TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 pp1–21
DOI: 10.26599/TST.2025.9010075
Volume x, Number x, xxxx xxxx

Research Article

CRE-RFGP: Relation Extraction for Chinese Medical Records

Qijie Qian, Xin Xu, Bo Li, Baoquan Ren, Yuan Miao, Yun Liu, Yong'an Guo*, and Shenqi Jing*

Abstract: The vast amount of available biomedical data has become a massive treasure trove of knowledge, offering significant potential for extracting valuable information. Extracting large-scale medical entities and relations from biomedical data is of great significance for constructing medical knowledge graphs, facilitating intelligent assistant diagnosis in the medical field, and enabling various other applications. Recent methods have achieved considerable progress, but there are still inherent limitations. First, these approaches extract entities or entity pairs without considering the relations between entities, which may lead to extracting redundant and/or incorrect entities. Second, they are unable to handle nested entities caused by the two-pointer method used for entity pair extraction. This is more challenging when dealing with Chinese medical records in which the occurrence of nested entities is more popular. Third, they encounter difficulties when subjects and objects are overlapping between different relations in a sentence. To tackle the above challenges, this paper proposes a novel relation extraction approach named CRE-RFGP. Specifically, it decomposes the relation extraction task into three components, including relation-first decoder, global entity extraction, and subject–object alignment. The decoder is utilized to predict and filter relations and restrict entity extraction based on predicted relations. A relation-specific attention mechanism and a global pointer network are employed to effectively handle the problem of information overlapping and object/subject nesting. An entity correspondence matrix is introduced to align subjects, objects, and their relations into triples. The matrix not only reduces the complexity of relation extraction but also plays an important part in improving the precision of extracted triples. Comprehensive evaluations are conducted experimentally based on two Chinese medical datasets. Comprehensive experiments on two Chinese medical datasets demonstrate that CRE-RFGP outperforms eight state-of-the-art approaches.

Key words: relation extraction; entity extraction; Chinese medical record; relation-first; global pointer; subject–object alignment

1 Introduction

In medical society, a vast amount of biomedical data is available nowadays. Those data have become a massive treasure trove of knowledge, offering significant support for downstream applications like

• Qijie Qian, Xin Xu, Baoquan Ren, and Yong'an Guo are with School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China. E-mail: 2022010311@njupt.edu.cn; xuxincd88@126.com; renbq88@126.com; guo@njupt.edu.cn.
• Bo Li and Yuan Miao are with College of Arts, Business, Law, Education, and Information Technology, Victoria University, Melbourne 3011, Australia. E-mail: bo.li@vu.edu.au; yuan.miao@vu.edu.au.
• Yun Liu is with the Center of Data Management, The First Affiliated Hospital, Nanjing Medical University, Nanjing 210029, China. E-mail: liuyun@njmu.edu.cn.
• Shenqi Jing is with Department of Medical Informatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, China. E-mail: jingshenqi@njmu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2024-08-22; revised: 2025-04-22; accepted: 2025-04-23

smart healthcare, intelligent medical diagnosis, and medical consultations[1]. A great portion of such biomedical data is textual data[2], which serve as foundational resources to build up medical knowledge graphs to serve downstream applications[3]. However, those textual data are highly unstructured[4], and need to be pre-processed. Recently, how to extract medical entities and corresponding relations between entities has attracted massive attention from academia[5–7].

Extracting entities and relations from medical records is more challenging than common relation extraction[8]. That is because medical relation extraction mainly contains a large number of professional terms, abbreviations, and specific grammatical structures in this special field. At the same time, the extraction of relations in the medical field needs to take into account the complex relations between different entities, such as patients and diseases, symptoms and diseases, drugs and diseases, etc. Unfortunately, due to the inherent complexity of the Chinese language, it is more challenging to conduct relation extraction for Chinese medical records[9]. For example, a sentence in Chinese has no boundary between characters and/or words. The different granularity of Chinese input can have a significant impact on the model. Considering the large volume of medical records to be processed, manually processing them to extra entities and relations is infeasible, i.e., it is too time-consuming and economically burdensome. Recently, researchers began to employ natural language processing (NLP) techniques to identify entities and relations from unstructured texts[10, 11]. Compared with previous methods, there is a substantial saving of human and material resources with this approach.

Relation and its relevant entities can be presented in the form of triples {subject, relation, object}, in which the subject entity possesses a specific relation towards the object entity. As shown in Fig. 1, the given Chinese medical record (one sentence) contains four triples, like {Type 2 diabetes, case classification, diabetes}. In practice, a piece of medical record usually contains multiple relation triples. More importantly, many relation triples may share the same entity. For example, the entity Type 2 diabetes is involves in all four triples shown in Fig. 1. It is the subject entity in the first triple and is the object entity in the rest three triples. The situation that multiple triples have the same entity is referred to as entity overlap[12]. In addition, characters

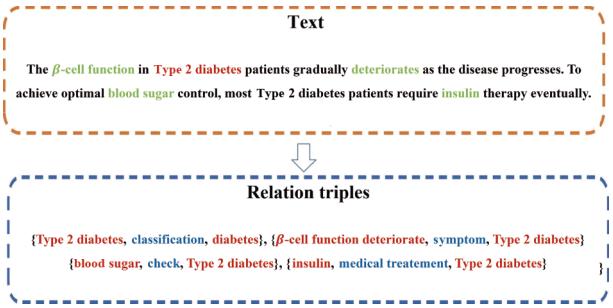


Fig. 1 An example of relation triple extraction for Chinese medical record. The English sentences and triples were translated from Chinese via Google Translate for ease of understanding.

of one entity can be a substring of another entity, e.g., multiple different relations. For example, the entity diabetes is part of entity Type 2 diabetes. This is referred to as entity nesting[7].

Broadly, modern methods of NLP-based relation extraction can be classified into two categories: Pipeline-based methods and joint-based methods. The pipeline-based methods perform two independent and sequential tasks to extract relation triples[13], including NER[14, 15] and relation classification[16]. Usually suffering from error propagation[14, 17], such methods encounter challenges when performing relation classification after recognizing all entities in the text, e.g., errors in entity recognition would propagate to relation classification and undermine the performance[17]. Moreover, the pipeline-based models overlooked the interaction between the two tasks[18]. In contrast, joint extraction models seek to extract entities and relations simultaneously in an end-to-end manner[12, 19, 20]. They decompose the entire tasks into several subtasks and solve the problems through multi-task learning frameworks[21, 22].

Despite these advancements, existing relation extraction methods still face key challenges in practical applications, particularly in Chinese medical record processing. These records contain specialized domain knowledge, involve complex sentence structures, and often exhibit overlapping and nested entities, making conventional relation extraction techniques less effective. Furthermore, the high prevalence of noisy and incomplete data in real-world clinical settings exacerbates these challenges, demanding a more robust and adaptable approach.

Recently, such methods have achieved attractive performance but still have drawbacks. First, in general, the quantity of entities in a record is much greater than

that of relations. If entities are extracted first, there will be many incorrect and redundant combinations of entities[11]. For example, CasRel[22] employs a span-based method to extract the subject entity and then extract the object entity based on potential relations. However, performing relation classification by mapping every relation to every potential entity pair would result in massive redundant and incorrect triples[23]. In addition, CasRel's span-based extraction focuses solely on the start and end positions of entities, leading to an inability to handle nested entities[7]. Moreover, its subject–object alignment mechanism is inefficient since it can only process one subject at a time[12]. Second, span-based entity extraction only focuses on the start and end positions of entities without understanding the underlying semantics. This results in difficulties in handling nested entities[7]. For example, PRGC[12] uses a two-pointer method to extract subject and object entities during model decoding. Similar to CasRel, it still focuses only on the start and end positions of entities and thus could not handle nested entities. Similarly, OneRel[7] uses a score-based classifier and relation-specific horns strategy to extract triples but suffers from extracting redundant entities. Third, regarding the subject–object alignment, many methods adopt the nearest neighbour principle based on heuristic algorithms or design complex decoders to combine entity pairs, which leads to high complexity and low efficiency[12]. For example, TPLinker[16] uses a complex decoder to evade exposure bias in subject–object alignment, but this results in sparse labels and low convergence rates.

To address these challenges, this paper proposes a novel relation extraction framework, CRE-RFGP, tailored specifically for Chinese medical records. CRE-RFGP introduces three key innovations as below.

(1) Relation-first decoder: Instead of utilizing all potential relations, CRE-RFGP selects a smaller set of useful relations with this decoder. Then, it uses selected relations to extract entities. This not only significantly reduces the search space but also extracts triples more accurately. As a result, lower computational complexity and better performance can be achieved. It is more useful when dealing with medical big data which contains numerous relations and entities, such as Chinese medical datasets.

(2) Global entity extraction: CRE-RFGP uses a relation-specific attention mechanism to gain sentence representations for each specific relation, allowing the utilization of semantic information. Then, CRE-RFGP combines it with a global pointer network, which is a relation-specific sequence tagging component[7], to extract entities, respectively. This helps tackle overlapping entities and nested entities.

(3) Subject–object alignment: CRE-RFGP employs a relation-independent entity correspondence matrix to perform the subject–object alignment. The matrix possesses correspondence scores between all potential subjects and objects. By examining this matrix, CRE-RFGP can find the correct entity pairs with high efficiency.

The main contributions of this paper are as follows:

- We propose CRE-RFGP, a novel relation triple extraction framework for handling Chinese medical records. Different from the aforementioned approaches, CRE-RFGP adopts three unique components to alleviate existing problems, improve extraction accuracy, and improve efficiency.
- We make a big step to solve the entity overlap and entity nesting problems. More specifically, CRE-RFGP combines a relation-specific attention mechanism with a global pointer network. The global pointer treats all characters in entity boundaries as a whole for discrimination and extraction. The attention mechanism further helps distil semantic information from those characters.
- We prototype CRE-RFGP and conduct comprehensive experiments based on two public Chinese medical datasets. Compared with eight state-of-the-art approaches, CRE-RFGP has the best performance in both accuracy and efficiency. Moreover, CRE-RFGP has more significant advantages over the other approaches when a sentence has multiple relation triples. This demonstrates that CRE-RFGP is more useful in practice.

The rest of this paper is organized as follows. Section 2 reviews representative studies published in recent years. Section 3 provides technical details of our CRE-RFGP framework. Section 4 experimentally evaluates the effectiveness of CRE-RFGP against eight state-of-the-art approaches based on two Chinese medical datasets. Finally, Section 5 concludes this paper and points out future work.

2 Related Work

2.1 General relation extraction

In terms of relation extraction, early research adopted

the pipeline-based process which consists of two independent steps: Using an entity extraction model to identify entities and then using a relation classification model to identify relations for identified entities[24, 25]. For instance, utilizing a long short-term memory network (LSTM) within a multi-channel recurrent neural network, SDP-LSTM[24] captures heterogeneous information along the shortest dependency path (SDP) between two entities. Then, the model identifies subject and object entities and proceeds to classify their relation. Nevertheless, the correlation between these two steps is disregarded by SDP-LSTM, leading to the propagation of errors. To tackle this issue, researchers have proposed joint models that enable the simultaneous extraction of entities and relations[26, 27].

Notably, these joint models usually have superior performance compared with the pipeline methods[28]. Traditional joint extraction methods are feature-based[18, 29–31]. For example, Xing et al.[30] proposed a method utilizing unsupervised word embeddings and sentence embeddings to identify phenotype features in biomedical data. Their method employs a dictionary-based and rule-based approach to identify genes. However, it heavily relies on manual feature engineering, requiring a significant amount of labour work. To reduce this manual effort, recent methods primarily rely on neural networks, which can be broadly categorized into table-filling methods, tagging methods, and Seq2Seq methods.

- **Table filling methods.** Methods in this category construct a table for each relation, where each cell in the table typically represents the start and end positions of the two entities (or even their types) that have a particular relation[16, 29]. An example of recent work following this approach is OneRel[32]. OneRel proposes a single-module, single-step decoding method for the joint extraction of entities and relations. It directly identifies triples and captures the interdependence among them. Due to the sparsity exhibited by the table, a significant number of redundant predictions are encountered by these methods, resulting in inefficiencies during the extraction process.
- **Tagging methods.** Well-designed strategies are typically employed by methods in this category to establish connections between entities and relations. To name a few, relations in the sentence are modelled as functions that map subjects to objects by CasRel[22]. However, it encounters difficulties in handling relation redundancy. To address this limitation, proposed as an improved cascaded tagging framework, FastRE[33] incorporates a mechanism for mapping types to relations, thereby enhancing tagging efficiency and reducing relation redundancy. Additionally, a position-dependent adaptive threshold strategy is employed to enhance tagging accuracy. Similarly, SAHT[34] uses a subject attention mechanism that relates to subject features to create a specific representation for each subject. Then, it utilizes an object multi-relation tagger to extract objects and relations based on these representations. Nevertheless, these methods can not handle nested and overlapping entities.
- **Sequence-to-sequence (Seq2Seq) methods.** Methods in this category attempt to generate all triples sequentially. Seq2Rel[3] employs a novel linearization pattern to address complexities that were overlooked by previous Seq2Seq methods, such as core relations and *n*-ary relations. REBEL[2] is an autoregressive approach that formulates relation extraction as a Seq2Seq task and retrains the encoder-decoder transformer (BART) using a new dataset. However, these methods still suffer from challenges related to relation redundancy and low efficiency for subject–object alignment.

2.2 Relation extraction for Chinese medical records

Recently, many efforts were made to solve the relation extraction problems for Chinese medical data. One notable study is the knowledge-guided distance supervision (KGDS) model proposed by Zhao et al.[35]. This model adopts entity-type alignment instead of the traditional entity alignment approach to extract coarse-grained relations. Furthermore, Zhao et al.[35] enhanced the model by incorporating knowledge from both the knowledge base and dataset of electronic medical records to address the challenge of relation ambiguity. Another notable approach is the attention-based model introduced by Zhang et al.[36]. This model leverages a multi-hop attention mechanism to extract diverse semantic information from Chinese medical records. By generating multiple-weight vectors for each sentence using an attention mechanism, the model enables the generation of distinct semantic representations for each sentence. Li et al.[6] proposed a document-level relation extraction model that learns relevant semantic information between document titles, abstracts, and SDPs through cross-attention mechanisms. To differentiate the importance of entities

in different sentences, they calculated the weights of occurrence sentences and adjacent entity sentences using a Gaussian probability distribution. Unfortunately, the above approaches could not handle redundant predictions and nested entities.

Zhao et al.[37] proposed a unified multi-task learning framework to solve the relation redundancy problem. A type attention mechanism was employed to extract the subject, which provided explicit prior type information. Then, a subject-aware relation prediction was introduced, selecting relevant relations by considering both global and local semantics. Finally, a question-generated question answering (QA) method was utilized to extract the object in an object-driven manner[38]. While this method successfully addresses the problem of relation redundancy, the subject–object alignment process is overly complex and inefficient. Gao et al.[39] proposed ERGM, a multi-stage joint entity and relation extraction model with global entity matching, to improve entity-pair alignment. However, its multi-stage extraction and matching process still increases the complexity of subject–object alignment and may introduce redundant relation judgments. PRGC[12] employs a double-pointer method for entity extraction but fails to handle nested entities. TPLinker[16] utilizes a token-to-entity linking scheme to facilitate entity extraction and align the subject entity with the object entity based on a given relation. However, TPLinker still suffers from redundant relation judgments, over-complex subject–object alignment, and limitations associated with span-based extraction schemes.

Different from existing studies, our CRE-RFGP prioritizes relation decoding to solve the entity redundancy problem. Then, it uses a relation-specific attention mechanism and a global pointer network to handle overlapping and nested entities. In light of the discoveries indicating that relations are commonly triggered by context rather than entities, inspiration was drawn[8, 12], CRE-RFGP uses a novel entity alignment mechanism to generate triplets more efficiently.

3 Approach

In this section, we formally define the relation extraction problem and then provide detailed explanations of our CRE-RFGP model. A general structure of CRE-RFGP is shown in Fig. 2.

3.1 Problem definition

For the task of relation triple extraction, the input is ([CLS], X, [SEP]), in which $X = (x_1, x_2, \ldots, x_n)$ is a sentence consisting n tokens, and [CLS] and [SEP] are two special tags indicating the start and end of the sentence, respectively. The tokens corresponding to these two special tags are represented as x_{cls} and x_{sep}, respectively. Considering a set of predefined relation types denoted as R, the objective is to extract all potential triples represented as $T(X) = (e_i, r_{ij}, e_j)$, where e_i and e_j are character sequences that represent the subject and object, respectively, and $r_{ij} \in R$ is a relation in R between e_i and e_j. We define the three key components of CRE-RFGP as below.

Relation-first decoder. With a non-autoregressive decoder, the module predicts potential relations and employs a binary classifier to eliminate irrelevant relations. For the given sentence X, the output of this component is $Y_r(X) = \{r_1, r_2, \ldots, r_m | r_i \in R\}$, in which Y_r is the predicted subset of relation, r_1 is a candidate relation, m is the size of the subset of candidate relations, and $|R|$ is the total number of relations.

Global entity extraction. Given a sentence X and a predicted candidate relation r_i, this component uses a relation-specific attention mechanism to measure the importance of each Chinese character to r_i, and obtain the sentence representation S_r for r_i. Then, CRE-RFGP uses the global pointer network to distinguish the start and end positions of the entity in the sentence. Note that it treats all characters between the start position and the end position as a whole. With S_r as input, the global pointer network outputs $Y_e(S_r) = \{(e_1, r_{12}, e_2), (e_2, r_{23}, e_3), \ldots, (e_i, r_{ij}, e_j)\}$, where Y_e is a set of candidate triples, e_i is the subject, e_j is the object, r_{ij} is the relation between subject and object. We call (e_i, e_j) a candidate entity pair.

Subject–object alignment. Given a sentence X, this component predicts the correspondence scores between the subjects and objects. A subject–object pair existing in the correct triple will receive a higher score, while other pairs have lower scores. Let M represent the global correspondence matrix. The output of this component is $Y_s(X) = \{M \in R^{n \times n}\}$, in which Y_s is the set of entity correspondence matrices, and M is an entity correspondence matrix with a size of $n \times n$.

3.2 Encoder

The encoding module aims to extract textual features

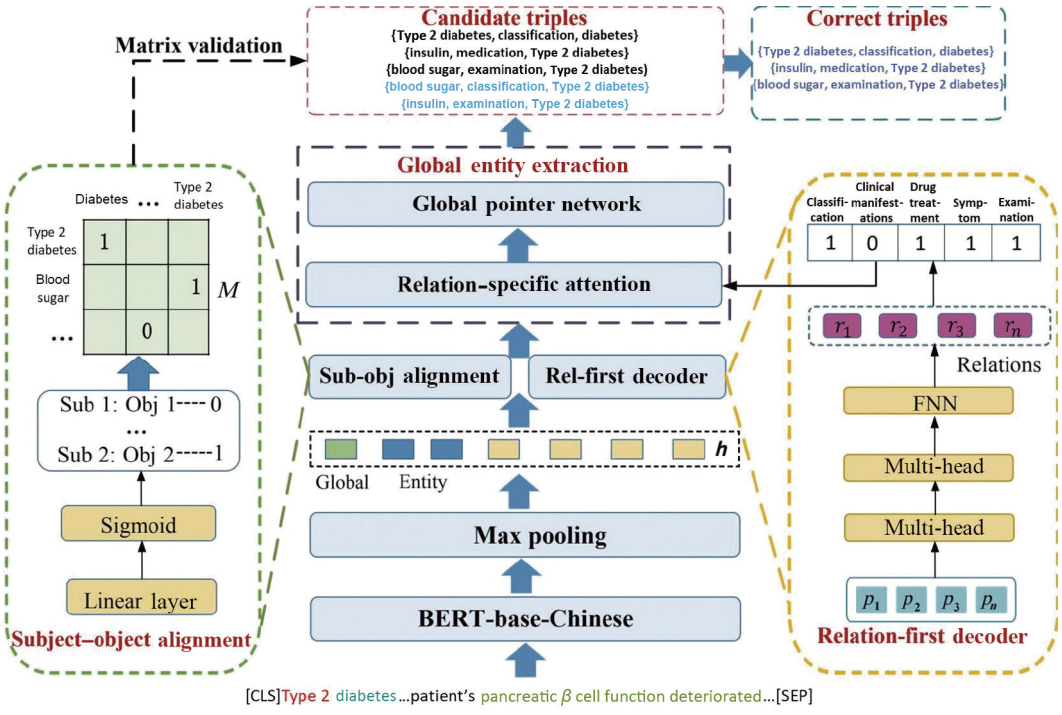


Fig. 2 General structure of CRE-RFGP. Given a sentence in Chinese medical records, predicting a subset of potential relations and a global correspondence matrix that signifies the alignment between subjects and objects is accomplished by CRE-RFGP. Then, it applies a relation-specific attention mechanism to each potential relation in the sentence and uses a global pointer network for entity extraction. Next, it enumerates all candidate triples, e.g., five triples in this example. Finally, it applies the constraint of global correspondence and identifies three correct triples. Sub denotes subject, Obj denotes object, Rel denotes relation, and FNN denotes feedforward neural network.

from the sentence and use these features as the input of the decoding module. The pre-trained language model BERT-base-Chinese[40] is used as the encoder. The input vectors of the BERT-base-Chinese are character embedding[41] and position embedding[41] of each character in the sentences. Note that CRE-RFGP deals with one single sentence each time; therefore, the sentence segmentation embedding is not needed. It can be easily iterated to handle medical records with multiple sentences. The deep semantic representation of the sentence can be obtained by the input vector through a multi-level Transformer[40]. The transformer is represented here as Trans(x), where x represents the input vector. The state vector of the encoded sentence can be represented as $h^{(\ell)}$, calculated as Eqs. (1) and (2),

$$h_0 = S\,W_s + W_p \tag{1}$$

$$h^{(\ell)} = \text{Trans}(h^{(\ell-1)}), \quad \ell = 1, 2, \ldots, N \tag{2}$$

where ℓ denotes the index of the Transformer layer. The matrix S represents the one-hot vectors[41] of the input sentence, representing the attention mask of the sentence, W_s is the word embedding matrix. The position embedding matrix W_p and the hidden layer's state vector $h^{(\ell)}$ are represented in the given context. The sentence semantic representation is obtained by using the state vector from the last hidden layer of BERT-base-Chinese. However, it could be extended to other encoders, such as RoBERTa[42, 43].

3.3 Relation-first decoder

As shown in Fig. 2, the relation-first decoder is shown in the yellow box, where $\{p_1, p_2, \ldots, n\}$ are all the potential relations. Previous works performed entity extraction for all relations[11, 39]. In contrast, our method differs from theirs as we first predict a subset of candidate relations in the sentence and then only extract entities based on these target relations. Via a non-autoregressive decoder, the model makes predictions of potential relations and employs a binary classifier to eliminate irrelevant relations.

Potential relation extractor. CRE-RFGP uses a transformer-based decoder to predict potential relations[5]. Initialized with a learnable embedding $Q \in \mathbb{R}^{n_q \times d}$ of n_q, the decoder's input is based on the

maximum number of relations in the sentence, and d is the dimension of the vector. Given the state vector $h_\alpha \in \mathbb{R}^{n_q \times d}$ of a sentence with n tags and an output vector $H^r \in \mathbb{R}^{n_q \times d}$, the predicted relation is gained from Eqs. (3) and (4),

$$H_\alpha^{\text{avg}} = \text{Avgpool}(h_\alpha) \tag{3}$$

$$p_i = \text{Softmax}(W_r H_\alpha^{\text{avg}} + b_r) \tag{4}$$

where Avgpool(·) is the average pooling operation[18], $W_r \in \mathbb{R}^{|R| \times d}$ and $b_r \in \mathbb{R}^{|R|}$ are learnable parameters. The total number of relation types is represented by $|R|$. $\{r_1, r_2, \ldots, r_n\}$ are the obtained relations. h_α denotes the hidden-state representation output by the α-th Transformer layer, and H_α^{avg} denotes the sentence-level representation obtained by applying average pooling to h_α. In Eq. (4), p_i denotes the predicted probability of the i-th relation type, Softmax(·) is used to normalize the output into a probability distribution over all relation types, and ($W_r \in \mathbb{R}^{|R| \times d}$) and ($b_r \in \mathbb{R}^{|R|}$) are learnable parameters. Here, d is the dimension of the hidden vector. $r_1, r_2, \ldots, r_n$ are the obtained candidate relations.

Candidate relation assessment. Now, CRE-RFGP filters out irrelevant relations to effectively generate candidate relations. From the output vector matrix H^r of the decoder and the embedding of [CLS]. This component generates the set of candidate relations by predicting a relation confidence vector $\boldsymbol{v}^{\text{rel}}$ using the binary classifier.

$$\boldsymbol{v}^{\text{rel}} = \sigma(\text{TW}_s[H^r; x_{\text{CLS}}] + b_s) \tag{5}$$

where TW_s is trainable weight, b_s is a bias, and $\sigma(\cdot)$ is the sigmoid function[12]. A greater value obtained from the activation function indicates higher confidence that the relation is included in the sentence, and vice versa. We model the task of candidate relation assessment as a multi-label binary classification task. If the probability exceeds the threshold λ_1, label 1 will be assigned to the corresponding relation, otherwise, label 0 will be assigned. Irrelevant relations can be filtered out and a subset $R_n = \{r_1, r_2, \ldots, r_n\} \in \mathbb{R}$ is predicted while discarding most of the negative samples.

3.4 Global entity extraction

Relation-specific attention mechanism. Each Chinese character in sentence $X = (x_1, x_2, \ldots, x_n)$ holds different weights under different relations. We propose a relation-specific attention mechanism to allocate disparate weights to contextual characters for each relation. Attention scores are obtained as

$$e_{ik} = V^{\text{T}} \tanh(W_r r_k + W_\alpha h_\alpha + W_x x_i) \tag{6}$$

$$\mu_{ik} = \frac{\exp(e_{ik})}{\sum_{i=1}^{n} \exp(e_{ik})} \tag{7}$$

where r_k is the trainable embedding for the k-th relation, and V, W_r, W_α, $W_x \in \mathbb{R}^{|R| \times d}$ are transformation matrices. This approach allows the attention score to gauge both the importance of each character for relational expression and its contribution to the overall sentence. Then, the weighted sum of every individual character generates the sentence's representation s_k for a specific relation type r_k,

$$s_k = \sum_{i=1}^{n} \mu_{ik} x_i \tag{8}$$

Global pointer. Figure 3 provides an example of the global pointer used by CRE-RFGP. Assuming that CRE-RFGP needs to identify four entities in a sentence: {disease: Diabetes}, {disease: Type 2 diabetes}, and {drug: insulin}. In the matrix, the positions with a value of 1 represent the positions of entities. Specifically, the row number in the matrix represents the start position of an entity, and the column number indicates the end position of an entity. This way, the global pointer identifies the start and end position of the entity as a whole, without separately recognizing the start and end.

The global pointer network implements entity recognition via two steps: Extraction and classification. Extraction refers to extracting entity spans, while classification involves determining the type of each entity. For simplicity, CRE-RFGP adopts a single entity type recognition strategy and iterates the process to handle multiple entity types. It employs a dot-product attention mechanism[5] and Relative Position Encoding (RoPE)[22] to generate the scoring matrix shown in Fig. 2. The introduction of RoPE relative positions is to increase sensitivity to entity length and span, enabling better differentiation of nested entities.

is h_i, the decoding vectors $q_{i,c}$ and $k_{i,c}$ are as

$$q_{i,c} = W_c^q h_i + b_c^q \tag{9}$$

$$k_{i,c} = W_c^k h_i + b_c^k \tag{10}$$

where $q_{i\alpha}$ and $k_{i\alpha}$ are new vectors obtained from each vector in the sentence vector sequence through different linear transformations, W_c^q and W_c^k are linear

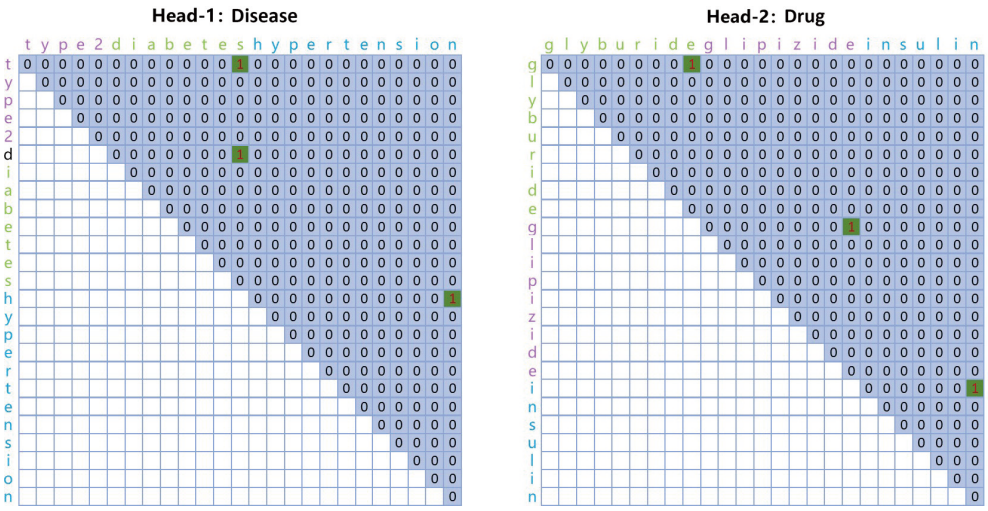


Fig. 3 An illustration of using a global pointer network to extract a diagram of entities specific to a certain attribute.

transformation matrices, i represents the position of the i-th Chinese character in the sentence after linear transformation, where c denotes the index of the entity type, it represents the c-th attribute in the attribute set, and b_c^q and b_c^k are trainable parameters. Based on the above, we can obtain two decoding vectors $q_c = [q_{1,c}, q_{2,c}, \ldots, q_{n,c}]$ and $k_c = [k_{1,c}, k_{2,c}, \ldots, k_{n,c}]$. Finally, a bilinear transformation is applied to score the i-th row and the j-th column in the c-th matrix, as

$$s_c(i,j) = q_{i,c}^{\mathrm{T}} k_{j,c} \tag{11}$$

where $s_c(i,j)$ represents the score of the entity with attribute c for the continuous segment in the sentence from position i to position j. A score greater than 1 indicates that the extracted entity pair is a potentially correct entity pair.

3.5 Subject–object alignment

After the operation of the global pointer network, CRE-RFGP obtains all possible subjects and objects. Then, by utilizing the global correspondence matrix[7], the determination of the correct pairs of subject and object is enhanced. The detailed process is as follows. First, CRE-RFGP exhaustively enumerates all potential subject–object pairs. Second, it checks the corresponding scores for each pair in the global matrix, keeping the value if it exceeds a certain threshold λ_2, otherwise filtering it out.

Given a sentence, as shown by the green matrix M in Fig. 2, the 2D global corresponding matrix has a size of $n \times n$. Each entry in the matrix represents both the start position of the paired subject–object entities and the confidence level of this pair. A higher value indicates a higher probability that the corresponding entity pair is part of a correct triple. For example, if the value for diabetes and Type 2 diabetes is high in the first row and first column, it indicates that they are part of the correct triple, such as {Type 2 diabetes, diabetes}. The value at each position in the matrix is obtained as

$$P_{i_{\mathrm{sub}},j_{\mathrm{obj}}} = \sigma\left(\mathrm{TW}_g\left[h_i^{\mathrm{sub}}; h_j^{\mathrm{obj}}\right] + b_g\right) \tag{12}$$

Forming a potential subject–object pair, h_i^{sub} and $h_j^{\mathrm{obj}} \in \mathbb{R}^{d \times n}$ represent the encoded representations of the i-th and j-th tokens in the input sentence, respectively. $\mathrm{TW}_g \in \mathbb{R}^{2d \times n}$ denotes a trainable weight, and σ represents the sigmoid function[12].

3.6 Loss function

In the optimization process, the model is jointly trained in a multi-tasking manner, with the parameters of the encoder shared among the tasks. However, for the relation-first decoder, there is no suitable way to rank those potential relations. Therefore, we adopt binary matching loss[44]. To find the best match between the ground truth and predicted relations, we seek the ranking strategy with the lowest loss Π^*.

$$\Pi^* = \arg\min_{\Pi \in \Pi(n_q)} \left(-\sum_{i=1}^{n_q} (I(y_i^r) \cdot P_{\Pi(i)}^r(y_i^r)) \right) \tag{13}$$

where $P_{\Pi(i)}^r$ represents the probability distribution of the $\Pi(i)$-th relation in the predicted relations, $\Pi(n_q)$ represents the space of all ranking strategies, and y_i^r represents the ground truth relations. $I(y_i^r)$ is a switch function: If $y_i^r \neq \varnothing$, $I(y_i^r) = 1$, otherwise 0. We define the loss of the relation-first decoder as

$$\mathcal{L}_{\text{rel}} = -\sum_{i=1}^{n_q} \log P^{r}_{\Pi^*(i)}(y^r_i) \tag{14}$$

where $P^{r}_{\Pi^*(i)}$ represents the probability distribution of the $\Pi(i)$-th relationship in the predicted relations under the ranking strategy that minimizes the loss. To predict entity pairs, the cross-entropy loss[45] is used as the loss function of global entity extraction.

$$\mathcal{L}_{\text{entity}} = \ln\left(1 + \sum_{(i,j)\in P_c} e^{-s_c(i,j)}\right) + \ln\left(1 + \sum_{(i,j)\in Q_c} e^{s_c(i,j)}\right) \tag{15}$$

$$\mathcal{L}_{\text{global}} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[y_{ij}\ln P_{i_{\text{sub}}j_{\text{obj}}} + (1-y_{ij})(\ln P_{i_{\text{sub}}j_{\text{obj}}})\right] \tag{16}$$

where $\mathcal{L}_{\text{entity}}$ denotes the loss function of global entity extraction, $\mathcal{L}_{\text{global}}$ denotes the loss function of the global correspondence matrix, y_{ij} is the confidence score of whether the i-th entity and the j-th entity belong to the same triple, P_c is the position for all entities of type c in the sample, Q_c is the position for all entities of type non-c in the sample, and $P_{i_{\text{sub}}j_{\text{obj}}}$ is the value of each position in the corresponding matrix of the entity. In Eqs. (14)–(16), denotes the natural logarithm with base e. The total loss $\mathcal{L}_{\text{total}}$ is the sum of these three parts.

$$\mathcal{L}_{\text{total}} = \omega_{\text{rel}}\mathcal{L}_{\text{rel}} + \omega_{\text{entity}}\mathcal{L}_{\text{entity}} + \omega_{\text{global}}\mathcal{L}_{\text{global}} \tag{17}$$

Performance could be further improved by carefully adjusting the weight of each sub-loss, but for simplicity, we assign equal weights to the three parts, i.e., $\omega_{\text{rel}} = \omega_{\text{entity}} = \omega_{\text{global}} = 1$.

4 Experimental Evaluation

We experimentally evaluate the effectiveness CRE-RFGP against state-of-the-art approaches in this section.

4.1 Experimental settings

4.1.1 Datasets

We conduct experiments on the CMeIE[4] and DiaKG[46] datasets. The CMeIE dataset is obtained from the CHIP2020 evaluation task, which includes training corpora for pediatrics and 100 common diseases. The training corpus for pediatrics consists of 504 pediatric diseases, while the training corpus for common diseases includes 109 common diseases. In total, the dataset consists of nine categories of medical entities. Additionally, the dataset also consists of 44 types of relations. The numbers of entities and relation types are presented in Table 1. Due to the large size of the CMeIE dataset, considering the resource constraints of the machine and the training time, we only extract some of the datasets in the experiment, including the training set (5000) and the test set (1200).

The DiaKG dataset is derived from 41 published diabetes guidelines and consensuses. It covers a wide range of research topics and hot areas, including basic research, clinical research, and so on. It has been considered as an authoritative resource for building the diabetes knowledge base. This dataset defines a total number of 18 entity types and 15 medical relations, containing 1892 sentences, 22 050 entities, and 6890 relations. It is currently the first annotated dataset related to diabetes in Chinese. The definitions and details of relation types are shown in Table 2.

4.1.2 Evaluation metrics and comparison models

Following previous work[11, 12, 16, 32, 39], we consider an extended relation triple to be correct only when it precisely matches the ground truth, which means that the subject, object, and the relation between the entity are all matched. Meanwhile, we report the standard precision, recall, and F1-score for all competing models. The definitions of these three metrics are as

$$\text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{18}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{19}$$

Table 1 Statistics of the CMeIE dataset.

Statistic/category		Training	Test
Number of relations		44	44
Number of sentences		17 924	5602
Number of tuples		54 286	17 512
Entity overlap type	Normal	6931	2116
	EPO※	1572	268
	SEO†	10 993	3486
Number of tuples in a sentence	1	6713	2036
	2	3711	1147
	3	2304	699
	4	1635	494
	⩾ 5	3561	1223

※Entity pair overlap (EPO) refers to the existence of more than one relationship between two entities.

†Single entity overlap (SEO) refers to an entity being involved in two relation triples simultaneously.

Table 2 Statistics of the DiaKG dataset. ADE denotes adverse drug event.

Relation	Number
Test_items_Disease	1171
Anatomy_Disease	1072
Drug_Disease	1315
Method_Drug	185
Treatment_Disease	354
Pathogenesis_Disease	130
Test_Disease	271
Operation_Disease	37
Class_Disease	854
Reason_Disease	164
Duration_Drug	61
Symptom_Disease	283
Amount_Drug	195
ADE_Drug	693
Frequency_Drug	103
Total	6890

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{20}$$

where TP stands for true positive, which means the sample is predicted as a positive sample and is actually a positive sample. FP means false positive, which means the sample is predicted as a positive, but it is actually a negative sample. FN means false negative, which means the sample is predicted as a negative sample, but it is actually a positive sample.

We compare our CRE-RFGP with the following eight state-of-the-art models proposed in recent years, as well as two models for Chinese.

- **ETL-Span**[47] applies a novel decomposition method. Distinguishing all head entities, it utilizes a fragment labelling approach to identify the corresponding tail entities and relations.
- **RSAN**[39] adopts a relation-based attention network. To build specific sentence representations for individual relations, relation-aware attention mechanisms are employed. Sequence labelling is subsequently conducted to extract the head and tail entities accordingly.
- **CasRel**[22] is a novel cascaded binary tagging framework. Relations in a sentence are modelled as functions that map subjects to objects.
- **TPLinker**[16] is a single-stage joint entity extraction model. Joint extraction is formulated as a problem of linking token pairs. A new handshake tagging scheme is also introduced, which aligns boundary tags for pairs of entities to mitigate exposure bias.
- **PRGC**[12] has a component to pre-indicate potential relations and then applies relation-specific sequence labelling combinations to handle overlapping probabilities between subjects and objects.
- **OneRel**[32] is a new joint relation extraction model. A fine-grained triple classification problem is introduced as a transformation for joint extraction. A score-based classifier and a horns tagging strategy are the components of the model.
- **ETEMA**[48] uses BERT to mine the semantic information of event sentences, and uses attention and semantic information to interactively combine event information with its contextual information.
- **BERT-PAGG**[49] integrates entity location information into local features through a piecewise convolutional neural network, uses the attention mechanism to capture more effective semantic features, and finally regulates the transmission of information flow through the gating mechanism.

4.1.3 Parameters

In the experiment, we utilize the BERT-base-Chinese model[40] obtained from the Huggingface version of the Transformer as the pre-trained model. The learning rate for model training is set to 5×10^{-5}, the batch size is 8, and the number of epochs is 50. We employ the AdamW algorithm to optimize gradient descent[50], which ensures stability throughout the iteration process, while also accelerating convergence and preventing overfitting. The weight decay value is set to 0.01. The output feature dimension of the BERT-base-Chinese is 768. In the dataset, the number of learnable embeddings denoted as n_q, is set to 15/12. The relation filtering threshold λ_1 is set to 0.4, and the entity matching threshold λ_2 is set to 0.6. Considering that some of the model weights are randomly initialized at the beginning of training, selecting a large learning rate may lead to instability. Hence, we incorporate an optimization method for the learning rate called warmup[51] during training. In the initial few epochs, a small learning rate is chosen to gradually adjust the weight distribution and accumulate some prior knowledge. As training progresses, the learning rate linearly increases, facilitating accelerated learning. This approach mitigates the issue of excessively large learning rates disrupting distribution stability, thereby enhancing the training effectiveness of the model.

4.2 Overall performance

The experimental results are presented in Table 3. It is easy to find that our CRE-RFGP achieves the best performance over all those competing models in all cases. Compared with CRE-RFGP, the ETL-Span model employing fragment tagging strategy also exhibits significant performance limitations. This is because the fragment-based tagging approach also struggles with complex entities and overlapping relations. Overall, CRE-RFGP outperforms ETL-Span by 20.9%, 27.1%, and 24.0%, in precision, recall, and F1-score, respectively, on the CMeIE dataset. Besides, it outperforms ETL-Span by 15.7%, 19.1%, and 17.4%, in precision, recall, and F1-score, respectively, on the DiaKG dataset.

As introduced before, the CasRel model adopts a binary tagging strategy, and the TPLinker model uses a pointer tagging strategy. In contrast, our CRE-RFGP employs a global pointer tagging strategy. Experimental results show that CRE-RFGP achieved significant improvements compared with CasRel and TPLinker. On the CMeIE dataset, CRE-RFGP outperforms CasRel in precision, recall, and F1-score by 9.7%, 9.9%, and 9.8%, respectively. On the DiaKG dataset, CRE-RFGP outperforms CasRel by 9.6%, 9.3%, and 9.4% in precision, recall, and F1-score, respectively. On the CMeIE dataset, CRE-RFGP outperforms TPLinker by 8.7%, 6.7%, and 7.7% in precision, recall, and F1-score, respectively. On the DiaKG dataset, CRE-RFGP outperforms TPLinker by 9.6%, 7.2%, and 8.3% in precision, recall, and F1-score, respectively. This outcome highlights the effectiveness of incorporating a relation-specific attention mechanism to obtain relation-specific sentence representations. These representations integrate crucial information and can help better interpret key information within complex sentence structures.

Both the RSAN model and PRGC model adopt a relation-specific sequence tagging method. On the CMeIE dataset, CRE-RFGP outperforms RSAN by 11.6%, 11.3%, and 11.4% in precision, recall, and F1-score, respectively. On the DiaKG dataset, CRE-RFGP outperforms RSAN by 11.9%, 12.6%, and 12.5% in precision, recall, and F1-score, respectively. CRE-RFGP outperforms PRGC by 7.0%, 7.5%, and 7.2% in precision, recall, and F1-score, respectively, on the CMeIE dataset. Moreover, on the DiaKG dataset, CRE-RFGP outperforms PRGC by 8.2%, 5.2%, and 6.6% in precision, recall, and F1-score, respectively. This indicates that the global pointer network used by our CRE-RFGP is more useful for handling the length and span of entities compared with sequence labelling methods, making it more beneficial for extracting overlapping and nested triples.

CRE-RFGP performs better than the OneRel model on both datasets. Specifically, on the CMeIE dataset, CRE-RFGP outperforms OneRel by 6.5%, 5.2%, and 5.9% in precision, recall, and F1-score, respectively. On the DiaKG dataset, CRE-RFGP surpasses OneRel by 6.2%, 5.2%, and 5.6% in precision, recall, and F1-score, respectively. There are two reasons behind this. First, compared with OneRel, the relation detection module in CRE-RFGP significantly reduces irrelevant relations. OneRel generates tabular features for each

Table 3 Comparison of the proposed CRE-RFGP against the state-of-the-art methods on the CMeIE and DiaKG datasets. Bold values indicate the best results in each column. The upward arrows (↑) and values in parentheses indicate the relative improvement of CRE-RFGP over the corresponding baseline model, which is calculated as (CRE-RFGP − baseline) / baseline × 100%.

(Unit: %)

Model	CMeIE			DiaKG		
	Precision	Recall	F1-score	Precision	Recall	F1-score
ETL-Span	60.7 (↑20.9%)	58.7 (↑27.1%)	59.7 (↑24.0%)	59.2 (↑15.7%)	56.1 (↑19.1%)	57.6 (↑17.4%)
RSAN	65.8 (↑11.6%)	67.0 (↑11.3%)	66.4 (↑11.4%)	61.2 (↑11.9%)	59.3 (↑12.6%)	60.2 (↑12.5%)
CasRel	66.9 (↑9.7%)	67.9 (↑9.9%)	67.4 (↑9.8%)	62.5 (↑9.6%)	61.1 (↑9.3%)	61.8 (↑9.4%)
TPLinker	67.5 (↑8.7%)	69.9 (↑6.7%)	68.7 (↑7.7%)	62.5 (↑9.6%)	62.3 (↑7.2%)	62.4 (↑8.3%)
PRGC	68.6 (↑7.0%)	69.4 (↑7.5%)	69.0 (↑7.2%)	63.3 (↑8.2%)	63.5 (↑5.2%)	63.4 (↑6.6%)
OneRel	68.9 (↑6.5%)	70.9 (↑5.2%)	69.9 (↑5.9%)	64.5 (↑6.2%)	63.5 (↑5.2%)	64.0 (↑5.6%)
BERT-PAGG	63.2 (↑16.1%)	62.0 (↑20.9%)	62.6 (↑18.2%)	59.5 (↑15.1%)	58.5 (↑14.2%)	59.0 (↑14.6%)
ETEMA	65.8 (↑11.6%)	67.8 (↑10.0%)	66.8 (↑10.8%)	62.0 (↑10.5%)	61.0 (↑9.5%)	61.5 (↑9.9%)
CRE-RFGP (ours)	**73.4**	**74.6**	**74.0**	**68.5**	**66.8**	**67.6**

relation, which means that filtering out negative correlations provides additional effectiveness compared with models that perform entity extraction under each relation. Second, CRE-RFGP incorporates a relation-specific attention mechanism, which is crucial for relation triple extraction. On the other hand, OneRel only assigns trainable weights to these relations, failing to fully explore the semantic information of the relations. In conclusion, due to the limited occurrence of nested entities and the relatively small number of predefined relations, CRE-RFGP performs relatively poorly on the DiaKG dataset. Hence, the full advantage of CRE-RFGP may not be fully reflected in this dataset. Conversely, on the CMeIE dataset with 53 relation types, CRE-RFGP achieves an F1-score advantage of 8.5% over OneRel. This observation emphasizes the superior capability of CRE-RFGP with multi-relation datasets.

On both datasets, CRE-RFGP outperformed the BERT-PAGG model. Specifically, on the CMeIE dataset, compared with BERT-PAGG, CRE-RGP is better than 16.1%, 20.9%, and 18.2% in terms of accuracy, recall, and F1-score, respectively. On the DiaKG dataset, compared with BERT-PAGG, CRE-RGP was better than 15.1%, 14.2%, and 14.6% in terms of accuracy, recall, and F1-score, respectively. On both datasets, CRE-RFGP outperformed the ETEMA model. Specifically, on the CMeIE dataset, CRE-RFGP is better than 11.6%, 10.0%, and 10.8% in accuracy, recall, and F1-score compared with ETEMA. On the DiaKG dataset, CRE-RFGP is better than ETEMA in terms of accuracy, recall, and F1-score by 10.5%, 9.5%, and 9.9%, respectively. Both ETEMA and BERT-PAGG models only use the attention mechanism to capture more effective semantic information but neither consider the problem of relation redundancy nor solve the problem of relation overlapping and nesting, resulting in experimental results that are not as good as CRE-RFGP.

4.3 Performance in complex scenarios

To comprehensively evaluate CRE-RFGP's ability to address entity overlapping/nesting issues and extract multiple relations simultaneously from more complex sentences, we split the two datasets into specific subsets according to the characteristics of entities in them. The F1-scores of all competing models on the two datasets are presented in Tables 4 and 5. It can be observed from Tables 4 and 5 that the relation-specific attention mechanism in CRE-RFGP contributes to solving the problem of entity overlapping and the global pointer extraction module contributes to the problem of entity nesting.

The comparison results illustrate that CRE-RFGP excels in handling normal, SEO, and EPO challenges, indicating that CRE-RFGP is capable of effectively handling complex relations within complex sentences. Furthermore, when it comes to extracting multiple relations simultaneously, CRE-RFGP consistently maintains a high F1-score, indicating its capability to extract multiple triplets concurrently. This highlights the model's proficiency in handling cases where multiple relations exist within the same sentence or context.

Table 4 F1-scores of different models with different overlapping patterns (normal, entity pair overlap (EPO), and single entity overlap (SEO)) on the CMeIE and DiaKG datasets. Bold values indicate the best results in each column. The upward arrows (↑) and values in parentheses indicate the relative improvement of CRE-RFGP over the corresponding baseline model, which is calculated as (CRE-RFGP − baseline) / baseline × 100%.

(Unit: %)

Model	CMeIE			DiaKG		
	Normal	SEO	EPO	Normal	SEO	EPO
ETL-Span	64.3 (↑13.4%)	63.4 (↑22.1%)	56.1 (↑34.4%)	57.3 (↑13.8%)	59.1 (↑17.1%)	53.4 (↑24.7%)
RSAN	65.0 (↑12.2%)	65.2 (↑18.7%)	66.9 (↑12.7%)	57.5 (↑13.4%)	57.8 (↑19.7%)	59.6 (↑11.7%)
CasRel	67.1 (↑8.6%)	67.8 (↑14.2%)	68.8 (↑9.6%)	58.0 (↑12.4%)	58.5 (↑18.3%)	59.8 (↑11.4%)
TPLinker	67.9 (↑7.4%)	69.3 (↑11.7%)	70.1 (↑7.6%)	60.2 (↑8.3%)	64.8 (↑6.8%)	61.5 (↑8.3%)
PRGC	69.3 (↑5.2%)	68.8 (↑12.5%)	70.2 (↑7.4%)	62.3 (↑4.7%)	65.2 (↑6.1%)	61.5 (↑8.3%)
OneRel	69.4 (↑5.0%)	71.1 (↑8.9%)	70.0 (↑7.7%)	62.9 (↑3.7%)	65.6 (↑5.5%)	62.2 (↑7.1%)
BERT-PAGG	65.8 (↑10.8%)	60.3 (↑28.4%)	59.8 (↑26.1%)	61.2 (↑6.5%)	58.6 (↑18.1%)	55.2 (↑20.1%)
ETEMA	68.4 (↑6.6%)	66.1 (↑17.1%)	64.0 (↑17.8%)	62.0 (↑5.2%)	59.4 (↑16.5%)	58.6 (↑13.7%)
CRE-RFGP (ours)	**72.9**	**77.4**	**75.5**	**65.2**	**69.2**	**66.6**

Table 5 F1-scores of different models with different numbers of triples on the CMeIE and DiaKG datasets. *n* is the number of triples in a sentence. Bold values indicate the best F1-score in each column.

(Unit: %)

Model	CMeIE					DiaKG				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n \geqslant 5$	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n \geqslant 5$
ETL-Span	67.3	63.9	60.4	62.5	62.6	58.5	56.2	52.9	54.3	54.7
RSAN	68.5	67.0	67.4	66.1	65.1	60.3	60.0	59.1	58.5	57.5
CasRel	65.8	67.1	68.6	68.9	60.8	62.0	61.6	60.2	60.8	58.1
TPLinker	70.9	68.2	68.9	67.9	68.4	63.3	60.9	61.5	62.2	61.6
PRGC	70.0	69.8	69.3	69.8	68.3	62.5	63.5	64.0	62.2	62.3
CRE-RFGP (ours)	**72.9**	**73.3**	**74.1**	**75.6**	**76.3**	**65.3**	**65.6**	**67.5**	**67.8**	**68.0**

As observed from the experimental results in Table 4, CRE-RFGP outperforms other models in handling entity overlapping issues. On the CMeIE dataset, CRE-RFGP achieves F1-scores of 72.9%, 77.4%, and 75.5% in the normal, SEO, and EPO cases, respectively. On the DiaKG dataset, CRE-RFGP achieves F1-scores of 65.2%, 69.2%, and 66.6% in the normal, SEO, and EPO cases, respectively. Compared with CasRel and ETL-Span, which solely adopt sequence tagging strategies, CRE-RFGP outperforms ETL-Span by 13.4%, 22.1%, and 34.4% in the normal, SEO, and EPO cases, respectively, on the CMeIE dataset. On the DiaKG dataset, CRE-RFGP outperforms ETL-Span by 13.8%, 17.1%, and 24.7% in the normal, SEO, and EPO cases, respectively. On the CMeIE dataset, CRE-RFGP outperforms CasRel by 8.6%, 14.2%, and 9.6% in the normal, SEO, and EPO cases, respectively. On the DiaKG dataset, CRE-RFGP outperforms CasRel by 12.4%, 18.3%, and 11.4% in the normal, SEO, and EPO cases, respectively. This is because of the complex hierarchical structure of overlapping and nested entities, where relying solely on ordinary sequence tagging strategies is insufficient to handle such complexity. However, CRE-RFGP incorporates a relation attention mechanism before entity extraction to integrate relation information and utilizes global pointers to extract entities, both of which contribute to handling complex scenarios effectively.

Compared with RSAN and PRGC, which combine sequence tagging strategies with relation information, CRE-RFGP outperforms RSAN by 12.2%, 18.7%, and 12.7% in the normal, SEO, and EPO cases, respectively, on the CMeIE dataset. On the DiaKG dataset, CRE-RFGP outperforms RSAN by 13.4%, 19.7%, and 11.7% in the normal, SEO, and EPO cases, respectively. On the CMeIE dataset, CRE-RFGP outperforms PRGC by 5.2%, 12.5%, and 7.4% in the normal, SEO, and EPO cases, respectively. On the DiaKG dataset, CRE-RFGP outperforms PRGC by 4.7%, 6.1%, and 8.3% in the Normal, SEO, and EPO cases, respectively. This is because incorporating relation information improves the performance of extracting overlapping entities, but it still cannot handle nested entities. The TPLinker and OneRel models, which utilize table filling, also demonstrate bad performance in extracting overlapping and nested entities. On the CMeIE dataset, the CRE-RFGP outperforms TPLinker by 7.4%, 11.7%, and 7.6% in the normal, SEO, and EPO cases. On the DiaKG dataset, the CRE-RFGP outperforms TPLinker by 8.3%, 6.8%, and 8.3% in the normal, SEO, and EPO cases. On the CMeIE dataset, the CRE-RFGP model outperforms OneRel by 5.0%, 8.9%, and 7.7% in the normal, SEO, and EPO cases. On the DiaKG dataset, the CRE-RFGP outperforms OneRel by 3.7%, 5.5%, and 7.1% in the normal, SEO, and EPO cases. Because the table-filling method treats the start and end of an entity individually, OneRel is not very sensitive to the length and span of the entity. As a result, it is hard for OneRel to perform as well as CRE-RFGP when extracting nested entities.

On the CMeIE dataset, the CRE-RFGP model outperformed ETEMA by 10.8%, 28.4%, and 26.1% in normal, SEO and EPO cases, respectively. On the DiaKG dataset, CRE-RFGP outperformed ETEMA by 6.5%, 18.1%, and 20.1% in normal, SEO, and EPO cases, respectively. On the CMeIE dataset, the CRE-RFGP model outperformed BERT-PAGG by 6.6%, 17.1% and 17.8% in normal, SEO, and EPO cases, respectively. On the DiaKG dataset, CRE-RFGP outperformed BERT-PAGG by 5.2%, 16.5%, and 13.7% in normal, SEO, and EPO cases, respectively. Because ETEMA only uses the attention mechanism to focus on the importance between words to simulate

entity information, there is no method such as sequence annotation or pointer annotation. As a result, ETEMA struggles to perform as well as CRE-RFGP when it comes to extracting overlapping and nested entities. The BERT-PAGG model uses the location information and attention mechanism of entities, which can better identify overlapping entities, but it is difficult to extract nested entities.

These experimental results demonstrate that the adopted global entity extraction approach effectively addresses the issues of entity overlap and unclear boundaries. Furthermore, the model significantly reduces redundancy in entity output. Specifically, CRE-RFGP performs best in resolving SEO issues, followed by EPO and normal.

4.4 Performance with different tuple numbers

The experimental results in Table 5 represent datasets divided into subsets containing n tuples in a sentence, aiming to investigate the performance of CRE-RFGP with different numbers of tuples in a sentence. The results indicate that CRE-RFGP outperforms other models when handling complex sentences, especially when the sentences contain more tuples. CRE-RFGP demonstrates strong performance, achieving the highest values when there are five or more tuples, with precision, recall, and F1-scores of 77.1%, 75.5%, and 76.3%, respectively. In the CMeIE dataset, CRE-RFGP consistently achieves an F1-score above 73% in scenarios with any number of tuples, surpassing other models' performance. This again confirms the effectiveness of CRE-RFGP. In the DiaKG dataset, CRE-RFGP achieves precision, recall, and F1-score above 65% in scenarios with any number of tuples, all of which are higher than the F1-scores of the comparative models. The highest values are obtained when there are five or more tuples, with precision, recall, and F1-scores of 69.3%, 66.8%, and 68%, respectively.

Furthermore, it is worth noting that the performance of CRE-RFGP in multi-tuple scenarios improves with an increasing number of tuples in a sentence. This demonstrates that CRE-RFGP, utilizing global pointers for entity extraction, is effective in handling the extraction of multiple tuples.

4.5 Impact of parameters

Our model introduces two parameters: λ_1 and λ_2. λ_1 represents the threshold employed in the relation decoding module, aimed at filtering a subset of relations using a non-autoregressive decoder. This addresses the problem of entity redundancy, which arises from both entity extraction and redundant operations on all relations. By controlling the probability threshold for relation prediction, λ_1 ensures that only high-confidence relations are retained, effectively reducing noise and improving precision. λ_2 serves as a global threshold value utilized during the construction of the entity mapping matrix. Its purpose is to pre-filter entity correspondences, ensuring that only the most reliable entity pairs are considered for relation extraction. This additional filtering step validates the extracted entity triplets generated by the global pointer network, thereby preventing incorrect entity pairings and improving the model's robustness against noisy data. Consequently, this approach enables a greater number of correct triplets to be obtained while eliminating spurious relations.

Now, let us examine the impact of different values of these two parameters on the final results. Figure 4a shows the variation trend of F1-score vs λ_2 when λ_1 is equal to 0.4. It can be observed that when λ_1 remains constant, as λ_2 gradually increases, the F1-score of CRE-RFGP initially rises from 69.3% to 73.9%, and then continues to increase until reaching the highest value of 74.0%. Afterwards, when λ_2 exceeds 0.6, the F1-score starts to decline, indicating that an excessively high λ_2 filters out too many valid entity correspondences, leading to a loss in recall. Figure 4b illustrates the variation trend of F1-score vs λ_1 when λ_2 is equal to 0.4. It can be seen that when λ_2 remains constant, as λ_1 gradually increases, the F1-score of CRE-RFGP initially rises from 21% to 74.0%, then stabilizes until it starts to decline after λ_1 reaches 0.6. This trend suggests that moderate filtering of relations enhances model performance, but overly aggressive filtering leads to missed correct relations, negatively affecting recall.

When the relation threshold λ_1 is set to 0, regardless of the value of the entity correspondence threshold λ_2, the obtained F1-scores remain around 20%. This indicates that when λ_1 is set to 0, no filtering of improbable relations occurs, resulting in a large number of redundant relations. These redundant relations ultimately overwhelm the entity extraction process, leading to poor model performance. This result demonstrates the effectiveness of the relation

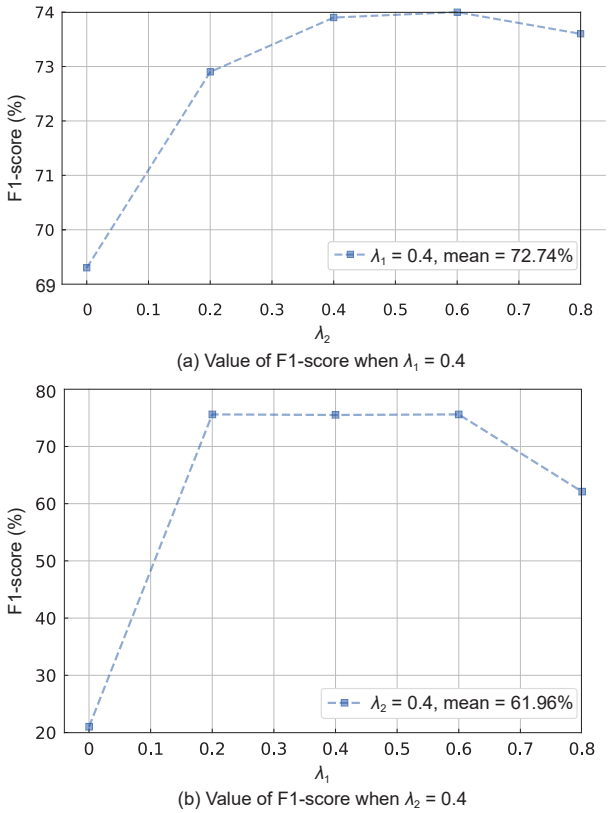


Fig. 4 Parameter sensitivity analysis of CRE-RFGP with respect to the relation filtering threshold λ_1 and the entity correspondence threshold λ_2. Subfigure (a) reports the F1-score under different λ_2 values when λ_1 is fixed at 0.4, and subfigure (b) reports the F1-score under different λ_1 values when λ_2 is fixed at 0.4.

decoding module in removing noise and improving relation precision. As λ_1 increases, the F1-score stabilizes around 74.0% and decreases by only 1% compared with the best result, suggesting that a well-tuned λ_1 helps balance relation filtering and model recall.

From the experimental results shown in Fig. 4, it can be observed that as the entity correspondence threshold λ_2 increases, the F1-score decreases by 16% compared with the best result. This indicates that when λ_2 is set too high, only a few correct entity correspondences remain in the matrix, while many valid entity pairs may fall below the filtering threshold (e.g., a threshold of 0.8 eliminates many correct entity pairs). This means that many correct triplets are mistakenly excluded, leading to a significant drop in recall and demonstrating that an excessively high λ_2 negatively impacts model performance. On the other hand, when λ_2 is set to 0, the F1-score of CRE-RFGP remains around 69.5%, which is approximately 7% lower than the best experimental result. This discrepancy demonstrates that without proper entity correspondence filtering, erroneous entity pairings are introduced, reducing precision. The effectiveness of the entity mapping matrix module is evident from these results, as it provides an essential filtering step to refine triplet extraction.

Based on Fig. 4, the best performance of the model is achieved when $\lambda_1 = 0.4$ and $\lambda_2 = 0.6$, yielding an F1-score of 76.2%. Compared with other parameter settings, F1-score variations remain within 1.5%, suggesting that CRE-RFGP maintains stable performance across different configurations. Consequently, based on experimental observations, we adopt $\lambda_1 = 0.4$ and $\lambda_2 = 0.6$ as the optimal parameter settings in CRE-RFGP, striking a balance between relation filtering and entity correspondence selection to achieve superior relation extraction performance.

4.6 Ablation study

To further showcase the effectiveness of each unique design in CRE-RFGP, ablation experiments are performed. Each time one component is removed while keeping the other components unchanged. The results on both datasets are summarized in Tables 6 and 7, respectively.

First, we remove the relation-first decoder component from our model and perform the entity extraction using each relation from both datasets. The

Table 6 Ablation study of CRE-RFGP on the CMeIE dataset. Bold values indicate the best results in each column.

(Unit: %)

Model	Precision	Recall	F1-score
CRE-RFGP	**73.4**	**74.6**	**74.0**
Without relation-first decoder	70.3	72.5	71.4
Without relation-specific attention	56.6	72.3	63.5
Without global pointer	62.4	71.6	66.7
Without global correspondence matrix	67.0	72.3	69.5

Table 7 Ablation study of CRE-RFGP on the DiaKG dataset. Bold values indicate the best results in each column.

(Unit: %)

Model	Precision	Recall	F1-score
CRE-RFGP (ours)	**68.5**	**66.8**	**67.6**
Without relation-first decoder	65.2	66.0	65.5
Without relation-specific attention	54.6	66.3	59.8
Without global pointer	60.5	64.5	62.4
Without global correspondence matrix	63.9	66.2	65.0

results show a decrease of 4.2% in precision, 2.8% in recall, and 3.5% in F1-score on the CMeIE dataset. On the DiaKG dataset, CRE-RFGP's precision, recall, and F1-score decreased by 4.8%, 1%, and 3%, respectively. This is because using individual relations for predicting entities in a large relation set dataset introduces significant entity and relation redundancy, resulting in a decrease in model performance. This experiment demonstrates the effectiveness of the relation-first decoder module in the overall model, improving precision.

Second, by removing the relation-specific attention mechanism module, sentence representations are no longer constructed for specific relations. Instead, relation embeddings are directly employed to guide the process of entity extraction. The results show a decrease of 22.8% in precision, 3.1% in recall, and 14.2% in F1-score on the CMeIE dataset. On the DiaKG dataset, CRE-RFGP's precision, recall, and F1-score decreased by 20.3%, 0.7%, and 11.5%, respectively. This is because relation embeddings only learn shallow co-occurrence of triples, leading to more triple predictions but lower precision. In contrast, our relation-specific attention mechanism can capture fine-grained semantic features in sentences, providing specific relation-based global pointer entity extraction for subsequent modules.

When we remove the global pointer module and adopt a dual-pointer-based approach to detect the start and end positions of entities using pointer labels. The results showed a decrease of 15.0% in precision, 4% in recall, and 9.9% in F1-score on the CMeIE dataset. On the DiaKG dataset, CRE-RFGP's precision, recall, and F1-score decreased by 11.7%, 3.4%, and 7.7%, respectively. This is because the pointer-based approach only focuses on the start and end positions of entities and is not sensitive to entity length and span. However, the global pointer module considers the length of each entity, treating the start and end as a single position label. This approach can improve precision and effectively address nested entities.

When we remove the global correspondence matrix module and instead use a heuristic nearest neighbour principle to combine subjects and objects. The results show a decrease of 8.7% in precision, 3% in recall, and 6.1% in F1-score on the CMeIE dataset. On the DiaKG dataset, CRE-RFGP's precision, recall, and F1-score decreased by 6.7%, 0.9%, and 3.8%, respectively. This is because, without the entity matrix module, the model predicts a higher number of triple predictions due to many mismatched entity pairs, which affects the model's performance by losing the constraints imposed by the matrix.

4.7 Model efficiency

From Table 8, it can be observed that the relation-first decoder and global correspondence matrix in CRE-RFGP contribute to the overall model efficiency. On the CMeIE dataset, without the relation-first decoder, CRE-RFGP has 788 656 candidate relations. However, with the relation-first decoder, the number of candidate relations is reduced to 56 265. In the CMeIE dataset, the search space is reduced by 92.9%, and in the DiaKG dataset, it is reduced by 83.8%. This reduction is attributed to the fact that without the relation-first decoder, each sentence needs to undergo entity extraction for all relations. With the relation-first decoder, each sentence filters out a large portion of irrelevant relations, significantly reducing the search space of CRE-RFGP.

Theoretically, without the global correspondence matrix, CRE-RFGP adopts the most common heuristic nearest neighbour algorithm for entity alignment. This algorithm aligns one pair of entities at a time, requiring a loop to search for each subject and subsequently search for its corresponding object. If a sentence contains k entity pairs, this algorithm needs to perform k iterations, resulting in a time complexity of $O(kn^2)$. On the other hand, if the global correspondence matrix is utilized, although the time complexity of searching the matrix is also $O(n^2)$, the key difference lies in the fact that the global correspondence matrix can handle all entity pairs in a sentence at once. Therefore, its overall time complexity becomes $O(n^2)$, which leads to a k-fold improvement in time complexity.

4.8 Time consumption discussion

As shown in Table 9, CRE-RFGP requires less training

Table 8 Effect of the relation-first decoder on the number of candidate relations.

Dataset	CRE-RFGP	Number of candidate relations
CMeIE	Without relation-first decoder	788 656
	With relation-first decoder	56 265
DiaKG	Without relation-first decoder	28 380
	With relation-first decoder	9460

and inference time than PRGC and OneRel. On the CMeIE dataset, CRE-RFGP reduces the training time from 428 s for PRGC and 399 s for OneRel to 285 s, corresponding to speedups of approximately 1.50× and 1.40×, respectively. On the DiaKG dataset, CRE-RFGP reduces the training time from 286 s for PRGC and 239 s for OneRel to 191 s, corresponding to speedups of approximately 1.50× and 1.25×, respectively. This improvement is primarily due to two key factors:

- **Computational complexity reduction.** By utilizing a global entity alignment matrix, the final relation extraction step requires only matrix-based verification, eliminating the need for additional iterative computations.
- **Efficient multitask learning.** During training, CRE-RFGP employs a multitask learning strategy, where the encoder parameters are shared across tasks. This reduces the total number of model parameters, optimizing memory usage and speeding up gradient updates.

For inference speed, CRE-RFGP also outperforms competing models. On the CMeIE dataset, it achieves an inference speed 1.2 times faster than PRGC and 1.1 times faster than OneRel, while on the DiaKG dataset, it is 1.2 times faster than both PRGC and OneRel. The reduction in parameter count directly contributes to this improvement. Since CRE-RFGP requires fewer computational and storage resources, it minimizes memory footprint and allows for faster execution during inference. These results highlight the model's efficiency, making it more suitable for real-world applications where both high accuracy and fast response times are essential.

4.9 Case study

To visually validate the effectiveness of our model, we present two examples from the test set and extraction results generated by PRGC and CRE-RFGP, as shown in Fig. 5.

In the first example, there are four correct triples in this sentence. CRE-RFGP successfully identifies four correct relation triples, but PRGC only successfully recognizes three relation triplets. Although the entity pair in triple {oral, ADE_Drug, insulin secretagogues} is correct, the relation between entities is incorrectly identified by PRGC. In the second example, CRE-

Table 9 Efficiency comparison of different models on the CMeIE and DiaKG datasets. Bold values indicate the best results in each column. For training time, inference time, and number of parameters, smaller values are better; for F1-score, larger values are better.

Dataset	Model	Training time (s)	Inference time (s)	Number of parameters	F1-score (%)
CMeIE	PRGC	428	12	86 056	69.0
	OneRel	399	11	75 120	69.9
	CRE-RFGP (our)	**285**	**10**	**63 421**	**74.0**
DiaKG	PRGC	286	12	110 503	63.4
	OneRel	239	12	92 225	64.0
	CRE-RFGP (our)	**191**	**10**	**76 932**	**67.6**

Text	Ground Truth	PRGC	CRE-RFGP
CMeIE: 1705 In **oral** hypoglycemic medications, **insulin secretagogues** can all cause **hypoglycemia**, among which **sulfonylureas** are particularly common. **Glitinides** are relatively less common due to their short pharmacological duration of action.	{oral, method_drug, insulin secretagogues}, {oral, method_drug, sulfonylurea}, {hypoglycemia, ADE_Drug, sulfonylurea}, {hypoglycemia, ADE_Drug, Glitinide}	{oral, method_drug, insulin secretagogues}, {oral, method_drug, sulfonylurea}, {hypoglycemia, ADE_Drug, sulfonylurea}, {oral, ADE_Drug, Insulin secretagogues} ✖	{oral, method_drug, insulin secretagogues} {oral, method_drug, sulfonylurea}, {hypoglycemia, ADE_Drug, sulfonylurea}, {hypoglycemia, ADE_Drug, glitinide}
DiaKG: 106 Many patients with **acute myeloid leukemia** experience **respiratory distress**. **Acute myeloid leukemia** is caused by **bone marrow infiltration**, **anemia**, or systemic inflammatory cytokine-related effects.	{acute myeloid leukemia, cause, bone marrow infiltration}, {acute myeloid leukemia, clinical manifestation, respiratory distress}, {acute myeloid leukemia, cause, anemia }	{Acute myeloid leukemia, cause, bone marrow infiltration}, {acute myeloid leukemia, clinical manifestation, respiratory distress}, {acute myeloid leukemia, cause, anemia}, {acute myeloid leukemia, cause, respiratory distress} ✖	{Acute myeloid leukemia, cause, bone marrow infiltration}, {acute myeloid leukemia, clinical manifestation, respiratory distress}, {acute myeloid leukemia, cause, anemia}

Fig. 5 Comparison of PRGC and CRE-RFGP. Examples are collected from the CMeIE and DiaKG datasets. In the original sentences, correct entities are in bold and underlined, and the correct relations are colored. The red crosses indicate incorrect relations.

RFGP successfully identifies three correct relation triples while PRGC makes an error in predicting the relations between entities in the triple {acute myeloid leukemia, case, trouble breathing}.

Through the two cases studied in Fig. 5, we observe that PRGC can successfully extract entities but sometimes cannot identify their relations correctly. This is because the model tends to memorize the positions of entities rather than understanding their basic semantics. However, in the CRE-RFGP model, the method of extracting entities using a relation-specific attention mechanism combined with a global pointer network not only considers the positions of subject–object entities but also leverages the attention mechanism to capture relational information between entities. This enables CRE-RFGP to perform better in the entity extraction task.

5 Conclusion

This paper proposed an end-to-end NLP model named CRE-RFGP for extracting relations from Chinese medical records. Specifically, CRE-RFGP leverages pre-trained BERT-base-Chinese to obtain character embeddings. It utilizes a relation-first decoder to prioritize the prediction of potential relations and extract a subset of relations. Then, a relation-specific attention mechanism is employed to capture deep semantic features of relations in medical records. To handle entity overlapping and nesting between entities, a global pointer network is utilized. Finally, an entity mapping matrix is designed to align subjects and objects into triplets, ensuring low complexity. Experimental results demonstrated that CRE-RFGP achieves state-of-the-art performance on two Chinese medical datasets specifically designed for relation extraction evaluation. Besides, CRE-RFGP can also effectively handle numerous complex scenarios where entities are overlapping and/or one entity is nested in another one.

Despite its strong performance, CRE-RFGP still faces challenges in certain scenarios, leading to extraction errors. We analyzed the incorrect predictions and identified three primary sources of errors as below.

- **Complex sentence structures with implicit relations.** In some cases, Chinese medical records contain long and convoluted sentences with implicit relations that are difficult to infer directly. Example: A sentence describing a treatment outcome without explicitly stating the causal link between the drug and the disease may lead to relation misclassification. While CRE-RFGP captures explicit relationships well, implicit relations relying on external knowledge remain challenging.
- **Ambiguity in domain-specific terminology.** Some medical terms have multiple meanings depending on the context, leading to incorrect entity classification. Example: The term “hepatitis” might be linked to different causes (viral, autoimmune, and drug-induced) without explicit disambiguation in the text. Future improvements could involve integrating domain-specific knowledge graphs to improve term disambiguation.
- **Errors in nested entity extraction.** Although CRE-RFGP performs well on overlapping entities, it occasionally struggles with deeply nested entities where a single entity belongs to multiple hierarchical structures. Example: “Diabetic nephropathy” includes both “diabetes” and “nephropathy”, yet in some cases, the model incorrectly associates it with only one disease. This indicates a need for improving contextual dependency modeling to enhance entity segmentation.

In the future, we will design a lower-complexity labeling method based on the global pointer labeling strategy. In addition, considering that reinforcement learning has not been widely applied in the field of relation extraction, we will further explore incorporating reinforcement learning to solve the overlapping relation extraction tasks. Furthermore, we plan to integrate external medical knowledge bases to improve model robustness against ambiguous terminology. Additionally, addressing nested entity issues by enhancing contextual representations and leveraging hierarchical learning strategies will be a key direction for improvement.

Acknowledgment

This work was supported by the Leading-edge Technology Program of Jiangsu Natural Science Foundation (No. BK20202001), the Key R&D Plan of Jiangsu Province (Social Development) (Nos. BE2020721 and BE2020798), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX23_1089).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] D. Sousa and F. M. Couto, Biomedical relation extraction with knowledge graph-based recommendations, *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 4207–4217, 2022.

[2] P. L. H. Cabot and R. Navigli, REBEL: Relation extraction by end-to-end language generation, in *Proc. Findings of the Association for Computational Linguistics*, Punta Cana, Dominican Republic, 2021, pp. 2370–2381.

[3] J. Giorgi, G. Bader, and B. Wang, A sequence-to-sequence approach for document-level relation extraction, in *Proc. 21st Workshop on Biomedical Language Processing*, Dublin, Ireland, 2022, pp. 270–284.

[4] T. Zhang, N. Li, Y. Zhou, W. Cai, and L. Ma, Information extraction of Chinese medical electronic records via evolutionary neural architecture search, in *Proc. 2023 IEEE Int. Conf. Data Mining Workshops*, Shanghai, China, 2023, pp. 396–405.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[6] Z. Li, H. Chen, R. Qi, H. Lin, and H. Chen, DocR-BERT: Document-level R-BERT for chemical-induced disease relation extraction via Gaussian probability distribution, *IEEE J. Biomed. Health Inf.*, vol. 26, no. 3, pp. 1341–1352, 2022.

[7] H. Li, M. Cheng, Z. Yang, L. Yang, and Y. Chua, Named entity recognition for Chinese based on global pointer and adversarial training, *Sci. Rep.*, vol. 13, no. 1, p. 3242, 2023.

[8] X. Sun, Q. Guo, and S. Ge, GFN: A novel joint entity and relation extraction model with redundancy and denoising strategies, *Knowl.-Based Syst.*, vol. 300, p. 112137, 2024.

[9] Q. Liu, L. Zhang, G. Ren, and B. Zou, Research on named entity recognition of Traditional Chinese Medicine chest discomfort cases incorporating domain vocabulary features, *Comput. Biol. Med.*, vol. 166, p. 107466, 2023.

[10] F. Lu, Q. Qi, and H. Qin, Joint extraction of Uygur medicine knowledge with edge computing, *Tsinghua Science and Technology*, vol. 30, no. 2, pp. 782–795, 2025.

[11] Z. Li, L. Fu, X. Wang, H. Zhang, and C. Zhou, RFBFN: A relation-first blank filling network for joint relational triple extraction, in *Proc. 60th Annu. Meeting of the Association for Computational Linguistics: Student Research Workshop*, Dublin, Ireland, 2022, pp. 10–20.

[12] H. Zheng, R. Wen, X. Chen, Y. Yang, Y. Zhang, Z. Zhang, N. Zhang, B. Qin, X. Ming, and Y. Zheng, PRGC: Potential relation and global correspondence based joint relational triple extraction, in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. Natural Language Processing*, Virtual Event, 2021, pp. 6225–6235.

[13] D. Zelenko, C. Aone, and A. Richardella, Kernel methods for relation extraction, in *Proc. 2002 Conf. Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA, 2002, pp. 71–78.

[14] L. Ratinov and D. Roth, Design challenges and misconceptions in named entity recognition, in *Proc. 13th Conf. Computational natural Language Learning*, Boulder, CO, USA, 2009, pp. 147–155.

[15] E. F. T. K. Sang and F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in *Proc. 7th Conf. Natural Language Learning at HLT-NAACL 2003*, Edmonton, Canada, 2003, pp. 142–147.

[16] Y. Wang, B. Yu, Y. Zhang, T. Liu, H. Zhu, and L. Sun, TPLinker: Single-stage joint extraction of entities and relations through token pair linking, in *Proc. 28th Int. Conf. Computational Linguistics*, Barcelona, Spain, 2020, pp. 1572–1582.

[17] R. Bunescu and R. Mooney, A shortest path dependency kernel for relation extraction, in *Proc. Human Language Technology Conf. and Conf. Empirical Methods in Natural Language Processing*, Vancouver, Canada, 2005, pp. 724–731.

[18] Q. Li and H. Ji, Incremental joint extraction of entity mentions and relations, in *Proc. 52nd Annu. Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, 2014, pp. 402–412.

[19] W. Xue, X. Xia, P. Wan, P. Zhong, and X. Zheng, Adversarial attack on object detection via object feature-wise attention and perturbation extraction, *Tsinghua Science and Technology*, vol. 30, no. 3, pp. 1174–1189, 2025.

[20] S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu, Joint extraction of entities and relations based on a novel tagging scheme, in *Proc. 55th Annu. Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 2017, pp. 1227–1236.

[21] M. Miwa and M. Bansal, End-to-end relation extraction using LSTMs on sequences and tree structures, in *Proc. 54th Annu. Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2016, pp. 1105–1116.

[22] X. Wang, Z. Liu, and J. Liu, Joint relational triple extraction with enhanced representation and binary tagging framework in cybersecurity, *Comput. Secur.*, vol. 144, p. 104001, 2024.

[23] H. Zheng, R. Wen, X. Chen, Y. Yang, Y. Zhang, Z. Zhang, N. Zhang, B. Qin, X. Ming, and Y. Zheng, PRGC: Potential relation and global correspondence based joint relational triple extraction, in *Proc. 59th Annu. Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. Natural Language Processing*, Virtual Event, 2021, pp. 6225–6235.

[24] Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin, Classifying relations via long short term memory networks along shortest dependency paths, in *Proc. 2015 Conf. Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 1785–1794.

[25] Y. Sun, J. Wang, H. Lin, Y. Zhang, and Z. Yang, Knowledge guided attention and graph convolutional networks for chemical-disease relation extraction, *IEEE/ACM Trans. Computat. Biol. Bioinf.*, vol. 20, no. 1, pp. 489–499, 2023.

[26] Z. Jia, Z. Yan, W. Han, Z. Zheng, and K. Tu, Modeling instance interactions for joint information extraction with neural high-order conditional random field, in *Proc. 61st Annu. Meeting of the Association for Computational Linguistics*, Toronto, Canada, 2023, pp. 13695–13710.

[27] J. Wang and W. Lu, Two are better than one: Joint entity and relation extraction with table-sequence encoders, in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing*, Virtual Event, 2020, pp. 1706–1721.

[28] Z. Yan, Z. Jia, and K. Tu, An empirical study of pipeline vs. joint approaches to entity and relation extraction, in *Proc. 2nd Conf. the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th Int. Joint Conf. Natural Language Processing*, Virtual Event, 2022, pp. 437–443.

[29] B. Stanoev, G. Mitrov, A. Kulakov, G. Mirceva, P. Lameski, and E. Zdravevski, Automating feature extraction from entity-relation models: Experimental evaluation of machine learning methods for relational learning, *Big Data Cogn. Comput.*, vol. 8, no. 4, p. 39, 2024.

[30] W. Xing, J. Qi, X. Yuan, L. Li, X. Zhang, Y. Fu, S. Xiong, L. Hu, and J. Peng, A gene–phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach, *Bioinformatics*, vol. 34, no. 13, pp. i386–i394, 2018.

[31] B. Stanoev, G. Mitrov, A. Kulakov, G. Mirceva, P. Lameski, and E. Zdravevski, Automating feature extraction from entity-relation models: Experimental evaluation of machine learning methods for relational learning, *Big Data Cogn. Comput*., vol. 8, no. 4, p. 39, 2024.

[32] Y. M. Shang, H. Huang, and X. Mao, OneRel: Joint entity and relation extraction with one module in one step, in *Proc. 36th AAAI Conf. Artificial Intelligence*, Virtual Event, 2022, pp. 11285–11293.

[33] G. Li, X. Chen, P. Wang, J. Xie, and Q. Luo, FastRE: Towards fast relation extraction with convolutional encoder and improved cascade binary tagging framework, in *Proc. 31st Int. Joint Conf. Artificial Intelligence*, Vienna, Austria, 2022, pp. 4201–4208.

[34] Y. Zhao and X. Li, A subject-aware attention hierarchical tagger for joint entity and relation extraction, in *Proc. 34th Int. Conf. Advanced Information Systems Engineering*, Leuven, Belgium, 2022, pp. 270–284.

[35] Q. Zhao, D. Xu, J. Li, L. Zhao, and F. A. Rajput, Knowledge guided distance supervision for biomedical relation extraction in Chinese electronic medical records, *Expert Syst. Appl.*, vol. 204, p. 117606, 2022.

[36] T. Zhang, H. Lin, M. M. Tadesse, Y. Ren, X. Duan, and B. Xu, Chinese medical relation extraction based on multi-hop self-attention mechanism, *Int. J. Mach. Learn. Cyber.*, vol. 12, no. 2, pp. 355–363, 2021.

[37] T. Zhao, Z. Yan, Y. Cao, and Z. Li, A unified multi-task learning framework for joint extraction of entities and relations, in *Proc. 35th AAAI Conf. Artificial Intelligence*, Virtual Event, 2021, pp. 14524–14531.

[38] W. Liu and Z. Wang, Event relation extraction based on heterogeneous graph attention networks and event ontology direction induction, *Tsinghua Science and Technology*, DOI: 10.26599/TST.2024.9010104.

[39] C. Gao, X. Zhang, L. Li, J. Li, R. Zhu, K. Du, and Q. Ma, ERGM: A multi-stage joint entity and relation extraction with global entity match, *Knowl.-Based Syst*., vol. 271, p. 110550, 2023.

[40] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*, Virtual Event, 2020, pp. 38–45.

[41] H. Hu, W. Zhao, W. Zhou, and H. Li, SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding, *IEEE Trans. Pattern Anal. Machi. Intell.*, vol. 45, no. 9, pp. 11221–11239, 2023.

[42] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, A robustly optimized BERT pre-training approach with post-training, in *Proc. 20th Chinese National Conf. Computational Linguistics*, Huhhot, China, 2021, pp. 1218–1227.

[43] C. Gao, X. Zhang, L. Li, J. Li, R. Zhu, K. Du, and Q. Ma, ERGM: A multi-stage joint entity and relation extraction with global entity match, *Knowl.-Based Syst*., vol. 271, p. 110550, 2023.

[44] K. Detroja, C. K. Bhensdadia, and B. S. Bhatt, A survey on relation extraction, *Intell. Syst. Appl.*, vol. 19, p. 200244, 2023.

[45] Z. Zhang and M. R. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels, in *Proc. 32nd Int. Conf. Neural Information Processing Systems*, Montréal, Canada, 2018, pp. 8792–8802.

[46] D. Chang, M. Chen, C. Liu, L. Liu, D. Li, W. Li, F. Kong, B. Liu, X. Luo, Q. Jin, et al., DiaKG: An annotated diabetes dataset for medical knowledge graph construction, in *Proc. 6th China Conf. Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction*, Guangzhou, China, 2021, pp. 308–314.

[47] D. Sui, X. Zeng, Y. Chen, K. Liu, and J. Zhao, Joint entity and relation extraction with set prediction networks, *IEEE*

Trans. Neural Netw. Learn. Syst., vol. 35, no. 9, pp. 12784–12795, 2024.

[48] F. Yang, S. Xu, P. Li, and Q. Zhu, Chinese event temporal relation extraction on multi-dimensional attention, in *Proc. 2023 Int. Joint Conf. Neural Networks*, Gold Coast, Australia, 2023, pp. 1–8.

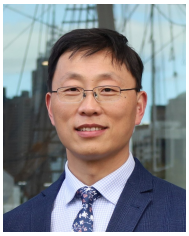
[49] B. Xu, S. Li, Z. Zhang, and T. Liao, BERT-PAGG: A Chinese relationship extraction model fusing PAGG and entity location information, *PeerJ Comput. Sci*., vo. 9, p. e1470, 2023.

[50] J. Shenouda, R. Parhi, K. Lee, and R. D. Nowak, Variation spaces for multi-output neural networks: Insights on multi-task learning and network compression, *J. Mach. Learn. Res.*, vol. 25, no. 1, p. 231, 2024.

[51] J. Li, K. Shuang, J. Guo, Z. Shi, and H. Wang, Enhancing semantic relation classification with shortest dependency path reasoning, *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1550–1560, 2023.

Qijie Qian received the MS degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2021. He is currently working toward the PhD degree at Nanjing University of Posts and Telecommunications, Nanjing, China. His main fields of research include artificial intelligence, Internet of Things, and future networks.

Xin Xu received the MS degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2024. He is currently an engineer at Nanjing University of Posts and Telecommunications, Nanjing, China. His main fields of research include artificial intelligence, natural language processing, and knowledge extraction.



Bo Li received the BS and MS degrees from Shandong Normal University, Jinan, China, in 2003 and 2010, respectively, and the PhD degree from Swinburne University of Technology, Melbourne, Australia, in 2023. He is currently a lecturer at Victoria University, Melbourne, Australia. His research interests include software engineering, edge computing, and network security.

Baoquan Ren received the PhD degree from the Institute of China Electronic System Engineering Corporation, Beijing, China, in 2010. He is now a senior engineer at School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include Internet of Things and wireless communication technology.

Yuan Miao received the PhD degree from Tsinghua University, Beijing, China, in 1996. He is currently a professor at Victoria University, Melbourne, Australia. His research interests are in human knowledge modeling, natural language comprehension, and decision support.

Yun Liu received the PhD degree from Nanjing Medical University, Nanjing, China, in 2000. She is currently a professor, a chief physician, and a PhD supervisor at the Center of Data Management, The First Affiliated Hospital, Nanjing Medical University, Nanjing, China. She is mainly engaged in medical informatics, hospital information management, and medical artificial intelligence research.

Yong'an Guo received the PhD degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2018. He is currently an associate professor at School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include Internet of Things, edge intelligence, and future networks.

Shenqi Jing received the PhD degree from Nanjing Medical University, Nanjing, China, in 2021. He is currently a senior engineer at Department of Medical Informatics, Nanjing Medical University, Nanjing, China. He is mainly engaged in medical informatics, knowledge map, and medical artificial intelligence research.