applied sciences

MDPI

Article

Improving Chinese Named Entity Recognition by Interactive Fusion of Contextual Representation and Glyph Representation

Ruiming Gu [1,*], Tao Wang [2], Jianfeng Deng [2] and Lianglun Cheng [1,*]

[1] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China
[2] School of Automation, Guangdong University of Technology, Guangzhou 510006, China
* Correspondence: 2112005028@mail2.gdut.edu.cn (R.G.); llcheng@gdut.edu.cn (L.C)

Abstract: Named entity recognition (NER) is a fundamental task in natural language processing. In Chinese NER, additional resources such as lexicons, syntactic features and knowledge graphs are usually introduced to improve the recognition performance of the model. However, Chinese characters evolved from pictographs, and their glyphs contain rich semantic information, which is often ignored. Therefore, in order to make full use of the semantic information contained in Chinese character glyphs, we propose a Chinese NER model that combines character contextual representation and glyph representation, named CGR-NER (Character–Glyph Representation for NER). First, CGR-NER uses the large-scale pre-trained language model to dynamically generate contextual semantic representations of characters. Secondly, a hybrid neural network combining a three-dimensional convolutional neural network (3DCNN) and bi-directional long short-term memory network (BiLSTM) is designed to extract the semantic information contained in a Chinese character glyph, the potential word formation knowledge between adjacent glyphs and the contextual semantic and global dependency features of the glyph sequence. Thirdly, an interactive fusion method with a crossmodal attention and gate mechanism is proposed to fuse the contextual representation and glyph representation from different models dynamically. The experimental results show that our proposed model achieves 82.97% and 70.70% F1 scores on the OntoNotes 4 and Weibo datasets. Multiple ablation studies also verify the advantages and effectiveness of our proposed model.

Keywords: Chinese named entity recognition; Chinese character glyph; interactive fusion; crossmodal attention; glyph representation

Citation: Gu, R.; Wang, T.; Deng, J.; Cheng, L. Improving Chinese Named Entity Recognition by Interactive Fusion of Contextual Representation and Glyph Representation. *Appl. Sci.* **2023**, *13*, 4299. https://doi.org/10.3390/app13074299

Academic Editor: Francisco García-Sánchez

Received: 20 December 2022
Revised: 21 March 2023
Accepted: 24 March 2023
Published: 28 March 2023

1. Introduction

Named entity recognition aims to identify entities such as names, places and organizations from unstructured texts. As the basic task of information extraction, NER will directly affect the downstream tasks such as relation extraction [1], entity linking [2] and the question answering system [3]. NER is often treated as a sequence labeling task; that is, each unit in the text is labeled with a predefined entity label [4]. In recent years, NER models based on deep learning have shown outstanding performance and gradually become the mainstream method [5].

Compared with English, Chinese texts lack natural word separators and explicit word boundaries. Therefore, Chinese NER needs to recognize the boundaries and types of entities at the same time, which is more difficult than English NER [6]. One way to implement Chinese NER is to segment sentences into word sequences first, and then recognize entities based on word sequences. However, the recognition effect of this method is vulnerable to the negative impact of segmentation errors. In order to avoid entity prediction failure caused by word segmentation errors, Chinese NER tasks generally model based on characters.

In English, words with the same prefix or suffix can often be classified into one category. For example, words such as geologist, pianist and physicist have the suffix "-ist",

which can represent entities engaged in a certain occupation. Chiu and Ma et al. [7,8] used a convolutional neural network (CNN) to capture the spelling features of prefixes and suffixes in English words; experiments showed that these spelling features can enhance the model's understanding of word semantics and improve the recognition performance of the NER model. Large-scale pre-trained language models, such as BERT [9] and GPT [10], use the word-piece method to split a word into multiple pieces, which can effectively express out-of-vocabulary words while reducing the size of the vocabulary, thus improving the semantic representation ability of the model. Different from English, Chinese characters are ideographic characters and have no prefix, suffix, root and other characteristics of English words. Chinese characters evolved from hieroglyphs, and their glyph structure contains rich semantic information. For example, "花 (flower)", "荷 (lotus)" and "茶 (tea)" all contain the radical "艹 (grass)", indicating that these three characters are related to herbaceous plants. This means that when there are characters out-of-vocabulary, we can deduce the approximate semantics of out-of-vocabulary characters from the glyph features of Chinese characters, such as radicals. At the same time, in Chinese naming culture, it is also preferred to use radicals with Wu-Xing (five elements) or positive meanings as part of the name. For example, radicals such as "火 (fire)" and "日 (sum)" often appear in names, while radicals such as "疒 (epilepsy)", which represent diseases, rarely appear in names. These patterns may be useful for identifying the Chinese names of people and businesses. Moreover, Chinese characters are mostly polysemous, meaning that one Chinese character has multiple meanings, and most of them need to be combined with other Chinese characters into words to have a more clear semantic representation. For example, the Chinese character "行" in "银行 (bank)" and "爬行 (crawling)" have completely different semantics. In "银行", "行" is a noun, indicating an institution, while in "爬行", "行" is a verb, indicating walking. Moreover, the potential word formation knowledge between the glyphs of adjacent characters can enrich the semantic information of Chinese characters and alleviate the problem of polysemy. For instance, in "爪" (often representing an action related to the hand), the radical of "爬", can help to express that "行" is a verb, indicating the meaning of walking. However, most of the existing Chinese NER models focus on the contextual semantic features of the character, and ignore the rich semantic information contained in the Chinese character glyph.

In addition, the simple concatenation of contextual semantic features from text and glyph features from images can easily cause information redundancy, and the model cannot make full use of the complementary information between different modal features. This multimodal feature fusion problem mainly focuses on video question answering [11], multimodal sentiment analysis [12], etc. In order to fully fuse features from different modals, Zadeh et al. successively proposed a tensor fusion network [13], memory fusion network [14] and dynamic fusion graph [15]. Further, Tsai et al. [16] proposed MulT, which uses multiple crossmodal Transformers to fuse the three modal features of text, audio and vision. At the heart of MulT is the use of a crossmodal attention mechanism to repeatedly strengthen the features from two modals, so that MulT can fully learn useful and complementary information from different modals. However, these methods focus on the fusion of sentence-level features, while the Chinese NER task needs to complete the fusion of character-level features.

The questions of how to better utilize the semantic knowledge contained in Chinese character glyphs and their sequences, and how to better integrate contextual semantic features from text and glyph features from images, are two of the main challenges facing Chinese NER currently. Therefore, in order to make full use of the semantic knowledge contained in Chinese character glyphs and fuse the extracted semantic features of glyphs into the character-based Chinese NER model to improve the recognition performance of the model, this paper proposes a novel Chinese NER model based on contextual representation and glyph representation, named CGR-NER. First, for the contextual representation, CGR-NER uses BERT to dynamically generate contextual semantic representations of characters to solve the polysemy problem; for the glyph representation, a hybrid neural

network CBHNN, which combines a three-dimensional convolutional neural network and bi-directional long short-term memory network, is proposed to extract the semantic information contained in a Chinese character glyph, the potential word formation knowledge between adjacent glyphs and the contextual semantic and global dependent features of glyph sequences from the perspective of images. Second, to make the representations from different modals complement each other and dynamically adjust the contributions of different representations in different contexts, an interactive fusion method with a crossmodal attention and gate mechanism is proposed, so as to realize the information interaction between the contextual representation and glyph representation and generate a fusion representation. Finally, the fusion representation is input into the BiLSTM-CRF network for Chinese named entity recognition. In summary, the main tasks and contributions of this paper are as follows:

- A novel hybrid neural network CBHNN is proposed to extract and utilize the rich semantic knowledge contained in Chinese character glyphs and glyph sequences.
- An interactive fusion method with a crossmodal attention and gate mechanism is proposed to realize the feature complementarity among different modal representations and dynamically control the fusion of contextual representation and glyph representation.
- Sufficient experiments are conducted on two mainstream Chinese NER datasets, and the results show the advantages and effectiveness of our model. The results of several ablation experiments show that our model still has good recognition performance in the case of insufficient training data or out-of-vocabulary character interference.

The rest of this paper is organized as follows: Section 2 describes the work related to the Chinese NER task; Section 3 introduces the architecture of CGR-NER; Section 4 presents our experimental details and ablation studies; finally, Section 5 gives the conclusions and future works.

2. Related Work

In traditional NER work, rule-based, dictionary-based and statistical machine learning methods are mainly used, such as the hidden Markov model [17], conditional random field (CRF) [18], etc. Such traditional NER models strongly rely on manual features. When facing different fields, manual features often need to be redesigned, resulting in poor model migration performance. With the development of deep learning, the end-to-end neural network model not only has better performance in NER tasks, but is also independent of manual features, so it has become the mainstream method in NER tasks. Huang et al. [19] proposed an end-to-end BiLSTM-CRF model by combining BiLSTM and CRF, and achieved excellent performance in sequence labeling tasks such as NER. Based on BiLSTM-Attention-CRF and a contextual representation combining the character level and word level, Ali et al. [20] proposed CaBiLSTM for Sindhi named entity recognition, achieving the best results on the SiNER dataset without relying on additional language-specific resources.

2.1. Chinese NER Based on External Resource Enhancement

In order to further improve the recognition performance of NER models, researchers try to integrate external knowledge such as a lexicon or syntactic characteristics into the model. Zhu et al. [21] proposed SDI-NER, which uses a graph neural network to learn syntactic dependency knowledge and integrates it into the BiLSTM-CRF model, so as to enhance the understanding of sentence semantics by the NER model. Li et al. [22] proposes FLAT, which causes potential matching words to interact with characters directly by improving the position encoding and attention mechanism, and improves the parallel computing efficiency of the NER model. Nie et al. [23] proposed AESINER, which integrates part-of-speech labels, syntactic constituents and dependency relations in sentences into an NER model through a key-value memory network. Xue et al. [24] proposed PLTE, which models all characters and matched words in parallel with batch processing by extending

the Transformer encoder. Hu et al. [25] proposed KGNER, a model that incorporates a knowledge graph into the Chinese NER task, and uses a new position encoding method and masking matrix to limit the impact of knowledge graph information on the original semantic meaning of sentences. Liu et al. [26] proposed LEBERT, a model that integrates the matched lexicon knowledge into the underlying layer of BERT through a lexicon adapter, achieving the deep fusion of character and vocabulary knowledge. Liu et al. [27] proposed MW-NER, which integrates word embeddings and n-gram embeddings through two independent attention networks to obtain multi-granular word information. It also attempts different fusion strategies to integrate character information and multi-granular word information, reducing the interference of word segmentation errors and noise in model performance. However, the performance of such methods is easily affected by the quality of external resources, such as the accuracy of parsing tools. At the same time, for entity recognition tasks in some special fields, it is necessary to build the corresponding lexicon or knowledge graph first, and then provide these methods for entity recognition.

2.2. Chinese NER Based on Chinese Character Glyph Enhancement

Chinese characters evolved from hieroglyphs, and their glyphs also contain rich semantic information. The work of extracting this semantic information to enhance the recognition performance of Chinese NER models can be divided into two methods, including the decomposition of Chinese characters and image processing.

2.2.1. The Decomposition of Chinese Characters

The decomposition of Chinese characters is to use radicals, wu-bi or structural components to represent Chinese character glyphs. For example, for Chinese character "茶", its radical is "艹"; wu-bi are "AWSU" and structural components are "艹人木". A typical method is ME-CNER [28], which improves the NER performance by using a CNN to capture potential glyph information in a radical sequence. Wu et al. [29] proposed MECT, which introduces the structural components of Chinese characters into FLAT, and designs a Cross-Transformer network to fuse the information of characters, radicals and words. Li et al. [30] proposed MFE-NER, which uses wu-bi to help generate a glyph embedding, and combines it with character and phonetic features to generate multi-feature fusion embedding, so as to improve the performance of the Chinese NER model and reduce the influence caused by character substitution. Lv et al. [31] proposed StyleBERT to improve the semantic expression ability of a Chinese pre-training language model by integrating word, pinyin, wu-bi and chaizi information. However, the method based on the decomposition of Chinese characters ignores the overall structure and spatial characteristics of Chinese characters, because the structural components of Chinese characters are not all horizontally distributed as with English words, and the distribution of structural components of Chinese characters can be vertical, surrounding, half-surrounding, etc., with obvious spatial characteristics. For example, "叶 (leaf)" and "困 (trap)" have the same radical "口" and structural components "口十", but the spatial distribution of the structural components is different.

2.2.2. Image Processing

In the work based on image processing, the semantic information of Chinese character glyphs is learned by extracting the visual features of Chinese character images, which avoids the shortcomings of the decomposition of Chinese characters method. Meng et al. [32] proposed Glyce, a representation vector of Chinese character glyphs. Glyce uses Chinese character images of different periods and writing styles to enrich the visual features of Chinese character glyphs; it also designs the Tianzige-CNN network to extract the semantic information contained in the glyph images, and updates the best records of multiple Chinese NER datasets. Song et al. [33] treat the problem of extracting semantic information from Chinese character glyph images as an image classification problem and propose the GlyNN network to encode Chinese character images. Then, the glyph feature obtained is

used as an additional representation of the NER model. Sun et al. [34] proposed ChineseBERT, which integrates Chinese character images with different writing styles and Chinese character pinyin into the pre-training of the language model, effectively enhancing the semantic expression ability of the Chinese pre-training language model. However, these methods encode each Chinese character glyph independently, ignoring the potential word formation knowledge between adjacent glyphs. Moreover, the representation sequences of different modals have their unique contextual semantics and global dependency information, and the above methods do not explicitly extract this semantic information from the glyph representation sequences. In addition, the above methods simply concatenate the contextual representation and glyph representation from different modals, ignoring the interactive information between the two representations.

In comparison, our approach fully excavates and makes use of the semantic knowledge contained in Chinese character glyphs. First, we propose CBHNN, which uses the characteristics of 3DCNN, which simultaneously extracts spatial and temporal dimension information, to learn the semantic knowledge contained in Chinese character glyphs and the potential word formation knowledge between adjacent glyphs, and explicitly captures the contextual semantics and global dependence features of a glyph sequence through BiLSTM to generate the glyph representation. Second, an interactive fusion method with a crossmodal attention and gate mechanism is proposed to realize the feature complementarity among the contextual representation and glyph representation, reducing the information redundancy and improving the recognition performance of the model.

3. CGR-NER Architecture

Chinese NER is conventionally regarded as a typical sequence labeling task: for an input sentence $S = \{c_1, c_2, \ldots, c_n\}$ with n tokens, we output its corresponding named entity label sequence $Y = \{y_1, y_2, \ldots, y_n\}$ in the same length. Following this paradigm, the CGR-NER architecture mentioned in this paper is shown in Figure 1. It is composed of four layers: a representation layer, fusion layer, semantic encoding layer and decoding layer. We will introduce CGR-NER in detail in this section.

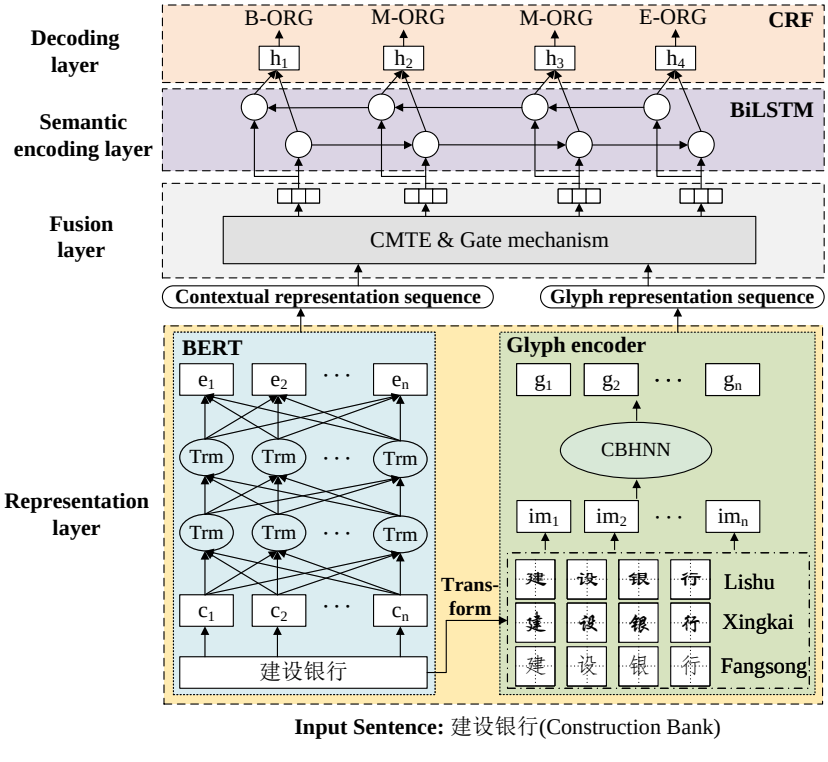

Figure 1. The overview of the proposed CGR-NER model.

3.1. Representation Layer

In the representation layer, each character of the sentence S will be converted into the corresponding contextual representation and glyph representation.

3.1.1. Contextual Representation

In previous studies, Word2Vec [35] was often used to generate a contextual representation. However, the contextual representation generated by Word2Vec was context-independent and could not effectively represent polysemy words. To solve this problem, CGR-NER uses BERT to dynamically generate a contextual representation. BERT is a pre-trained language representation model, which is stacked by multiple layers of Transformer encoders. It no longer uses the traditional one-way language model or shallow concatenation of two one-way language models for training, but uses the masked language model to train the two-way Transformer encoder, and finally generates a deep two-way language representation that can fuse contextual semantic information. The process of generating the contextual representation sequence $E_c = \{e_1, e_2, \ldots, e_n\}$ corresponding to sentence S can be denoted as

$$E_c = BERT(S) \tag{1}$$

where e_i is the contextual representation corresponding to the ith character of sentence S, and n is the sentence length. $E_c \in \mathbb{R}^{n \times d_c}$, d_c is the dimension of the contextual representation.

3.1.2. Glyph Representation

From the perspective of image processing, we propose a hybrid neural network named CBHNN that combines 3DCNN and BiLSTM to learn the semantic information contained in Chinese character glyphs and obtain the glyph representation of each character, as shown in Figure 2. The input of CBHNN is a sequence of glyph images $P = \{p_1, p_2, \ldots, p_n\}$, for which the characters of the sentence S need to be converted into corresponding glyph images first. In order to enrich the visual features of Chinese character glyphs, three glyph images with different styles, Lishu, Xingkai and Fangsong, were used to concatenate the dimensions of the image channel. Therefore, the glyph image size of CBHNN input was $3 \times 24 \times 24$. Then, three 3DCNN blocks were used to capture the glyph feature contained in the glyph image. Each 3DCNN block contains a 3DCNN, ReLU function and 3D max-pooling operation. The number of output channels of each block was 64, 128, 256, and the output of the last block was batch-normalized and reshaped to obtain the glyph feature vector of 256 dimensions. Different from 2DCNN, 3DCNN can simultaneously extract the feature information of spatial and temporal dimensions, which means that CBHNN can not only extract the glyph visual feature in the Chinese character image, but also capture the glyph visual features of adjacent characters at the same time, so that the model can learn the potential word formation knowledge between glyphs. It makes the same Chinese character dynamically generate different glyph feature vectors in different contexts, which is beneficial to alleviate the problem of polysemy. Finally, CBHNN uses BiLSTM to extract contextual semantics and global dependency features from the glyph feature vector sequence, to obtain the glyph representation sequence $E_g = \{g_1, g_2, \ldots, g_n\}$. The process of obtaining the glyph representation sequence corresponding to the sentence S can be denoted as

$$E_g = CBHNN(P) \tag{2}$$

where g_i is the glyph representation corresponding to the ith character of sentence S. $E_g \in \mathbb{R}^{n \times d_g}$, d_g are the dimensions of the glyph representation.

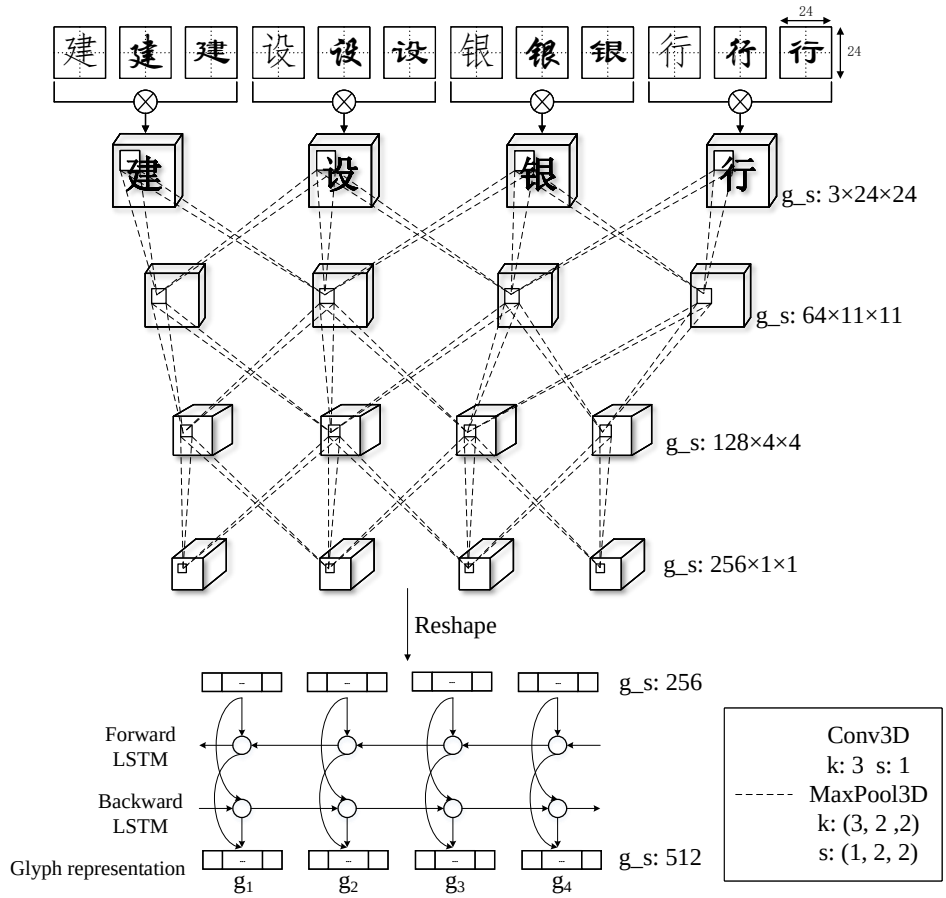

Figure 2. Architecture of CBHNN. "g_s" represents the tensor size of output from each network. "Conv3D", "MaxPool3D", "k" and "s"" stand for 3DCNN, 3D max-pooling, kernel size and stride.

3.2. Fusion Layer

After obtaining the contextual representation sequence and glyph representation sequence, we further propose an interactive fusion method with a crossmodal attention and gate mechanism to obtain a semantic enhanced fusion representation sequence, as shown in Figure 3. We first use two crossmodal Transformer encoders (CMTE) to complete the feature complementarity between the two modal representations, and then use a gate mechanism to dynamically adjust the contributions of different modal representations in different contexts.

At the heart of CMTE to realize the interaction between different modal features is that the source modal feature can adaptively learn the complementary information from the target modal feature through the crossmodal attention mechanism. In the following, we take the contextual representation adaptively learning potential complementary information from the glyph representation as an example, which is denoted as $g \rightarrow c$ (see $CMTE_{g \rightarrow c}$ in Figure 3). In this example, the contextual representation sequence E_c is used as the source modal feature sequence, and the glyph representation sequence E_g is the target modal feature sequence. Formally, $CMTE_{g \rightarrow c}$ computes feed-forwardly for $i = 1, \ldots, D$ layers of crossmodal attention blocks. In the crossmodal attention block of each layer, the crossmodal attention is calculated first:

$$U_{g \rightarrow c}^{[0]} = E_c W_c \tag{3}$$

$$Q_c^{[i-1]} = LN(U_{g \rightarrow c}^{[i-1]}) \tag{4}$$

$$K_g = V_g = LN(E_g W_g) \tag{5}$$

$$S_{g\to c}^{[i-1]} = MHCMA(Q_c^{[i-1]}, K_g, V_g) \tag{6}$$

where $W_c \in \mathbb{R}^{d_c \times d}$, $W_g \in \mathbb{R}^{d_g \times d}$ are transformation matrices, and LN means layer normalization. $MHCMA$ represents multi-head crossmodal attention, which is the key to achieving crossmodal interaction. To enhance the direction and distance perception ability of the model in the interaction process, we adopt the relative positional encoding scheme proposed by TENER [36], so $MHCMA$ can be calculated as follows:

$$MHCMA(Q_c^{[i-1]}, K_g, V_g) = [head^{(1)}; \ldots; head^{(n)}]W_o \tag{7}$$

$$Q^{(h)}, K^{(h)}, V^{(h)} = Q_c^{[i-1]}W_q^{(h)}, K_g^{d_k}, V_g W_v^{(h)} \tag{8}$$

$$R_{t-j} = [\ldots sin(\frac{t-j}{10000^{2m/d_k}})cos(\frac{t-j}{10000^{2m/d_k}})\ldots]^T \tag{9}$$

$$A_{t,j}^{rel} = Q_t^{(h)}K_j^T + Q_t^{(h)}R_{t-j}^T + uK_j^T + vR_{t-j}^T \tag{10}$$

$$head^{(h)} = softmax(A^{rel})V^{(h)} \tag{11}$$

where n is the number of heads, h is the index of heads, $d_k \times n = d$. $W_O \in \mathbb{R}^{d\times d}$, $W_q^{(h)} \in \mathbb{R}^{d\times d_k}$, $W_v^{(h)} \in \mathbb{R}^{d\times d_k}$, $u \in \mathbb{R}^{d_k}$, $v \in \mathbb{R}^{d_k}$ are learnable parameters. $K_g^{d_k}$ is the part of K_g allocated to each head. $R_{t-j} \in \mathbb{R}^{d_k}$ is the relative positional encoding, $m \in [0, \frac{d_k}{2}]$ in Equation (9). t, j is the index of the target token and context token, respectively. $Q_t^{(h)}$ is the query vector of token t, K_j is the key vector of token j. $Q_t^{(h)}K_j^T$ is the attention score between token t, j; $Q_t^{(h)}R_{t-j}^T$ is the bias of the token t in the relative position; uK_j^T is the bias of the token j; vR_{t-j}^T is the bias term for distance and direction between tokens t, j.

Then, a reset function r_a is used to dynamically control the residual connection of $Q_c^{[i-1]}$ and $S_{g\to c}^{[i-1]}$:

$$r_a = \sigma(W_{l_1}Q_c^{[i-1]} + W_{l_2}S_{g\to c}^{[i-1]} + b_r) \tag{12}$$

$$\bar{S}_{g\to c}^{[i-1]} = [r_a \circ Q_c^{[i-1]}] + [(1_a - r_a) \circ S_{g\to c}^{[i-1]}] \tag{13}$$

where σ is the sigmoid function, W_{l_1}, W_{l_2} are trainable parameters and b_r is the bias term. $\circ$ represents the element-wise multiplication operation and 1_a is a 1-matrix with its dimension matching r_a.

After this, the position-wise feed-forward can be calculated as follows:

$$U_{g\to c}^{[i]} = LN(\bar{S}_{g\to c}^{[i-1]}) + F_{\theta^{[i]}}(LN(\bar{S}_{g\to c}^{[i-1]})) \tag{14}$$

where F_θ represents the position-wise feed-forward network parameterized by θ.

In this process, E_c constantly interacts with E_g through multi-layer crossmodal attention blocks, and updates its representation sequence by learning the potential complementary feature in E_g. The process of the glyph representation adaptively learning complementary information from the contextual representation ($c \to g$) is similar to $g \to c$, except that the source modal feature sequence is E_g and the target modal feature sequence is E_c.

After using CMTE to complete the interaction between the two modal representations, we propose a gate mechanism to dynamically adjust the contributions of different modal

representations in different contexts to enhance the performance of the model. In detail, we use a gated threshold r_f to evaluate the outputs from the $CMTE_{g\to c}$ and the $CMTE_{c\to g}$ by

$$r_f = \sigma(W_{g\to c}U_{g\to c} + W_{c\to g}U_{c\to g} + b_f) \tag{15}$$

$$U_{fusion} = [r_f \circ U_{g\to c}] \oplus [(1_f - r_f) \circ U_{c\to g}] \tag{16}$$

where $W_{g\to c}$, $W_{c\to g}$ are trainable parameters and b_f is the bias term. $U_{g\to c}$ is the output of $CMTE_{g\to c}$ and $U_{c\to g}$ is the output of $CMTE_{c\to g}$. 1_f is a 1-matrix with its dimension matching r_f. $\oplus$ denotes the concatenation operation.

Finally, we obtain the semantic enhanced fusion representation sequence U_{fusion}.

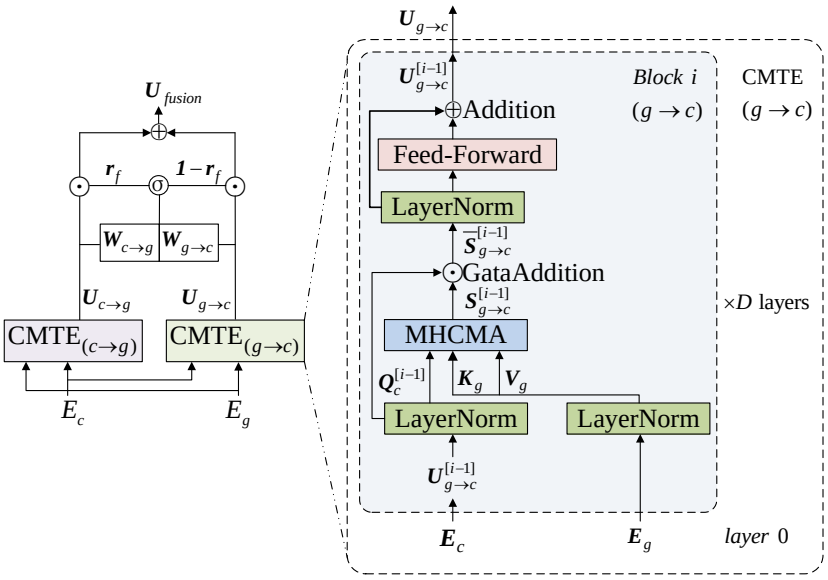

Figure 3. The architecture of the fusion layer consists of two crossmodal Transformer encoders and a gate mechanism. Each crossmodal Transformer encoder is stacked by multiple crossmodal attention blocks.

3.3. Semantic Encoding Layer

BiLSTM is suitable for modeling sequence problems. It not only effectively solves the problem of gradient disappearance and explosion when traditional recurrent neural networks deal with long-distance dependency information, but also captures more dependency features from the forward and backward direction of the sequence and outputs hidden feature vectors with both forward and backward sequence dependency information. CGR-NER uses BiLSTM to encode the glyph feature sequence and fusion representation sequence and extract their contextual semantics and global dependency features.

For the fusion representation sequence $U_{fusion} = \{u_1, u_2, \ldots, u_n\}$ output from the fusion layer, the process of BiLSTM to calculate the corresponding hidden state of each representation is as follows:

$$\overrightarrow{h_i} = \overrightarrow{LSTM}(u_i, \overrightarrow{h_{i-1}}) \tag{17}$$

$$\overleftarrow{h_i} = \overleftarrow{LSTM}(u_i, \overleftarrow{h_{i-1}}) \tag{18}$$

$$h_i = [\overrightarrow{h_i} : \overleftarrow{h_i}] \tag{19}$$

where $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ denote the hidden state of the forward and backward for u_i. The final output hidden state sequence is $H = \{h_1, h_2, \ldots, h_n\}$.

3.4. Decoding Layer

The conditional random field can effectively alleviate the label bias problem and is often used as the decoder for sequence labeling tasks. For the output $H = \{h_1, h_2, \ldots, h_n\}$ of the semantic encoding layer, assuming that its predicted label sequence is $\widehat{y} = \{\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n\}$, the probability of label $\widehat{y}_i$ is calculated by

$$\widehat{y}_i = \arg\max_{y_i \in \zeta} \frac{\exp(W_c \cdot h_i + b_c)}{\sum_{y_{i-1} y_i} \exp(W_c \cdot h_i + b_c)} \tag{20}$$

where ζ is the set with all named entity labels, and W_c and b_c are trainable parameters. In the process of decoding, we use the Viterbi algorithm to generate the optimal label sequence $\widehat{y}$.

4. Experiments and Analyses

We conduct extensive experiments on two mainstream Chinese NER datasets to evaluate the entity recognition performance of CGR-NER and compare it with the mainstream Chinese NER model. We present the experimental datasets, parameter setting, experimental results and analysis in this section.

4.1. Datasets

The effectiveness of CGR-NER is evaluated on two datasets from different fields, namely Weibo [37] and OntoNotes 4 [38]. The corpus of the Weibo dataset is collected from Chinese social media site Sina Weibo, and labeled with four entity labels: PER (person), ORG (organization), LOC (location) and GEO (geography). The corpus of the OntoNotes 4 dataset is collected from a news domain, including four entity labels: PER, ORG, LOC and GEO. The statistics of these datasets are shown in Table 1.

Table 1. Statistics of the datasets.

Dataset	Type	Train	Dev	Test
Weibo	Sentence	1.35 k	0.27 k	0.27 k
	Character	73.78 k	14.51 k	14.84 k
	Entity	1.90 k	0.39 k	0.42 k
OntoNotes 4	Sentence	15.72 k	4.30 k	4.35 k
	Character	491.90 k	200.51 k	208.07 k
	Entity	13.37 k	6.95 k	7.68 k

4.2. Parameter Setting and Evaluation Metrics

In our experiment, precision (P), recall (R) and F1 score (F1) are used as evaluation metrics. Only when the model correctly recognizes the boundary and type of entities at the same time can it represent correct recognition. In order to avoid experimental error and contingency, we tested each experiment ten times, and calculated the average values of the evaluation metrics as the final experimental results.

We manually adjusted the hyperparameters according to the recognition performance of the model on the development set. At the representation layer, we used BERT, a pre-trained language model released by Google, to generate a contextual representation and follow its default Settings. For CBHNN, we follow the settings in Section 3.1.2. For CMTE, d_k is set to 64, and the numbers of crossmodal attention heads and layers are set to 8 and 2, respectively. The hidden size of BiLSTM in the semantic encoding layer is set to 512. The maximum sentence length is 200, the batch size is 32 and the dropout rate is 0.4 for all datasets. The training epoch is set to 100 for the Weibo dataset, and 20 for OntoNotes 4. Adam is adopted to optimize the model, with a learning rate of 0.0004. The model with the highest F1 score on the development set will be the best, and the performance evaluation

will be conducted on the test set. All the experiments are conducted on an NVIDIA 3090 GPU. Table 2 shows the selection range of hyperparameters.

Table 3 shows the evaluation results on the development set of different kernel sizes (context window size) in CBHNN and different hidden sizes in the semantic encoding layer. We can find that the kernel size is set to 3 and the hidden size is set to 512, which are superior to other optional parameters. Therefore, our model selects the kernel size of 3 in CBHNN and the hidden size of 512 in the semantic encoding layer as the final parameters.

Table 2. The selection range of hyperparameters.

Hyperparameter	Value
CBHNN_kernel_size	3, 5, 7
CBHNN_hidden_size	128, 256, 512
CMTE_layer	1, 2, 3
CMTE_head	4, 8, 16
Dropout_rate	0.2, 0.3, 0.4, 0.5
Learning_rate	0.00001, 0.0001, 0.0004, 0.001
Semantic_Encoding_Layer_hidden_size	128, 256, 512

Table 3. The evaluation results of different hyperparameters on the development set. Among them, kernel size represents the size of the context window in CBHNN; hidden size represents the dimensions of the hidden state in the semantic encoding layer.

Hyperparameter		Weibo			OntoNotes 4		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Kernel Size	3	73.92	75.45	**74.68**	**82.52**	79.88	**81.18**
	5	**74.48**	74.68	74.58	80.8	**80.13**	80.47
	7	72.59	**75.97**	74.24	81.04	78.91	79.96
Hidden Size	128	72.24	**75.97**	74.06	78.31	**80.4**	79.34
	256	73.59	74.16	73.87	80.33	80.37	80.35
	512	**73.62**	75.71	**74.65**	**80.9**	79.84	**80.37**

4.3. Model Training

To demonstrate the training situation and computational cost of our model, we reimplement BERT-BiLSTM-CRF because it is the mainstream base model in the NER task, and we take it as the baseline. Figure 4 shows the variation tendency of evaluation metrics and loss value on the development set with the training epoch. As can be seen from Figure 4, on the Weibo dataset, the evaluation metrics and loss value exhibit significant fluctuations during the early stages of training. This is because the Weibo dataset has a relatively small amount of data, approximately one twelfth of the OntoNotes 4 dataset, so the model requires more training time to learn and adjust before it can stably demonstrate good performance. Meanwhile, compared with BERT-BiLSTM-CRF, the loss curve of CGR-NER is lower and smoother, indicating the better fit of the CGR-NER model. Moreover, to demonstrate the computational cost of CGR-NER, we also report the total number of parameters and the average time per epoch during training for both BERT-BiLSTM-CRF and CGR-NER in Table 4.

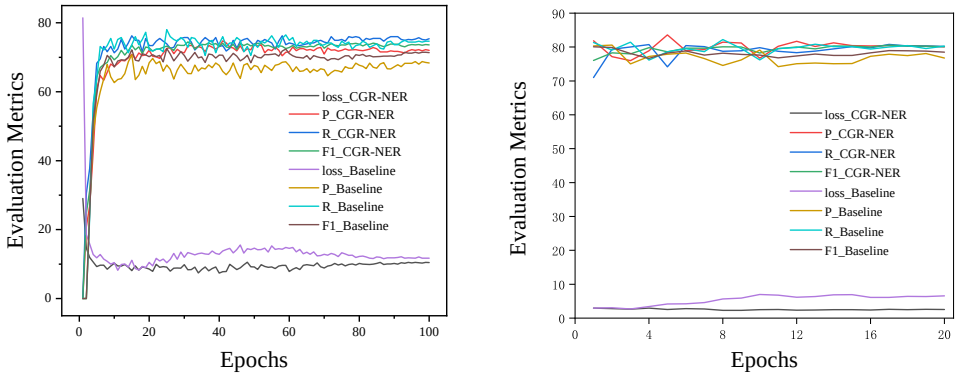

Figure 4. The variation tendency of evaluation metrics and loss value on the development set with the training epoch. (**a**) Weibo dataset; (**b**) OntoNotes 4 dataset.

Table 4. The total number of parameters and the average time per epoch during training.

Model	Weibo	OntoNotes 4	#Parameter
Baseline	14.88 s	130.90 s	426.27 M
CGR-NER	19.57 s	153.57 s	631.81 M

4.4. Effectiveness Study

We compare CGR-NER with several classic and recently published NER models introduced in the related work. Among them, SDI-NER, FLAT+BERT, AESINER, PLTE+BERT, LEBERT, KGNER and MW-NER enhance the recognition performance of the NER model by introducing a lexicon, syntax knowledge and a knowledge graph; MECT, StyleBERT, GlyNN, Glyce, MFE-NER and ChineseBERT enhance the recognition performance of the NER model by fusing the unique language features of Chinese characters, such as radical, wu-bi, the structural components of Chinese characters and pinyin. The experimental results are shown in Tables 5 and 6.

Table 5. Results on OntoNotes 4 dataset.

Model	Precision	Recall	F1 Score
MECT [29]	-	-	82.57
MW-NER [27]	-	-	81.51
LEBERT [26]	-	-	82.08
KGNER [25]	-	-	82.10
FLAT+BERT [22]	-	-	81.82
AESINER [23]	-	-	81.18
StyleBERT [31]	-	-	79.10
PLTE+BERT [24]	79.62	81.82	80.60
SDI-NER [21]	82.14	78.39	80.22
Glyce [32]	81.87	81.40	81.63
ChineseBERT [34]	80.03	83.33	81.65
BERT-BiLSTM-CRF	80.51	82.23	81.34
CGR-NER(ours)	82.16	83.79	**82.97**

Table 6. Results on Weibo dataset.

Model	Precision	Recall	F1 Score
SDI-NER [21]	-	-	68.28
FLAT+BERT [22]	-	-	68.55
AESINER [23]	-	-	69.78
MECT [29]	-	-	70.43
MW-NER [27]	-	-	69.41
LEBERT [26]	-	-	70.75
KGNER [25]	-	-	71.90
StyleBERT [31]	-	-	69.60
GlyNN [33]	-	-	69.20
PLTE+BERT [24]	72.00	66.67	69.23
MFE-NER [30]	70.36	65.31	67.74
Glyce [32]	67.68	67.71	67.60
ChineseBERT [34]	68.27	69.78	69.02
BERT-BiLSTM-CRF	69.86	67.20	68.40
CGR-NER	70.23	71.17	**70.70**

For the OntoNotes 4 dataset, the precision, recall and F1 score of our model surpass the baseline (i.e., BERT-BiLSTM-CRF) by 1.65%, 1.56% and 1.63%, respectively, indicating that the semantic information contained in the Chinese character glyphs can significantly improve the recognition performance of the model. For the second-best model, MECT, which fuses the lexicon and the structural information of Chinese characters, our model surpasses it by 0.4% for the F1 score. In addition, compared with Glyce, which also utilizes a CNN to extract semantic information from the visual features of glyphs, our model significantly improves by 1.34% for the F1 score. We attribute this improvement to the fact that CGR-NER fully utilizes the potential word formation knowledge between adjacent glyphs and the interaction information between the character representation and glyph representation.

For the Weibo dataset, the precision, recall and F1 score of our model surpass the baseline by 0.37%, 3.97% and 2.3%, respectively. KGNER and LEBERT achieved the first and second highest F1 scores, respectively. We speculate that this is because the corpus of the Weibo dataset is collected from social media, and the corpus is messy and noisy, which makes the NER task more challenging. Meanwhile, KGNER and LEBERT supplement additional semantic information for the NER model with the help of a knowledge graph and lexicon, respectively, so as to have better performance in the Weibo dataset. It should be noted that in some special fields, such as industrial robot entity recognition, the performance of these models may be limited due to the lack of corresponding knowledge graphs or lexicons. However, CGR-NER still achieved the third highest F1 score without the help of external knowledge, which proves that our model has a strong entity recognition ability. At the same time, CGR-NER does not rely on external knowledge, so it can be applied to a wider range of scenarios.

4.5. Ablation Study

4.5.1. Effect of Model Hyperparameters

In order to explore the effect of the kernel size in CBHNN and the number of cross-modal attention layers in CMTE on the performance of CGR-NER, we select different parameter values and evaluate them. The experimental results are shown in Tables 7 and 8.

Table 7 shows the impact of different kernel sizes on model performance. It can be seen that three different kernel sizes have achieved better performance than BERT-BiLSTM-CRF, indicating that different kernel sizes have positive effects on model performance. However, the F1 score does not increase as the kernel size increases. One possible reason is that the small kernel size is more focused on extracting local details, so the glyph features have richer semantic information.

Table 7. Performance of CGR-NER under different kernel sizes.

Kernel Size	Weibo			OntoNotes 4		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
3	70.23	**71.17**	**70.70**	**82.16**	**83.79**	**82.97**
5	**71.35**	69.47	70.37	80.28	83.40	81.82
7	68.72	70.69	69.70	81.08	82.69	81.88
BERT-BiLSTM-CRF	69.86	67.2	68.4	80.51	82.23	81.34

Table 8. Performance of CGR-NER under different crossmodal attention layers.

Layer Size	Weibo			OntoNotes 4		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
1	70.21	69.95	70.08	81.56	83.77	82.65
2	**70.23**	**71.17**	**70.70**	**82.16**	**83.79**	**82.97**
3	70.01	69.66	69.83	79.98	83.65	81.77
BERT-BiLSTM-CRF	69.86	67.2	68.4	80.51	82.23	81.34

It can be seen from Table 8 that the three different layer sizes have achieved better performance than BERT-BiLSTM-CRF. When the number of layers is set to 2, the F1 score on the two datasets reaches the best value. When the number of layers increases to 3, the performance of CGR-NER does not continue to be improved. One possible reason is that the excessive number of layers may cause the model to over-fit or even learn noise, resulting in reduced recognition performance.

4.5.2. Effect of Model Components

We conducted experiments on the Weibo and OntoNotes 4 datasets to evaluate the impact of various components contained in CGR-NER on model performance. As shown in Table 9, after removing the Chinese character glyph ('-w/o Glyph') from CGR-NER, the F1 scores on the Weibo and OntoNotes 4 datasets are significantly reduced by 2.3% and 1.63%, respectively, which verifies the effectiveness of the Chinese character glyph in improving the recognition performance of our model. When the gate mechanism is removed ('-w/o GM') and the contextual representation and glyph representation after CMTE interaction are directly concatenated, the F1 scores on Weibo and OntoNotes 4 are decreased by 0.6% and 0.22%, respectively, which indicates that the importance of the representation from different modals varies, and the gate mechanism could effectively balance them. After removing the CMTE ('-w/o CMTE') from the fusion layer, the F1 scores on the Weibo and OntoNotes 4 datasets are reduced by 0.54% and 0.28%, respectively, which indicates that the interaction information between different modal representations could help the model to learn more semantic knowledge, thus improving the recognition performance of the model. When CMTE and the gate mechanism are removed at the same time ('-w/o CMTE+GM'), and the contextual representation is directly concatenated with the glyph representation, the F1 scores on the Weibo and OntoNotes 4 datasets are reduced by 0.85% and 0.55%, respectively, verifying the necessity of CMTE and the gate mechanism.

Table 9. Ablation experiment results on Weibo and OntoNotes 4 datasets.

Method	Weibo			OntoNotes 4		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
CGR-NER	70.23	71.17	70.70	82.16	83.79	82.97
−w/o Glyph	69.86 (−0.37)	67.20 (−3.97)	68.40 (−2.3)	80.51 (−1.65)	82.23 (-1.56)	81.34 (−1.63)
−w/o GM	70.38 (0.15)	69.83 (−1.34)	70.10 (−0.6)	81.82 (−0.34)	83.70 (−0.09)	82.75 (-0.22)
−w/o CMTE	70.27 (0.04)	70.05 (−1.12)	70.16 (−0.54)	81.90 (−0.26)	83.52 (−0.27)	82.69 (−0.28)
−w/o CMTE+GM	69.96 (−0.27)	69.76 (−1.41)	69.85 (−0.85)	81.96 (−0.2)	82.89 (−0.9)	82.42 (−0.55)

4.5.3. Effect of CBHNN

To verify the effectiveness of CBHNN in capturing the semantic information contained in Chinese character glyphs, we used glyph encoders with different structures to generate a glyph representation, and conducted experiments on the Weibo and OntoNotes 4 datasets, as shown in Table 10. Among them, 'character' represents BERT-BiLSTM-CRF, which only has a contextual representation as input. Tianzige is Tianzige-CNN proposed by Glyce [32]. Tianzige-BiLSTM represents that BiLSTM is used to capture the hidden state of Tianzige output, which is used as the glyph representation input of the model. CBHNN^{CNN} represents the CBHNN with no BiLSTM. Other settings of CGR-NER remained unchanged.

According to Table 10, the model based only on the contextual representation as input achieves the worst result because it ignores the semantic information contained in the Chinese character glyphs. Compared with Tianzige, the F1 scores of CBHNN^{CNN} on Weibo and OntoNotes 4 are improved by 0.6% and 0.34%, respectively, for the reason that the CBHNN^{CNN} can not only capture the semantic information in Chinese character glyphs, but also learns the potential word formation knowledge between adjacent glyphs through 3D convolution, thus enriching the semantic information of the representation. After Tianzige and CBHNN^{CNN} are combined with BiLSTM, respectively, the performance of the NER model is further improved, indicating that the context semantics and global dependency features in the glyph sequence are helpful for entity recognition. In contrast, CBHNN, which combines 3DCNN and BiLSTM, achieves the best F1 scores with 70.70% and 82.97% on the Weibo and OntoNotes 4 datasets, respectively. Hence, the experimental results show that CBHNN can effectively learn the semantic information contained in Chinese character glyphs, the potential word formation knowledge between adjacent glyphs and the context semantics and global dependency features of glyph sequences, thus improving the recognition performance of our proposed model.

Table 10. Result of CGR-NER with different glyph encoders.

Model	Weibo			OntoNotes 4		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
character	69.86	67.20	68.40	80.51	82.23	81.34
Tianzige	70.23	69.57	69.89	81.31	82.99	82.14
Tianzige-BiLSTM	70.40	69.83	70.11	81.75	83.44	82.58
CBHNN^{CNN}	70.22	70.66	70.49	81.76	83.22	82.48
CBHNN	70.23	71.17	70.70	82.16	83.79	82.97

4.5.4. Influence of Training Set Size

To verify the performance of CGR-NER under the condition of insufficient training data, we randomly select 10∼90% samples from the training sets of the Weibo and OntoNotes 4 datasets as the training set, respectively, while keeping the development set and test set unchanged, and we use BERT-BiLSTM-CRF as the baseline to compare with CGR-NER. As can be seen from Figure 5, the performance of CGR-NER surpasses that of BERT-BiLSTM-CRF under different training set sizes, indicating that CGR-NER has a stronger learning ability. In addition, we can find that CGR-NER has a more obvious improvement effect on the Weibo dataset. When the training data amount is 10%, the F1 score of CGR-NER on the Weibo dataset is improved by around 10% compared with BERT-BiLSTM-CRF. This is because the corpus of the Weibo dataset is collected from social media, the total amount of data is small, and the text format is chaotic, making entity recognition more challenging. At the same time, the glyphs of Chinese characters can be used as a data supplement to provide additional semantic knowledge for CGR-NER, so that CGR-NER can still have an excellent fitting ability in the case of insufficient training data.

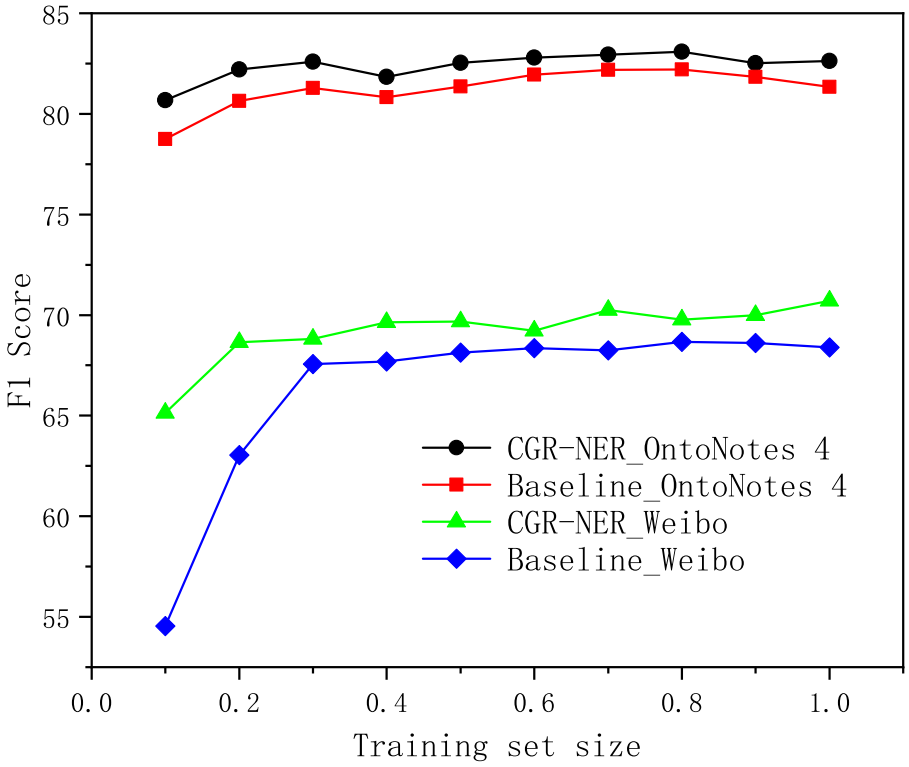

Figure 5. Performance of the model under different training set sizes. The training set size represents the percentage of training samples used in the original training set.

4.5.5. Influence of Out-of-Vocabulary Characters

To verify the recognition performance of the NER model under the condition of out-of-vocabulary characters, we divided the test set of each dataset into two parts: one containing out-of-vocabulary characters and the other containing no out-of-vocabulary characters. The experimental results are shown in Table 11. It can be found that the results of the model on the subset without out-of-vocabulary characters are better than those on the subset with out-of-vocabulary characters, indicating that it is still challenging for the NER model to recognize entities correctly in the presence of out-of-vocabulary characters. In addition, our CGR-NER outperforms BERT-BiLSTM-CRF, regardless of whether the subsets contain out-of-vocabulary characters. For the subset containing out-of-vocabulary characters, the F1 score of our model surpasses that of BERT-BiLSTM-CRF by 3.32% and 3.9% on Weibo and OntoNotes 4, respectively. One possible reason is that CGR-NER can infer the approximate

semantics of out-of-vocabulary characters by extracting the semantic information contained in the Chinese character glyph structure, so that the model can better understand the text semantics and improve the recognition performance.

Table 11. NER model performance on the test subset with and without out-of-vocabulary characters.

Model	Type	Weibo			OntoNotes 4		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
BERT-BiLSTM-CRF	-w/o OOV	69.39	69.69	69.48	81.71	82.85	82.26
	-w OOV	74.92	47.14	57.79	72.36	76.75	74.53
CGR-NER	-w/o OOV	70.70	72.27	71.43	82.33	83.71	83.01
	-w OOV	73.28	58.57	61.11	76.50	80.48	78.43

4.6. Case Study

Last but not least, we intuitively demonstrate the effectiveness of CGR-NER through two cases from the test set of the Weibo dataset. Table 12 shows the recognition results of CGR-NER and BERT-BiLSTM-CRF. In sentence 1, BERT-BiLSTM-CRF failed to identify the entity "萌萌 (Mengmeng)", but CGR-NER made a correct identification. In sentence 2, BERT only recognized the entity "哥 (brother)", but failed to recognize the entities "林日曦 (Lin Rixi)", "姗日曦 (Shan Rixi)" and "曉囈 (Xiao yi)", while CGR-NER correctly recognized all entities. We attribute the correct recognition of CGR-NER to the following reasons. First, entities such as "林日曦", "姗日曦", "曉囈" and "萌萌" have radicals such as "木 (wood)", "日 (sun)", "艹 (grass)" and "明 (bright)". The combination of these radicals often appears in Chinese naming, which is very consistent with the naming style of Chinese culture. Consequently, by capturing such features in Chinese character glyphs, CGR-NER can more accurately identify entities such as Chinese people and enterprise names. Secondly, although "曉" and "囈" belong to out-of-vocabulary characters, "曉" and "晓 (dawn)", "囈" and "呓 (balderdash)" have the same radical and a similar glyph structure, respectively. As a result, CGR can infer that they have similar semantics through their similar visual features, thereby reducing the impact of out-of-vocabulary characters on the model recognition performance and correctly identifying the entity "曉囈".

Table 12. Entity recognition results of NER model. Blue represents error recognition results.

Sentence 1	萌萌不是马哎哟小脸红红的 (Isn't Mengmeng? Oh, that little red face)
Gold label	萌(B-PER.NAM)萌(E-PER.NAM)不(O)是(O)马(O)哎(O)哟(O)小(O)脸(O)红(O)红(O)的(O)
BERT-BiLSTM-CRF	萌(O)萌(O)不(O)是(O)马(O)哎(O)哟(O)小(O)脸(O)红(O)红(O)的(O)
CGR-NER	萌(B-PER.NAM)萌(E-PER.NAM)不(O)是(O)马(O)哎(O)哟(O)小(O)脸(O)红(O)红(O)的(O)
Sentence 2	林日曦姗日曦，哥，新婚快乐呀。曉囈为 (Lin Rixi Shan Rixi, brother, happy wedding. Xiaoyi)
Gold label	林(B-PER.NAM)日(M-PER.NAM)曦(E-PER.NAM)姗(B-PER.NAM)日(M-PER.NAM)曦(E-PER.NAM)，(O)哥(S-PER.NOM)，(O)新(O)婚(O)快(O)乐(O)呀(O)。(O)曉(B-PER.NAM)囈(E-PER.NAM)为(O)
BERT-BiLSTM-CRF	林(O)日(O)曦(O)姗(O)日(O)曦(O)，(O)哥(S-PER.NOM)，(O)新(O)婚(O)快(O)乐(O)呀(O)。(O)曉(O)囈(O)为(O)
CGR-NER	林(B-PER.NAM)日(M-PER.NAM)曦(E-PER.NAM)姗(B-PER.NAM)日(M-PER.NAM)曦(E-PER.NAM)，(O)哥(S-PER.NOM)，(O)新(O)婚(O)快(O)乐(O)呀(O)。(O)曉(B-PER.NAM)囈(E-PER.NAM)为(O)

5. Conclusions

In order to make full use of the rich semantic information contained in Chinese character glyphs and improve the recognition performance of Chinese NER, this paper proposes a novel Chinese NER model named CGR-NER. In CGR-NER, a hybrid neural network

CBHNN is applied to capture the semantic information contained in Chinese character glyphs, the potential word formation knowledge between adjacent glyphs and the unique contextual semantics and global dependencies of glyph sequences. In addition, a fusion method with a crossmodal attention and gate mechanism is proposed to dynamically fuse the contextual representation from text and the glyph representation from images, so that the model can learn the complementary information between different modal representations. We conducted sufficient experiments on two mainstream Chinese NER datasets. The experimental results showed that CGR-NER achieved 70.70% and 82.97% F1 scores on the Weibo dataset and OntoNotes 4 dataset, which were increased by 2.3% and 1.63% compared with the baseline, respectively. At the same time, we conducted multiple groups of ablation experiments, proving that CGR-NER can still maintain good recognition performance in the condition of insufficient training data and out-of-vocabulary characters.

In addition, our model also has some limitations. Since CGR-NER carries out Chinese NER at the character level, it may lack the word-level features. In the future, we will consider how to more effectively integrate characters, glyphs and word-level information into the Chinese NER model, and continue to study the entity recognition effect of our model in specific domains to verify its robustness and generalization ability.

Author Contributions: Conceptualization, R.G. and T.W.; software, R.G.; validation, R.G. and J.D.; writing—original draft preparation, R.G. and J.D.; writing—review and editing, T.W. and L.C.; funding acquisition, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Program of NSFC-Guangdong Joint Funds (U2001201), Industrial core and key technology plan of ZhuHai City (ZH22044702190034HJL), the National Natural Science Foundation of China (Grant No. U1801263) and the National Key R&D Program of China (Grant No. 2019YFB1705500). Our work is also supported by Guangdong Provincial Key Laboratory of Cyber-Physical System (2020B1212060069).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nasar, Z.; Jaffry, S.W.; Malik, M.K. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.* **2021**, *54*, 1–39. [CrossRef]
2. Martins, P.H.; Marinho, Z.; Martins, A.F. Joint Learning of Named Entity Recognition and Entity Linking. *ACL* **2019**, *2*, 190–196.
3. Liu, Y.; Hashimoto, K.; Zhou, Y.; Yavuz, S.; Xiong, C.; Yu, P.S. Dense hierarchical retrieval for open-domain question answering. *arXiv* **2021**, arXiv:2110.15439.
4. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–70. [CrossRef]
5. Liu, P.; Guo, Y.; Wang, F.; Li, G. Chinese named entity recognition: The state of the art. *Neurocomputing* **2022**, *473*, 37–53. [CrossRef]
6. Ma, R.; Peng, M.; Zhang, Q.; Wei, Z.; Huang, X.J. Simplify the Usage of Lexicon in Chinese NER. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 July 2020; pp. 5951–5960.
7. Chiu, J.P.; Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]
8. Ma, X.; Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 1064–1074.
9. Kenton, J.D.M.W.C.; Toutanova, L.K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
10. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. Available online: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (accessed on 23 March 2023).

11. Lao, M.; Guo, Y.; Pu, N.; Chen, W.; Liu, Y.; Lew, M.S. Multi-stage hybrid embedding fusion network for visual question answering. *Neurocomputing* **2021**, *423*, 541–550. [CrossRef]
12. Gandhi, A.; Adhvaryu, K.; Poria, S.; Cambria, E.; Hussain, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **2023**, *91*, 424–444. [CrossRef]
13. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 1103–1114.
14. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory Fusion Network for Multi-view Sequential Learning. *arXiv* **2018**, arXiv:1802.00927.
15. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, VIC, Australia, 15–20 July 2018; pp. 2236–2246.
16. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Association for Computational Linguistics, Florence, Italy, 11 July 2019; Volume 2019; p. 6558.
17. Burger, J.D.; Henderson, J.; Morgan, W. Statistical named entity recognizer adaptation. In Proceedings of the COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002), Taipei, Taiwan, 31 August 2002; pp. 163–166.
18. Chen, W.; Zhang, Y.; Isahara, H. Chinese named entity recognition with conditional random fields. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, NSW, Australia, 22–23 July 2006; pp. 118–121.
19. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
20. Ali, W.; Kumar, J.; Xu, Z.; Kumar, R.; Ren, Y. Context-Aware Bidirectional Neural Model for Sindhi Named Entity Recognition. *Appl. Sci.* **2021**, *11*, 9038. [CrossRef]
21. Zhu, P.; Cheng, D.; Yang, F.; Luo, Y.; Huang, D.; Qian, W.; Zhou, A. Improving Chinese Named Entity Recognition by Large-Scale Syntactic Dependency Graph. *IEEE ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 979–991. [CrossRef]
22. Li, X.; Yan, H.; Qiu, X.; Huang, X.J. FLAT: Chinese NER Using Flat-Lattice Transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 June 2020; pp. 6836–6842.
23. Nie, Y.; Tian, Y.; Song, Y.; Ao, X.; Wan, X. Improving named entity recognition with attentive ensemble of syntactic information. *arXiv* **2020**, arXiv:2010.15466.
24. Mengge, X.; Yu, B.; Liu, T.; Zhang, Y.; Meng, E.; Wang, B. Porous lattice transformer encoder for Chinese NER. In Proceedings of the 28th International Conference on Computational Linguistics, Virtual, 8–13 December 2020; pp. 3831–3841.
25. Hu, W.; He, L.; Ma, H.; Wang, K.; Xiao, J. Kgner: Improving Chinese named entity recognition by bert infused with the knowledge graph. *Appl. Sci.* **2022**, *12*, 7702. [CrossRef]
26. Liu, W.; Fu, X.; Zhang, Y.; Xiao, W. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 5847–5858.
27. Liu, T.; Gao, J.; Ni, W.; Zeng, Q. A Multi-Granularity Word Fusion Method for Chinese NER. *Appl. Sci.* **2023**, *13*, 2789. [CrossRef]
28. Xu, C.; Wang, F.; Han, J.; Li, C. Exploiting multiple embeddings for Chinese named entity recognition. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 2269–2272.
29. Wu, S.; Song, X.; Feng, Z. MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 1529–1539.
30. Li, J.; Meng, K. MFE-NER: Multi-feature fusion embedding for Chinese named entity recognition. *arXiv* **2021**, arXiv:2109.07877.
31. Lv, C.; Zhang, H.; Du, X.; Zhang, Y.; Huang, Y.; Li, W.; Han, J.; Gu, S. StyleBERT: Chinese pretraining by font style information. *arXiv* **2022**, arXiv:2109.07877.
32. Meng, Y.; Wu, W.; Wang, F.; Li, X.; Nie, P.; Yin, F.; Li, M.; Han, Q.; Sun, X.; Li, J. Glyce: Glyph-vectors for Chinese character representations. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 2746–2757 .
33. Song, C.H.; Sehanobish, A. Using Chinese glyphs for named entity recognition (student abstract). In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34; pp. 13921–13922.
34. Sun, Z.; Li, X.; Sun, X.; Meng, Y.; Ao, X.; He, Q.; Wu, F.; Li, J. ChineseBERT: Chinese Pretraining Enhanced by Glyph and Pinyin Information. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; pp. 2065–2075.
35. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
36. Yan, H.; Deng, B.; Li, X.; Qiu, X. TENER: Adapting transformer encoder for named entity recognition. *arXiv* **2019**, arXiv:1911.04474.

37. Peng, N.; Dredze, M. Named entity recognition for Chinese social media with jointly trained embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 548–554.
38. Weischedel, R.; Pradhan, S.; Ramshaw, L.; Palmer, M.; Xue, N.; Marcus, M.; Taylor, A.; Greenberg, C.; Hovy, E.; Belvin, R.; et al. *LDC2011T03*; Ontonotes Release 4.0; Linguistic Data Consortium: Philadelphia, PA, USA, 2011.