer, we also adopt a variant of the cross-attention that incorporates the relative position.

Given the output $H_{L\leftarrow L}$ of the lattice Transformer, the radical-level embeddings E_R and the relative position embeddings E_P, we perform a linear mapping of $H_{L\leftarrow L}$ to obtain the matrix $\widetilde{Q}_L$ and we also map E_R to the matrices K_R and V_R. Then, we compute the cross-attention (CA) with the following Eqs. (12), (13), (14) and (15):

$$\mathrm{CA}_{L\leftarrow R}\left(H_{L\leftarrow L}, E_R, E_P\right) = \mathrm{softmax}\left(A_{L\leftarrow R}\right) V_R \tag{12}$$

$$A_{L\leftarrow R,ij} = \left(\tilde{Q}_{L,i} + u_R\right)^{\mathrm{T}} K_{R,j} + \left(\tilde{Q}_{L,i} + v_R\right)^{\mathrm{T}} E_{P_{ij}} W_R^P \tag{13}$$

$$\tilde{Q}_L = H_{L\leftarrow L} W_{L\leftarrow L}^Q \tag{14}$$

$$K_R, V_R = E_R\left[W_R^K, W_R^V\right] \tag{15}$$

where $W_{L\leftarrow L}^Q$, W_R^K, W_R^V, and $W_R^P \in \mathbb{R}^{d_{model}\times d_{head}}$, and $u_R, v_R \in \mathbb{R}^{d_{head}}$ are learnable parameters.

The subsequent calculations follow the same procedure as the lattice Transformer. Finally, we obtain the updated lattice-and-radical-aware character representations from the radical Transformer denoted as $H_{L\leftarrow R}$.

3.4. CRF decoder

After the Radical Transformer, the character representations are enriched by integrating both lexical and radical information. We apply a CRF [13] decoder to predict the output labels. Given a sentence $X = (x_1, x_2, ..., x_n)$, the probability of the output tag sequence $Y = \{y_1, y_2, ..., y_n\}$ is calculated as Eq. (16):

$$\Pr(Y \mid X) = \frac{\prod_{i=1}^n \varphi\left(y_{i-1}, y_i \mid X\right)}{\sum_{y'\in\mathcal{Y}} \prod_{i=1}^n \varphi\left(y'_{i-1}, y'_i \mid X\right)} \tag{16}$$

where $\mathcal{Y}$ is the set of all possible tags, $\varphi\left(y_{i-1}, y_i, X\right) = \exp(W_{y_{i-1},y_i} H_{L\leftarrow R} + b_{y_{i-1},y_i})$, and W_{y_{i-1},y_i}, b_{y_{i-1},y_i} are trainable parameters. The training loss is calculated as Eq. (17):

$$\mathcal{L} = -\sum_{i=1}^n \log\left(\Pr\left(y_i \mid X\right)\right) \tag{17}$$

4. Experiments

4.1. Experiment settings

4.1.1. Datasets

We evaluated our method on two public Chinese medical datasets of different types. (1) **CCKS2017 dataset.**[3] This dataset comes from real-world Chinese electronic medical records. 1,596 annotated records with 5 entity types are used for evaluation. This dataset is originally split into train and test sets, with 1,198 and 398 records, respectively. Since we need to tune the hyper-parameters of the model with a development set, we randomly select 198 records from the train set as the development set. (2) **TCM2020 dataset.**[4] This dataset comes from traditional Chinese medicine manuals. 700 annotated manuals with 13 entity types are used for evaluation. This dataset is split into train, development, and test sets, with 420, 140, and 140 manuals, respectively.

4.1.2. Baselines and evaluation metrics

As shown in Table 1, we compare our approach with the following three categories (see the "Features" column of the table) of state-of-the-art baseline methods. All these methods employ a CRF decoder to predict labels, with differences in feature selection and encoder design. (1) **Methods without using any lexical and radical information.** LSTM [11] and TENER [35] employ LSTM [9] and Transformer [30] encoders to model text sequence features, respectively. BERT+LSTM [11] and BERT+TENER combine the pre-trained model BERT [3] with LSTM and Transformer, respectively. (2) **Methods that use lexical information.** AT-LatticeLSTM [40] and FLAT [17] use LatticeLSTM [39] and Lattice Transformer [17] encoders to model lattice structure features, respectively. Adversarial training [6] is applied in AT-LatticeLSTM to enhance model stability. BERT+MKRGCN [34] model uses the pre-trained model pre-trained on medical texts, utilizes the lexical information from both the lexicon and the knowledge graph, and models the relational graphs constructed of words and characters via the relational graph convolutional network [26]. BERT+FLAT [17] combines the pre-trained model BERT with FLAT. (3) **Methods that use both lexical and radical information.** MLSFN [27] model considers not only lexical and radical information but also phonetic and syntactic information. For a fair comparison, we select the results from the paper [27] that utilizes only lexical and radical information. MLSFN fuses the radical and lexical information through concatenation in the embedding layer and then encodes the character sequence with LSTM. BERT+ LSTM* [18] uses the pre-trained

[3] https://www.biendata.xyz/competition/CCKS2017_2/.

[4] https://tianchi.aliyun.com/dataset/86819.

Table 1

Comparison of different approaches on CCKS2017 and TCM2020 datasets (%). Models in the first part initialize character embeddings using Word2Vec [23], and those in the second part (named with BERT+X) initialize character embeddings using BERT [3]. The "Features" column lists whether the model utilizes lexical and radical information, and specifies the type of information used. The "Encoder" column lists the encoder the model utilized. $*$ donates a variant of LSTM that integrates radical information in the embeddings layer and exploits the lexical information in a post-processing way. † means the improvements over baselines are statistically significant (p-value $<$ 0.05 based on the paired *t*-test).

Model	Features	Encoder	CCKS2017			TCM2020		
			P	R	F1	P	R	F1
LSTM [11]	-	LSTM	89.53	85.89	87.67	67.29	78.61	72.51
TENER [35]	-	Transformer	89.82	87.16	88.47	68.73	82.02	74.79
AT-LatticeLSTM [40]	lexical	LSTM	88.98	90.28	89.64	-	-	-
FLAT [17]	lexical	Transformer	90.24	90.65	90.44	67.63	**86.81**	76.03
MLSFN [27]	lexical+radical	LSTM	89.28	**91.00**	90.13	-	-	-
CLART (Ours)	lexical+radical	Transformer	**90.70**	90.81	**90.75**†	**69.41**	84.68	**76.39**†
BERT+LSTM [11]	-	LSTM	90.78	90.29	90.53	71.95	83.55	77.31
BERT+TENER [35]	-	Transformer	90.97	90.15	90.56	71.96	83.87	77.46
BERT+MKRGCN [34]	lexical	LSTM+GCN	-	-	91.13	-	-	-
BERT+FLAT [17]	lexical	Transformer	91.43	91.14	91.29	72.52	85.66	78.54
BERT+ LSTM* [18]	lexical+radical	LSTM	**91.76**	90.88	91.32	71.10	**87.19**	78.33
BERT+CLART (Ours)	lexical+radical	Transformer	91.50	**91.81**	**91.66**†	**73.53**	85.21	**78.94**†

model pre-trained on medical texts, integrates the radical information in the embedding layer, encodes the character sequence with LSTM, and exploits the lexical information in a post-processing way.

We use the standard Precision (P), Recall (R), and F1 score (F) metrics to evaluate these methods. The results reported in the paper are obtained through our reproduction of open-source code or from the reported results in the original paper. Specifically, we rerun LSTM and BERT+LSTM with the built-in code of the FastNLP[5] framework while we rerun TENER, FLAT, BERT+TENER and BERT+FLAT with the official open-source code. We rerun BERT+ LSTM* with the official open-source code on TCM2020. We use the reported results on CCKS2017 for AT-LatticeLSTM, MLSFN, BERT+MKRGCN and BERT+ LSTM*.

4.1.3. *Implementation details*

We implement the proposed CLART model based on the FastNLP framework. The implementation details are as follows: (1) **External resources**. For CLART, We use the Word2Vec [23] released by Yang et al. [36] to obtain character embeddings, and for BERT+CLART, we use the BERT [3] from FastNLP with the "cn-wwm-ext" version to obtain character embeddings. We use the entities in CMEKG,[6] which is a Chinese medical knowledge graph, as the lexicon. We use a public radical dictionary[7] to parse radicals for each Chinese character. Word embeddings and radical component embeddings are randomly initialized. All the character, word, and radical component embeddings are trainable during fine-tuning. (2) **Preprocess**. We convert the annotated data to the widely-adopted BIO format. For example, we convert the annotation for an entity "左膝 (Left knee)" with "Body" type to "B-Body" and "I-Body" for the characters "左" and "膝", respectively. We split the sentence by punctuation marks ".", "。", ";" if it contains more than 200 characters. (3) **Hyper-parameters**. We set the batch size to 10, the number of epochs to 50, and the learning rate to 0.001. We use SGD as the optimizer. For CCKS2017, we set the head number of the Transformer to 6, the head dimension to 20, and the dropout rate of the linear layer to 0.1. For TCM2020, we set the head number to 8, the head dimension to 20, and the dropout rate of the linear layer to 0.15. We conduct all experiments with RTX 8000 GPUs.

4.2. *Results and discussion*

4.2.1. *Overall results*

We compare our approach with baselines in three categories (see the "Features" column of Table 1): methods without using any lexical and radical information, methods that use lexical information, and methods that use both lexical and radical information. Besides, We also compare the models which initialize character embeddings using Word2Vec [23] and BERT [3], respectively. We run our proposed model 5 times with different seeds, and we report the average scores to ensure robust results.

The results without BERT embeddings are shown in the first part of Table 1. We have the following findings. (1) Models that use lexical information outperform the models without the lexicon (AT-LatticeLSTM vs. LSTM, FLAT vs. TENER), indicating the effectiveness of incorporating the lexical information for the different Chinese medical NER models. (2) MLSFN, which incorporates both lexical and radical information, outperforms the lexicon-enhanced model AT-LatticeLSTM. This indicates that, in addition to lexical information, radical information plays a significant role in enhancing performance. Additionally, it is worth noting that FLAT, which only uses lexical information, outperforms MLSFN. A possible reason is that FLAT, utilizing the Transformer architecture, can better integrate lexical information compared to MLSFN, which employs an LSTM encoder. (3) Our proposed model CLART outperforms all

[5] https://github.com/fastnlp/fastNLP.

[6] https://tianchi.aliyun.com/dataset/81506.

[7] https://github.com/kfcd/chaizi.

Table 2
Performance of different variants (%). LT and RT are short for lattice Transformer, and radical Transformer, respectively.

Model	CCKS2017			TCM2020		
	P	R	F1	P	R	F1
BERT+CLART	91.50	91.81	**91.66**	73.53	85.21	78.94
w/o RT	91.43	91.14	91.29	72.52	85.66	78.54
w/o LT	91.67	90.98	91.32	72.36	85.45	78.36
w/o LT and RT	90.39	91.23	90.80	70.87	84.76	77.20
Variant A	91.15	**92.17**	**91.66**	**73.73**	85.11	**79.01**
Variant B	**91.92**	90.84	91.37	72.36	**85.81**	78.51

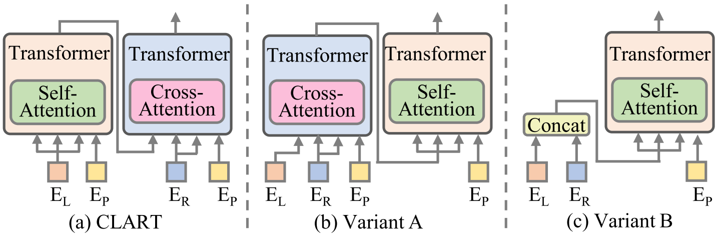


Fig. 3. (a) A simplified version of the proposed CLART model. (b) Variant A is a model to swap the order of the lattice and radical Transformer modules. (c) Variant B is a model to fuse the lattice and radical features by a basic concatenation operation.

baselines on both datasets, indicating that our model can more effectively integrate both lexical and radical information through the cascaded Transformer network.

The results with BERT embeddings are shown in the second part of Table 1. We have the following findings. (1) Models that use lexical information and BERT embeddings (BERT+MKRGCN and BERT+FLAT) outperform the model without the lexicon (BERT+LSTM and BERT+TENER). Besides, BERT+ LSTM*, which combines lexical information, radical information, and BERT embeddings, outperforms the lexicon-enhanced model BERT+MKRGCN. These comparisons indicate that the incorporation of lexical information and radical information still has a good complementary effect on the powerful character representation of BERT. (2) Although BERT+FLAT based on Transformer architecture does not utilize radical features, it still achieves comparable performance to BERT+ LSTM*, which illustrates the superiority of Transformer architecture. (3) Our proposed model BERT+CLART outperforms all baselines on both datasets, indicating that integrating both lexical and radical information with CLART is an effective way to boost the performance of BERT for Chinese medical NER.

4.2.2. Ablation study

To validate the effectiveness of the main components in the proposed method, we design ablation studies by removing different components from BERT+CLART as shown in the first part of Table 2. “w/o RT” means to remove the radical Transformer while keeping the lattice Transformer. In this case, the model degenerates to FLAT [17]. “w/o LT” means to remove the lattice Transformer while keeping the radical Transformer. “w/o LT and RT” means to remove both the lattice Transformer and radical Transformer. In this case, the model degrades to BERT+CRF. We have the following findings from the comparisons. When either the lattice Transformer or the radical Transformer is removed, there is a substantial drop in the F1 score. However, when both are removed, the decrease in the F1 score becomes even more significant. This clearly demonstrates the critical importance of both lexical and radical information in the Chinese medical NER task.

To evaluate the impact of different orders of the lattice and radical Transformer modules (*i.e.,* first lattice then radical vs. first radical then lattice) and the impact of different fusion methods (*i.e.,* concatenation vs. cross-attention) of the lattice and radical features, we further design two variants of the proposed CLART model. The proposed CLART model, Varian A and Variant B are shown in Fig. 3 (a), Fig. 3 (b) and Fig. 3 (c), respectively. In Variant A, we swap the order of the two modules of the CLART model while keeping the other configurations unchanged. In Variant B, we utilize a Transformer module similar to the lattice Transformer module, with the difference that we use concatenated embeddings of E_L and E_R as the input. The results of the comparisons are shown in the second part of Table 1. We have the following findings. The performance of Variant A is comparable to the proposed CLART model (BERT+CLART vs. Variant A), indicating that the order of the Transformer modules in our proposed model has little impact on the performance. The CLART model significantly outperforms Variant B (BERT+CLART vs. Variant B), indicating that modeling the dense interaction between the lattice and radical features via cross-attention leads to superior fusion results compared to a basic concatenation operation.

4.2.3. Comparisons on fine-grained entity types

We first analyze the distribution of representative radicals in each fine-grained entity type on the TCM2020 dataset to show the potential benefits of using radical features, as shown in Fig. 4. From the figure, we observe that there is a strong correlation between medical entity types and some specific radicals. For example, within the entity type of “Drug_Efficacy”, the radical “氵 ” exhibits

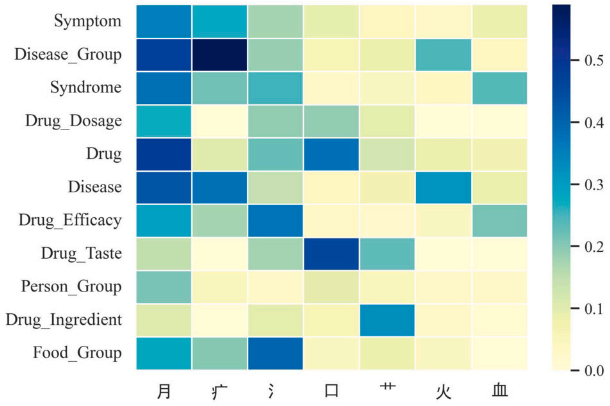


Fig. 4. Distribution of representative radicals on TCM2020 dataset.

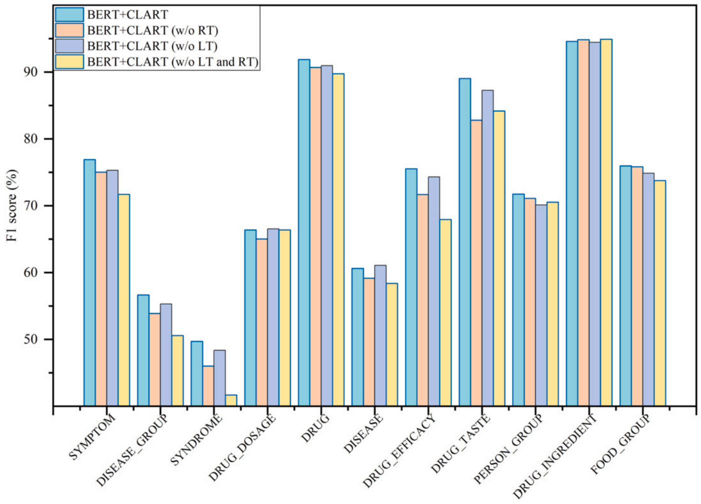


Fig. 5. Results of fine-grained entity types on TCM2020 dataset.

a higher frequency of occurrence, associated with entities such as “清热止血 (Clearing heat and stopping bleeding)”, “活血散寒 (Activating blood circulation and dispersing cold)”. Similarly, within the entity type of “Drug_Taste”, the radical “口” exhibits a higher frequency of occurrence, associated with entities such as “味甜 (Sweet taste)”, “味辛 (Spicy taste)”.

We then evaluate the performances of different CLART variants (introduced in Section 4.2.2) on the fine-grained entity types to further compare the impact of lexical and radical information, as shown in Fig. 5. We have the following findings: (1) The BERT+CLART model outperforms the other variants on most of the entity types. The trends are consistent with the overall results, indicating that the lexical and radical information can complement each other and bring greater improvement to the model. (2) In some entity types, such as “Drug_Efficacy”, BERT+CLART (w/o LT) outperforms BERT+CLART (w/o RT), indicating the radical information may bring more gain than the lexical information. A possible reason is that the distribution of radicals is more concentrated in these entity types, thus providing a better indicative capability. (3) For the entity type of “Drug_Taste”, BERT+CLART (w/o RT) performs worse than BERT+CLART (w/o LT and RT), possibly due to the noise introduced from the lexicon. However, BERT+CLART still outperforms BERT+CLART (w/o LT and RT) and achieves the best results, indicating the robustness of the model after incorporating both the lexical and radical information.

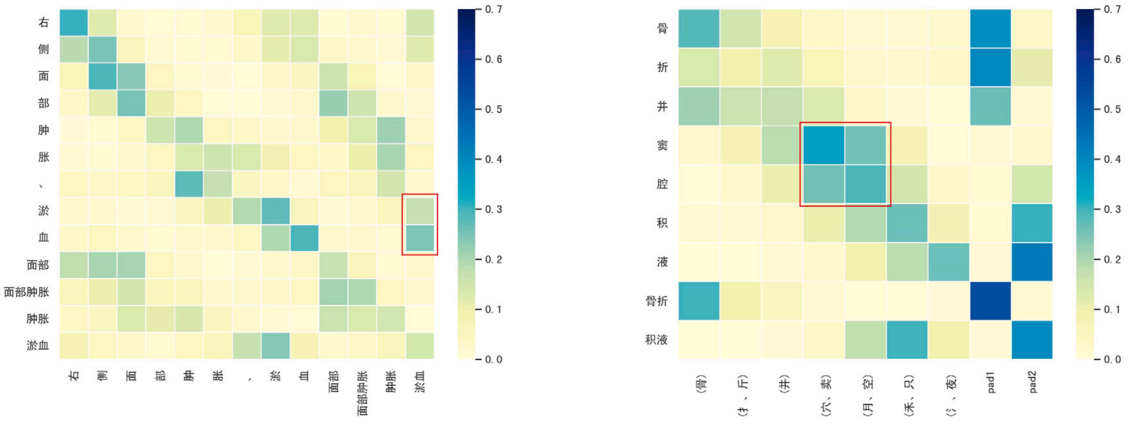
4.2.4. *Case study*

To validate the effectiveness of our model in leveraging lexical and radical information, we conduct a case study on the CCKS2017 test set, as shown in Table 3.

In the first case, CLART without lattice Transformer (w/o LT) fails to identify the symptom entity “淤血 (Congestion)”, whereas both CLART and CLART without radical Transformer (w/o RT) correctly recognize the entity. This is because the lexical word “淤血 (Congestion)” can help the model correctly identify the boundaries of the symptom entity “淤血 (Congestion)”. Furthermore, in this case, there are conflicting matched words, *i.e.*, “面部肿胀 (Facial swelling)” and “肿胀 (Swelling)”. “面部肿胀 (Facial swelling)” is the noise word. Both CLART and CLART (w/o RT) correctly identify the symptom entity “肿胀 (Swelling)”. This indicates that the

Table 3
Case study on the CCKS2017 test set. Contents with blue and red colors represent incorrect and correct entities, respectively. LT and RT are short for lattice Transformer, and radical Transformer, respectively.

Case 1									
Sentence (truncated)	右侧面部肿胀、淤血 (The right side of the face was swollen and congested)								
Matched words	面部(Face),面部肿胀(Facial swelling),肿胀(Swelling),淤血(Congest)								
Radical components	右: (一、丿、口),侧: (人、则),面: (面),部: (立、口、阝), 肿: (月、中),胀: (月、长),淤: (氵、於),血: (丶、皿)								
Characters	右	侧	面	部	肿	胀	、	淤	血
Gold labels	B-Body	I-Body	I-Body	I-Body	B-Symptom	I-Symptom	O	B-Symptom	I-Symptom
CLART	B-Body	I-Body	I-Body	I-Body	B-Symptom	I-Symptom	O	B-Symptom	I-Symptom
CLART (w/o RT)	B-Body	I-Body	I-Body	I-Body	B-Symptom	I-Symptom	O	B-Symptom	I-Symptom
CLART (w/o LT)	B-Body	I-Body	I-Body	I-Body	B-Symptom	I-Symptom	O	O	O
Case 2									
Sentence (truncated)	骨折并窦腔积液 (Fracture with sinus cavity effusion)								
Matched words	骨折(Fracture),积液(Effusion)								
radical	骨: (骨),折: (扌、斤),并: (并),窦: (穴、卖), 腔: (月、空),积: (禾、只),液: (氵、夜)								
Characters	骨	折	并	窦	腔	积	液		
Gold labels	B-Symptom	I-Symptom	O	B-Body	B-Body	O	O		
CLART	B-Symptom	I-Symptom	O	B-Body	B-Body	O	O		
CLART (w/o RT)	B-Symptom	I-Symptom	O	O	O	O	O		
CLART (w/o LT)	B-Symptom	I-Symptom	O	B-Body	B-Body	O	O		



(a) The attention weights of self-attention in lattice Transformer on case 1 (b) The attention weights of cross-attention in radical Transformer on case 2

Fig. 6. Visualizations of the attention weights on the two cases.

lattice Transformer in these two models can dynamically choose appropriate words, thus mitigating the impact of noise in matched words.

In the second case, CLART (w/o RT) fails to identify the body entity “窦腔 (Sinus cavity)”, whereas both CLART and CLART (w/o LT) correctly recognize the entity. This is because the radical components of “窦腔 (Sinus cavity)” contain radicals like “穴” and “月”, which are often associated with body entities. These radicals provide additional guiding information for the model to identify “窦腔 (Sinus cavity)” as a body type entity.

These two cases demonstrate that our model can effectively utilize both lexical and radical information to enrich the semantic representations of characters, thereby accurately identifying the medical entities. Moreover, we visualize the attention weights for these two examples, as shown in Fig. 6. The vertical and horizontal axes represent the queries and keys of attention, respectively. From Fig. 6-(a), we observe that the Chinese characters “淤 (Congestion)” and “血 (Blood)” have high attention weights on the lexical word “淤血 (Congestion)”. This indicates that the model utilizes lexical information to identify the symptom entity “淤血 (Congestion)”. From Fig. 6-(b), we observe that the Chinese characters “窦 (Sinus)” and “腔 (Cavity)” have high attention weights on the corresponding radicals "(穴, 卖)" and "(月, 空)", respectively. This indicates that the model utilizes radical information to identify the body entity "窦腔 (Sinus cavity)."

5. Conclusions

In this work, we present a novel CLART network based on the Transformer architecture, which effectively leverages both lexical and radical knowledge to achieve a deep understanding of Chinese medical texts and ensure accurate recognition of medical entities. Extensive experiments conducted on two public Chinese medical NER datasets validate the effectiveness of our proposed approach. In the future, we plan to enhance our model by incorporating additional types of knowledge, such as phonetic and syntactic information, which have been demonstrated to be effective for Chinese medical NER [27].

Author contribution statement

Yinlong Xiao: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Zongcheng Ji: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Jianqiang Li, Qing Zhu: Contributed reagents, materials, analysis tools or data.

CRediT authorship contribution statement

Yinlong Xiao: Conceptualization, Data curation, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing, Visualization. **Zongcheng Ji:** Conceptualization, Investigation, Supervision, Writing – review & editing. **Jianqiang Li:** Funding acquisition, Project administration, Supervision. **Qing Zhu:** Project administration, Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data associated with this study is publicly accessible at the following links. The link for TCM2020 dataset is https://tianchi.aliyun.com/dataset/86819, and the link for CCKS2017 dataset is https://www.biendata.xyz/competition/CCKS2017_2/.

Acknowledgements

This work is supported by the National Natural Science Foundation of China with project No. 62266041.

References

[1] L.J. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, CoRR, arXiv:1607.06450, http://arxiv.org/abs/1607.06450, 2016.

[2] Z. Dai, Z. Yang, Y. Yang, J.G. Carbonell, Q.V. Le, R. Salakhutdinov, Transformer-XL: attentive language models beyond a fixed-length context, in: A. Korhonen, D.R. Traum, L. Màrquez (Eds.), Association for Computational Linguistics, ACL, 2019, pp. 2978–2988, https://doi.org/10.18653/v1/p19-1285.

[3] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/n19-1423, arXiv:1810.04805.

[4] C. Dong, J. Zhang, C. Zong, M. Hattori, H. Di, Character-based LSTM-CRF with radical-level features for Chinese named entity recognition, in: C.Y. Lin, N. Xue, D. Zhao, X. Huang, Y. Feng (Eds.), NLPCC, Springer, 2016, pp. 239–250, https://doi.org/10.1007/978-3-319-50496-4_20.

[5] R.L. Gai, F. Gao, L.M. Duan, X.H. Sun, H.Z. Li, Bidirectional maximal matching word segmentation algorithm with rules, in: Advanced Materials Research, Trans Tech Publ., 2014, pp. 3368–3372.

[6] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: Y. Bengio, Y. LeCun (Eds.), ICLR, 2015, http://arxiv.org/abs/1412.6572.

[7] Q. Guo, X. Qiu, P. Liu, Y. Shao, X. Xue, Z. Zhang, Star-transformer, in: NAACL, 2019, pp. 1315–1325, https://doi.org/10.18653/v1/n19-1133, arXiv:1902.09113.

[8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, IEEE Computer Society, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[9] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[10] B. Hu, Z. Huang, M. Hu, Z. Zhang, Y. Dou, Adaptive threshold selective self-attention for Chinese NER, in: COLING, 2022, p. 1823.

[11] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, http://arxiv.org/abs/1508.01991, 2015.

[12] Y. Kim, Convolutional neural networks for sentence classification, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), EMNLP, ACL, 2014, pp. 1746–1751, https://doi.org/10.3115/v1/d14-1181.

[13] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: C.E. Brodley, A.P. Danyluk (Eds.), ICML, Morgan Kaufmann, 2001, pp. 282–289.

[14] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: K. Knight, A. Nenkova, O. Rambow (Eds.), NAACL, The Association for Computational Linguistics, 2016, pp. 260–270, https://doi.org/10.18653/v1/n16-1030.

[15] A. Lamurias, F.M. Couto, LasigeBioTM at MEDIQA 2019: biomedical question answering using bidirectional transformers and named entity recognition, in: Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019, 2019, pp. 523–527, https://doi.org/10.18653/v1/w19-5057.

[16] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, H. Xu, Research and applications: a comprehensive study of named entity recognition in Chinese clinical text, J. Am. Med. Inform. Assoc. 21 (2014) 808–814, https://doi.org/10.1136/amiajnl-2013-002381.

[17] X. Li, H. Yan, X. Qiu, X. Huang, FLAT: Chinese NER using flat-lattice transformer, in: ACL, Stroudsburg, PA, USA, Association for Computational Linguistics, 2020, pp. 6836–6842, https://www.aclweb.org/anthology/2020.acl-main.611, arXiv:2004.11795.

[18] X. Li, H. Zhang, X.H. Zhou, Chinese clinical named entity recognition with variant neural structures based on BERT methods, J. Biomed. Inform. 107 (2020) 103422, https://doi.org/10.1016/j.jbi.2020.103422.

[19] Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, H. Xu, Entity recognition from clinical texts via recurrent neural network, BMC Med. Inform. Decis. Mak. 17 (2017) 53–61, https://doi.org/10.1186/s12911-017-0468-7.

[20] L. Luo, N. Li, S. Li, Z. Yang, H. Lin, DUTIR at the CCKS-2018 Task1: a neural network ensemble approach for Chinese clinical named entity recognition, in: S. Hu, L. Zou (Eds.), Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2018), Tianjin, China, August 14-17, 2018, 2018, pp. 7–12, https://ceur-ws.org/Vol-2242/paper02.pdf, CEUR-WS.org.

[21] A. McCallum, D. Freitag, F.C.N. Pereira, Maximum entropy Markov models for information extraction and segmentation, in: P. Langley (Ed.), ICML, Morgan Kaufmann, 2000, pp. 591–598.

[22] A. McCallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in: W. Daelemans, M. Osborne (Eds.), NAACL, ACL, 2003, pp. 188–191, https://aclanthology.org/W03-0430/.

[23] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), ICLR, 2013, http://arxiv.org/abs/1301.3781.

[24] D.B. Nguyen, A. Abujabal, K. Tran, M. Theobald, G. Weikum, Query-driven on-the-fly knowledge base construction, Proc. VLDB Endow. 11 (2017) 66–79, https://doi.org/10.14778/3151113.3151119, http://www.vldb.org/pvldb/vol11/p66-nguyen.pdf.

[25] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training, 2018.

[26] M.S. Schlichtkrull, T.N. Kipf, P. Bloem, R. van den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: A. Gangemi, R. Navigli, M.E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, M. Alam (Eds.), The Semantic Web - 15th International Conference, Proceedings, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Springer, 2018, pp. 593–607.

[27] J. Shi, M. Sun, Z. Sun, M. Li, Y. Gu, W. Zhang, Multi-level semantic fusion network for Chinese medical named entity recognition, J. Biomed. Inform. 133 (2022) 104144, https://doi.org/10.1016/j.jbi.2022.104144.

[28] D. Sui, Y. Chen, K. Liu, J. Zhao, S. Liu, Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), EMNLP-IJCNLP, Association for Computational Linguistics, 2019, pp. 3828–3838, https://doi.org/10.18653/v1/D19-1396.

[29] I.J. Unanue, E.Z. Borzeshi, M. Piccardi, Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition, J. Biomed. Inform. 76 (2017) 102–109, https://doi.org/10.1016/j.jbi.2017.11.007.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017, pp. 5999–6009, http://arxiv.org/abs/1706.03762.

[31] Q. Wan, J. Liu, L. Wei, B. Ji, A self-attention based neural architecture for Chinese medical named entity recognition, Math. Biosci. Eng. 17 (2020) 3498–3511.

[32] Q. Wang, Y. Zhou, T. Ruan, D. Gao, Y. Xia, P. He, Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition, J. Biomed. Inform. 92 (2019), https://doi.org/10.1016/j.jbi.2019.103133.

[33] Y. Wu, M. Jiang, J. Xu, D. Zhi, H. Xu, Clinical named entity recognition using deep learning models, in: AMIA 2017, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 4–8, 2017, AMIA, 2017, https://knowledge.amia.org/65881-amiab-1.4254737/t003-1.4258387/f003-1.4258388/2730946-1.4258431/2731659-1.4258428.

[34] Y. Xiong, H. Peng, Y. Xiang, K.C. Wong, Q. Chen, J. Yan, B. Tang, Leveraging multi-source knowledge for Chinese clinical named entity recognition via relational graph convolutional network, J. Biomed. Inform. 128 (2022) 104035, https://doi.org/10.1016/j.jbi.2022.104035.

[35] H. Yan, B. Deng, X. Li, X. Qiu, TENER: adapting transformer encoder for named entity recognition, CoRR, arXiv:1911.04474, 2019.

[36] J. Yang, Y. Zhang, F. Dong, Neural word segmentation with rich pretraining, in: R. Barzilay, M.Y. Kan (Eds.), ACL, Association for Computational Linguistics, 2017, pp. 839–849, https://doi.org/10.18653/v1/P17-1078.

[37] J. Zhang, D. Shen, G. Zhou, J. Su, C.L. Tan, Enhancing HMM-based biomedical named entity recognition by studying special phenomena, J. Biomed. Inform. 37 (2004) 411–422, https://doi.org/10.1016/j.jbi.2004.08.005.

[38] Y. Zhang, H. Lin, Z. Yang, J. Wang, S. Zhang, Y. Sun, L. Yang, A hybrid model based on neural networks for biomedical relation extraction, J. Biomed. Inform. 81 (2018) 83–92, https://doi.org/10.1016/j.jbi.2018.03.011.

[39] Y. Zhang, J. Yang, Chinese NER using lattice LSTM, in: ACL, Association for Computational Linguistics, Stroudsburg, PA, USA, 2018, pp. 1554–1564, http://aclweb.org/anthology/P18-1144, arXiv:1805.02023.

[40] S. Zhao, Z. Cai, H. Chen, Y. Wang, F. Liu, A. Liu, Adversarial training based lattice LSTM for Chinese clinical named entity recognition, J. Biomed. Inform. 99 (2019), https://doi.org/10.1016/j.jbi.2019.103290.