sensors

MDPI

*Article*

# A Joint Extraction System Based on Conditional Layer Normalization for Health Monitoring

**Binbin Shi [1], Rongli Fan [1], Lijuan Zhang [2], Jie Huang [2], Neal Xiong [3], Athanasios Vasilakos [4], Jian Wan [2] and Lei Zhang [2,*]**

1 School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China; sbb_edu_stu@126.com (B.S.); fanrongli@zust.edu.cn (R.F.)
2 School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China; 121107@zust.edu.cn (L.Z.); huangjie@zust.edu.cn (J.H.); wanjian@zust.edu.cn (J.W.)
3 Department of Computer Science, Mathematics Sul Ross State University, Alpine, TX 79830, USA; xiongnaixue@gmail.com
4 Center for AI Research, University of Agder, 4879 Grimstad, Norway; thanos.vasilakos@uia.no
* Correspondence: leizhang@zust.edu.cn

**Abstract:** Natural language processing (NLP) technology has played a pivotal role in health monitoring as an important artificial intelligence method. As a key technology in NLP, relation triplet extraction is closely related to the performance of health monitoring. In this paper, a novel model is proposed for joint extraction of entities and relations, combining conditional layer normalization with the talking-head attention mechanism to strengthen the interaction between entity recognition and relation extraction. In addition, the proposed model utilizes position information to enhance the extraction accuracy of overlapping triplets. Experiments on the Baidu2019 and CHIP2020 datasets demonstrate that the proposed model can effectively extract overlapping triplets, which leads to significant performance improvements compared with baselines.

**Keywords:** joint extraction; talking-head attention; Chinese medical texts; RoBERTa; health monitoring



**Citation:** Shi, B.; Fan, R.; Zhang, L.; Huang, J.; Xiong, N.; Vasilakos, A.; Wan, J.; Zhang, L. A Joint Extraction System Based on Conditional Layer Normalization for Health Monitoring. *Sensors* **2023**, *23*, 4812. https://doi.org/10.3390/s23104812

Academic Editors: Hanfei Mei and Victor Giurgiutiu

Received: 16 April 2023
Revised: 10 May 2023
Accepted: 11 May 2023
Published: 16 May 2023



## 1. Introduction

The emerging field of health monitoring aims to detect and prevent disease early by combining electronic medical records with health indicators; in general, the idea of health monitoring is to accurately extract key information from electronic medical records or medical texts [1]. The key information is usually represented in the form of triplets; therefore, as a branch of NLP [2], entity relation joint extraction is exploited to extract relational triplets between entities from texts. In the field of health monitoring, entity relation extraction helps to discover hidden knowledge and associations using massive health databases in order to optimize and improve health monitoring services [3].

The modeling of extracted entities and relations in early works primarily used the pipeline model [4]. The pipeline model separates the extraction task into two different subtasks, which then use two separate models to reduce computational costs. Although the pipeline model is flexible and simplifies the processing flow, there are many limitations on the management of the two independent models. First, a mistaken result during entity extraction has an impact on the relation extraction task, which can degrade the pipeline model's performance. Second, the pipeline model ignores internal connections and dependencies between the entity recognition and relation extraction tasks [5]. Third, the pipeline model cannot provide an proper solution to the problem of overlapping entities in the named entity recognition task; this results in redundant information from candidate entities without relationships, which is detrimental to information extraction performance and imposes a considerable burden on optimal allocation of computing resources during

the relationship extraction task. Therefore, joint models that extract entities and relations simultaneously have been proposed to overcome the above disadvantages.

In recent years, research on joint models has been predominantly based on English public datasets, and there have been few studies based on Chinese medical datasets. Therefore, joint extraction of entity relationship encounters many challenges with respect to Chinese medical texts. A primary challenge is that, because many medical entities involve multiple words, it is difficult for joint models to identify these entities. More importantly, the language of medical descriptions is quite complex, especially in electronic medical records of patients' past and current medical history [6]. Furthermore, disease entities are associated with symptomatic entities. For example, the joint model extracts the triplet "cancer–symptom–pain". The relationship placing "symptom" between "cancer" and "pain" can be matched correctly; hence, a more efficient method for joint extraction of entities and relations can be designed by utilizing position information and correlations between entities to improve the accuracy of the joint model.

In this paper, we propose a novel pointer network model based on joint entity and relation extraction for Chinese medical texts. Because of the excellent performance of RoBERTa on sentence encoding tasks, RoBERTa is first utilized as the encoding layer to encode sentences, thereby enhancing the connection between entities and extracting positional information in sentences. Next, based on the information encoded by RoBERTa, the position information is applied to strengthen the feature relationships in sentences and to fuse the different features occurring between entities. Finally, the talking-head attention mechanism is introduced to enhance interaction between relationships. Compared with baseline models on the Baidu2019 and CHIP2020 datasets, the proposed model is able to consider the semantic features in sentences while achieving improved accuracy.

In summary, the contributions of our work can be summarized as follows:

1. We propose a joint entity and relation extraction model to effectively remedy the problem of overlapping triplets in Chinese datasets.
2. The proposed model combines position information with a talking-head attention layer, adding additional semantic information and enhancing interaction between relationships.
3. Our proposed model outperforms existing models in terms of the F1 value, realizing and improvement of 0.05 and 0.16 on the Baidu2019 and CHIP2020 datasets, respectively, and is able to extract entity relations in highly overlapping and complex sample datasets.

The rest of the paper is organized as follows. First, Section 2 introduces related works. In Section 3, the proposed joint model is described in detail. In Section 4, the datasets and the baseline models used for comparisons are described and the predictive performance of the proposed model is evaluated. Finally, we present our conclusions and discuss future research directions in Section 5.

## 2. Related Work

The existing relation extraction models in the field of medicine can be divided into pipeline models and joint models [7]. In pipeline models, the relation extraction task is divided into two independent subtasks and a model is constructed and trained for each subtask. On the other hand, joint models share parameters between the extraction results to simultaneously extract entities and relationships. In the field of medicine, early works addressed information extraction in a pipelined manner to enhance the accuracy of entity extraction and relation extraction. More specifically, the pipeline model has been applied to extract relational triplets in two separate steps: named entity recognition (NER) and relational classification (RC) [8].

In NER, during early sequence tasks the traditional methods are based on rules [9], dictionaries [10], and machine learning [11–13]. However, with the deepening of research on deep learning, these methods are becoming widely used in sequence modeling in the data-driven era, with examples including Long-Short-Term Memory (LSTM) [14],

Condition-Random Field (CRF), and Bidirectional Encoder Representations from Transformers (BERT) [15]. With the development of deep learning, Liu et al. [16] applied the Bi-LSTM-CRF model to the Chinese clinical medical entity recognition system, utilizing CRF to achieve high micro-average F1-scores on multiple English datasets. LSTM can be used to recognize entities in specific formats; however, the accuracy of LSTM models on Chinese datasets is not acceptable. Therefore, Gridach et al. [17] introduced a novel neural network architecture that investigated a combination of LSTM, CRF, word embeddings, and character-level representation for recognition of biomedical named entities. Considering the characteristics of Chinese words segmentation, the model named Lattice LSTM [18] combined character information and lexical information for the first time while avoiding the influence of segmentation errors on LSTM. Zhao et al. [19] proposed a novel Chinese clinical named entity recognition method combining lattice LSTM and adversarial training; this method can improve the robustness of neural models by increasing perturbations in order to avoid the influence of segmentation errors. Li et al. [20] proposed a BERT–BiLSTM–CRF model for the medical field that simultaneously considers the characteristics of Chinese medical word segmentation and medical dictionary characters. The construction of the BERT–BiLSTM–CRF model has been applied to the field of EMR as well. Specifically, Gao et al. [21] introduced a BERT Chinese pre-training model able to automate feature selection, then combined BiLSTM and CRF to optimize the Chinese NER algorithm and applied the model to process an electronic medical record dataset. Because the LSTM model ignores context information, Kong et al. [22] sought to exploit the context information of short-term and long-term memory for the NER task by designing a simple attention mechanism that can improve training efficiency based on multiple Convolutional Neural Networks (CNN) in parallel. However, Multi-CNN models have difficulty capturing the spatial information between words. Therefore, Wang et al. [23] proposed an adversarial training LSTM–CNN system, which they called ASTRAL, to exploit position information between adjacent words. Unlike existing NER methods, ASTRAL improves the model structure and training process by introducing a Gated CNN to fuse the information of adjacent words.

In terms of medical RC, existing methods rely on medical features that can easily cause errors to accumulate during relation extraction, which then degrades the accuracy of feature extraction systems such as Recurrent Neural Networks (RNN) or other neural network-based methods. Hence, Fei et al. [24] proposed a BiLSTM–RNN model to learn the semantic features for relation extraction, and verified that LSTM–RNN can achieve better performance than LSTM on feature extraction. To improve the performance of RNNs, Zhang et al. [25] proposed a model that can automatically learn the features of sentence sequences by combining the RNN and CNN approaches for extracting biomedical relationships; however, this model fails to exploit the dependency types among words. To address this issue, a dependency-driven approach was proposed in [26] for relation classification using an attentive graph convolutional network (A-GCN), which applies a graph convolutional network-based attention mechanism to distinguish the importance of different dependencies between words. However, A-GCN lacks reliability when coding long sentences. Wang at el. [27] designed a structural block method to encode blocks associated with entities; the structural block method is able to eliminate noise caused by irrelevant parts of sentences to enhance the representation of relevant words. While the structural block method has the advantage of being independent of long sentence context, it only encodes the sequential tokens within a block boundary. Fortunately, BERT is able to encode long sentences, which has a significant impact on the natural language processing. To this end, Zhang et al. [28] introduced a clinical language model that used BERT for context pretraining, focusing on the importance of embedding in sentences. However, their clinical language model ignored the critical role of important phrase information. Therefore, Xu et al. [29] introduced a relational classification model based on BERT and a gated multi-window attention network (BERT–GMAN) to construct a key phrase classification network to obtain multi-granularity phrase information and exploit element-wise max-pooling to

select the features of key phrases; BERT–GMAN showed greatly improved accuracy on the relational classification task. In addition, referencing subsequent improvements to BERT, the BioBERT model [30] has obtained excellent performance as a language representation model on multiple datasets. Nevertheless, the above-mentioned models ignore the connections between entity extraction and the relation extraction, which may lead to error propagation and decrease the accuracy of the extraction results.

Recently, many researchers have attempted to alleviate the problem of error propagation when jointly extracting entities and relations by exploiting complex semantic features with a single model. Based on complex semantic features, the initial joint model introduced the dependency syntax tree [31], which was able to effectively capture the features of sentences by stacking bidirectional tree-structured LSTM–RNNs on bidirectional sequential LSTM–RNNs. In this approach, the entities and relations are jointly represented to share parameters, allowing for entity and relation extraction in a single model. Based on previous work on dependency syntax trees [31], Katiyar et al. [32] replaced the dependency syntax tree with an LSTM network to determine different relations by comparing the similarity of entities with other entities. However, this approach ignores the entity boundary information in sentences. Hence, to enhance the accuracy of the pointer network in the process of decoding, Gu et al. [33] and Zeng et al. [34] introduced the copy mechanism to generate the relation and proposed the CopyNet and End2End Neural models, respectively. Among these, the CopyNet model uses a novel method of word generation based on the copying mechanism to choose proper sequences in the input and place them in the proper position in the output. On the other hand, the End2End Neural model utilizes the copy mechanism to jointly extract relational facts from sentences in the overlapping triplets class using different decoder strategies. Giannis et al. [35] proposed a multi-head selection model. This model considers the entity boundary information in sentences by introducing a CRF layer into the entity recognition task, thereby transforming the relationship extraction task into a multi-head selection problem. Although the multi-head selection model has demonstrated its effectiveness on multiple datasets, the model ignores robust generalizations during entity training as well as gaps between entities during the entity prediction process. To alleviate the model's problems with robust generalizations and gaps, adversarial training mechanisms [36] and soft label embedding [37] have been proposed. Unlike the existing joint models, ETL-span [38] is a novel framework that can adequately capture semantic dependencies between different steps to remove noise between pairs of entities. Dixit et al. [39] directly extracted span-level features on the basis of the ETL–SPAN model; this model is able to pay attention to overlapping entities, avoiding erroneous information transmission cascade; unfortunately, the performance of this model in extracting triplets from long sentences is not satisfactory. To address the issue of relation extraction in long sentences, Eberts et al. [40] used BERT to encode long sentences and fuse multiple features before classifying relationships. In recent years, with the help of BERT, researchers have paid more attention to the connections between entities and relationships. Luo et al. [41] proposed a neural network framework called the ATT–BiLSTM–CRF model, which uses an attention mechanism for biomedical joint extraction. The ATT–BiLSTM–CRF model effectively strengthens the connection between the biomedical entity and the relationship. Hong et al. [42] proposed a joint entity relation extraction model based on a graph convolutional network (GCN) to effectively distinguish the interaction between entities and relationships. Lai et al. [43] designed a new multi-attentional mechanism to improve the performance of graph attentional networks, although this did not result in any improvement on overlapping relationships. Considering the issue of overlapping relationships, Wei et al. [44] proposed the Casrel model for relationship extraction, then merged the original sequence information using on BERT. To address the challenge of overlapping relationships, in this paper we propose a novel model that utilizes RoBERTa to encode sentences in the encoding layer and exploit the position information in order to enhance the feature representation of Chinese sentences. In addition, conditional layer normalization is combined with talking-head attention to alleviate the problem of overlapping triplets.

## 3. Proposed Joint Model

In this section, we present a description of the proposed joint extraction model motivated by the previous work of Wei et al. [44], in which the head entity is defined randomly for relation extraction without completely traversing all entities during entity extraction. The proposed model can identify all possible triplets in a sentence, even where a few triplets may share the same entities or the same relations. More specifically, the proposed model extracts triplets using two modules: a RoBERTa encoding module, and a cascade decoding module. In the RoBERTa encoding module, RoBERTa is applied to fully extract sentence features and identify connections between words in sentences. In the cascade decoding module, subject extraction is applied to find all possible subjects in the sentence. Then, relation–object extraction is applied to find all relevant relations and their corresponding objects. The cascade decoding module is designed with multi-level training objectives, simplifying the entity extraction process. The specific block diagram of the proposed model is shown in Figure 1. Among them, $a_i(i = 1, 2, 3)$, $b_j(j = 1, 2, 3, 4)$, and $c_k(k = 1, 2, 3)$ represent "pancreatic cancer", "ultrasonic examination", and "pancreatic masses", respectively, in Chinese characters. The RoBERTa encoding module and cascade decoding module are elaborated upon in the following subsections.

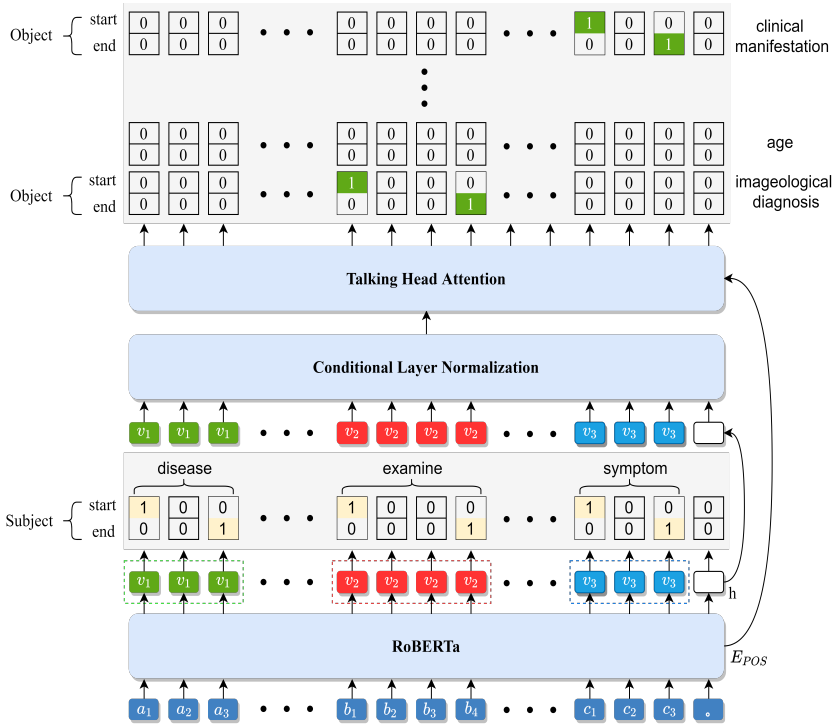



**Figure 1.** Block diagram of the proposed model.

### 3.1. RoBERTa Encoder

The RoBERTa encoder utilizes RoBERTa as an encoder to extract feature information from sentences. RoBERTa is a bidirectional coding representation algorithm based on the transformer algorithm for feature extraction and sentence modeling [45]. RoBERTa is able to learn deep representations by jointly conditioning on context, and can fine-tune additional output layers to perform efficiently on many downstream tasks.

In the encoder module, RoBERTa is applied to encode the sentence, which is then fed into subsequent decoder modules. Specifically, given a sentence $\mathbf{x}$ of length $n$, every word

$\mathbf{x}_t$ in the sentence can be transformed into a token embedding, segment embedding, and position embedding. Hence, the output $\mathbf{x}_t$ of RoBERTa is as follows:

$$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_n\} \quad (1)$$

$$\mathbf{x}_t = \mathbf{E}_S + \mathbf{E}_T + \mathbf{E}_P \quad (2)$$

where $\mathbf{E}_T$ represents a word vector $\mathbf{E}_{token}$, $\mathbf{E}_S$ represents a segment vector $\mathbf{E}_{segment}$, and $\mathbf{E}_P$ represents a position vector $\mathbf{E}_{position}$.

The detailed flow of RoBERTa in a Chinese sentence is shown in Figure 2, where $a_i(i = 1, 2, 3)$ and $b_j(j = 1, 2, 3, 4)$ represent "pancreatic cancer" and "ultrasonic examination", respectively, in Chinese characters. The twelve-layer transformer encoder of RoBERTa is utilized to encode the Chinese sentence with bidirectional coding representation, then three feature vectors are applied to reconstruct the sentence from the noisy data [46]. The feature vectors are then passed to the cascade module for subject extraction.

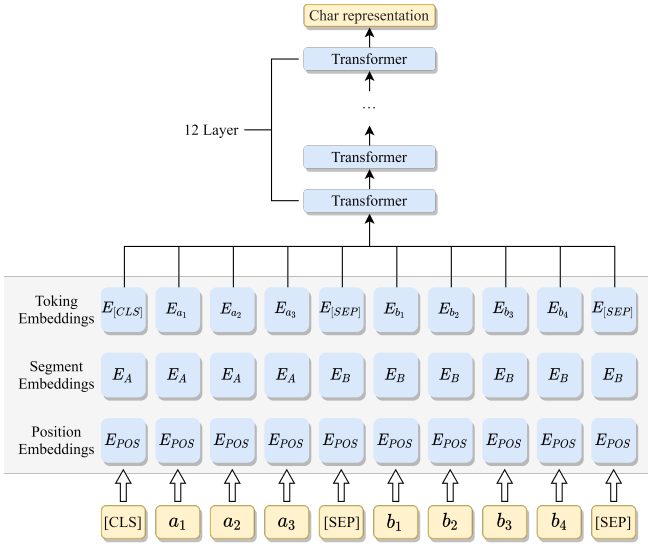



**Figure 2.** Block diagram of RoBERTa applied to a Chinese sentence.

### *3.2. Cascade Decoder*

The cascade module is adapted to extract triplets provided by the previous feature vectors. Specifically, the cascade decoder is divided into two cascaded steps, namely, subject extraction and relationship–object extraction. First, subject extraction detects subjects for each input sentence, including both head and tail entities. Second, relation–object extraction checks all possible relations to determine whether relations can be matched to the head and tail entities in the sentence. In addition, talking-head attention is utilized to obtain the relationships behind the conditional layer, which can enhance the accuracy of the cascaded steps. In the following subsections, the subject extraction and relationship–object extraction procedures are described in detail, as is the training algorithm used in the proposed model.

#### 3.2.1. Subject Extraction

In the subject extraction process, all possible subjects are recognized by directly decoding the feature vector **h** produced by the RoBERTa encoder. More specifically, this is essentially a binary classification problem that assigns each token a binary tag (0/1) that represents the start or end position of the identified subject by initializing a pointer network. Then, the subject is entered as a head entity in the module at the next level. The detailed operations used for subject extraction are as follows:

$$\begin{bmatrix} \mathbf{s}^{start} \\ \mathbf{s}^{end} \end{bmatrix} = \sigma\left(\begin{bmatrix} \mathbf{W}_s^{start} \\ \mathbf{W}_s^{end} \end{bmatrix} \times \mathbf{h} + \begin{bmatrix} \mathbf{b}_s^{start} \\ \mathbf{b}_s^{end} \end{bmatrix}\right) \quad (3)$$

where $\mathbf{s}^{start}$ represents the probabilities of the start position of all subjects for the input sentence and $\mathbf{s}^{end}$ represent the probabilities of the end position of all subjects for the input sentence. If the probability of a subject $s$ exceeds a certain threshold, the corresponding tag is assigned a value of 1; otherwise, it has a value of 0. Here, $\mathbf{h}$ is the encoded representation for the input sentence, $\mathbf{W}_s^{start}$ and $\mathbf{W}_s^{end}$ represent the weight matrix of the start and end positions in the full connection layer, and are updated automatically, $\mathbf{b}_s^{start}$ and $\mathbf{b}_s^{end}$ are the respective offset vectors of the start and end positions, and $\sigma$ is the sigmoid activation function used to map the output.

Next, the span of subject $s$ in the input sentence $\mathbf{x}$ is optimized using the following likelihood function $p_\theta(s \mid \mathbf{x})$:

$$p_\theta(s \mid \mathbf{x}) = \prod_{t \in (start, end)} \prod_{i=1}^{L} \left(s_i^t\right)^{R_1^t} \left(1 - s_i^t\right)^{R_2^t} \tag{4}$$

where $L$ is the length of the input sentence, $R_1^{start}$ and $R_2^{start}$ are marked as 1 and 0, respectively, if the subject start position is marked as 1 in the output start position sequence, $R_1^{end}$ and $R_2^{end}$ are marked as 0 and 1, respectively, if the subject end position is marked as 1 in the output end position sequence, and the parameter $\theta$ is $\left\{\mathbf{W}_s^{start}, \mathbf{b}_s^{start}, \mathbf{W}_s^{end}, \mathbf{b}_s^{end}\right\}$. Moreover, the function $p_\theta(s \mid \mathbf{x})$ is exploited to evaluate the subject extraction performance.

For subject detection, the match principle for the nearest start–end distance is adopted to decide the span of subjects. For example, as shown in Figure 1, the matrices with the marked start and end positions of the three entities "pancreatic cancer", "ultrasonic examination", and "pancreatic mass" are obtained after the RoBERTa encoding layer. As the start token matches the nearest end token, the result of the first entity is "pancreatic cancer". Based on the match strategy, the proposed model only considers those end tokens with positions behind the existing start token. More importantly, the match strategy can maintain the integrity of entities to the greatest extent possible.

#### 3.2.2. Relation–Object Extraction

Relation–object extraction simultaneously recognizes objects and their involved relations based on the previously obtained subjects. As shown in Figure 1, relation–object extraction consists of conditional layer normalization (CLN) and talking-head attention (THA). First, CLN is used to determine a specific category and randomly generate contexts based on this category. In particular, CLN [47] utilizes a fixed-length vector as a conditional scenario to incorporate the conditions **fi** and **fl** into normalization. Moreover, in CLN the feature $\mathbf{h}$ and two conditions are fused to combine relation features with the input entity features. Hence, the output $\hat{\mathbf{h}}$ of CLN is as follows:

$$\hat{\mathbf{h}} = [(\mathbf{h} - \mathbf{avg}) \div \mathbf{std}] \times \gamma + \beta \tag{5}$$

where **avg** is the mean value of $\mathbf{h}$, **std** is the standard deviation of $\mathbf{h}$, and $\beta$ and $\gamma$ are two dynamic matrices that are influenced by the input subject in the sentence. Two different matrices that can be transformed by the different entities for initializing $\beta$ and $\gamma$ in the same dimension are exploited. In addition, two matrices $\beta, \gamma$ and the feature $\mathbf{h}$ are combined to obtain the feature $\hat{\mathbf{h}}$ that is affected by subject $s$. When merging these vectors, it is crucial that the dimension output by CLN remains consistent with the original pretraining model.

In order to exploit the parameters of entity recognition while improving the accuracy of relation extraction, the output matrix is spliced into CLN using the position information from RoBERTa, then the subsequent THA utilizes the matrix. The proposed model combines a feature $\hat{\mathbf{h}}$ with the position information to obtain the matrix $\mathbf{H}$, as follows:

$$\mathbf{H} = \hat{\mathbf{h}} + \mathbf{E}_P \tag{6}$$

To enhance the effectiveness of feature extraction, the proposed model uses attention mechanisms. Compared with talking-head attention, multi-head attention [48] only focuses on the performance of each head, ignoring the relevance of the heads. The formulas for multi-head attention are as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{7}$$

$$\mathbf{head}_i = Attention\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right) \tag{8}$$

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(\mathbf{head}_1, \mathbf{head}_2 \cdots, \mathbf{head}_h)\mathbf{W}^O \tag{9}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are converted from $\mathbf{H}$, while $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, and $\mathbf{W}_i^V$ represent the respective weight parameters of $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ for the $i$th calculation; moreover, $d_k$ represents the dimension of $\mathbf{V}$, $softmax$ is the softmax activation function, the single-head attention $\mathbf{head}_i$ is calculated using Equation (8), and $Attention(\cdot)$ in Equation (8) is illustrated in Equation (7). Finally, by repeating Equation (8) $h$ times, the multiple attention $MultiHead(\cdot)$ is obtained based on the corresponding results for $h$ times, where $\mathbf{W}^O$ is the weight parameter and is automatically updated.

By linking the heads together, talking-head attention [49] can exploit more information from different representation subspaces at different positions. In addition, the information includes location information, syntax information, and other information. Hence, talking-head attention utilizes two additional learned matrices $\lambda_i^W$ and $\lambda_i^L$ to fuse head attention into talking-head attention. Therefore, the formulas for talking-head attention are as follows:

$$A(\mathbf{Q}, \mathbf{K}) = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \tag{10}$$

$$\mathbf{J}_i = \lambda_i^L softmax\left(\lambda_i^W A(\mathbf{Q}_i, \mathbf{K}_i)\right) \tag{11}$$

$$\mathbf{O} = TalkingHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(\mathbf{J}_1\mathbf{V}_1, \cdots, \mathbf{J}_h\mathbf{V}_h) \tag{12}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are converted from $\mathbf{H}$, $A(\mathbf{Q}_i, \mathbf{K}_i)$ represents the $i$th calculation of single-head attention, $softmax$ is the softmax activation function, and $\mathbf{J}_i$ indicates that single-head attention is associated with other attentions, where $\lambda_i^W$ and $\lambda_i^L$ can move information across attention heads by transforming the attention-logits and attention-weights, respectively. The output of talking-head attention $TalkingHead(\cdot)$ concatenates the calculation of attention for all heads.

Although the subjects are obtained by decoding the feature vector $\mathbf{h}$ during subject extraction, the head entity features are exploited during relation extraction to enhance the connection between the relations and the entities. Therefore, based on the features of the head entities in the sentences, the output of the relation extraction is as follows:

$$\begin{bmatrix} \mathbf{r}^{start} \\ \mathbf{r}^{end} \end{bmatrix} = \sigma\left(\begin{bmatrix} \mathbf{W}_r^{start} \\ \mathbf{W}_r^{end} \end{bmatrix} * (\mathbf{O} + \mathbf{v}_i) + \begin{bmatrix} \mathbf{b}_r^{start} \\ \mathbf{b}_r^{end} \end{bmatrix}\right) \tag{13}$$

where $\mathbf{r}^{start}$ and $\mathbf{r}^{end}$ represent the probabilities of the respective start and end positions of all relations in the input sentence; the corresponding token are marked as 1 if the probability of the start and end positions exceeds a certain threshold, and are 0 otherwise. Moreover, $\mathbf{v}_i$ represents the encoded representation vector between the start and end tokens

of the $i$th subject detected in the subject extraction module, $\mathbf{W}_r^{start}$ and $\mathbf{W}_r^{end}$ represent the weight matrix of start and end positions relative to the relations, $\mathbf{b}_r^{start}$ and $\mathbf{b}_r^{end}$ are the respective offset vectors of the start and end positions of the relations, and $\sigma$ is the sigmoid activation function.

More specifically, in order to achieve the fusion of $\mathbf{O}$ and $\mathbf{v}_i$ in Equation (13), it is necessary to ensure that the dimension of the two vectors remains consistent during the relation–object extraction process. For each subject $\mathbf{v}_i$, all subjects are traversed to extract triplets, while the subject is randomly selected for each sentence through the Casrel model [44]. Although the proposed model incurs higher computational cost during relation extraction, this results in improved accuracy during triplet extraction.

Next, the relation representations of the object $o$ and subject $s$ in the input sentence $\mathbf{x}$ are optimized using the following likelihood function $p_\theta(o \mid s, \mathbf{x})$:

$$p_\theta(o \mid s, \mathbf{x}) = \prod_{t \in \{start, end\}} \prod_{i=1}^{L} \left(r_i^t\right)^{R_1^t} \left(1 - r_i^t\right)^{R_2^t} \tag{14}$$

where $L$ is the length of the sentence, $R_1^{start}$ and $R_2^{start}$ are marked as 1 and 0, respectively, if the object start position is marked as 1 in the output start position sequence, $R_1^{end}$ and $R_2^{end}$ are marked as 0 and 1, respectively, if the object end position is marked as 1 in the output end position sequence, and the parameter $\theta$ is $\left\{\mathbf{W}_r^{start}, \mathbf{b}_r^{start}, \mathbf{W}_r^{start}, \mathbf{b}_r^{end}\right\}$. In addition, the function $p_\theta(o \mid s, \mathbf{x})$ is exploited to evaluate the relationship extraction performance.

As shown in Figure 1, for the output of each sentence a matrix is constructed to calculate the matching result between entities and relations. For example, the subject "pancreatic cancer" compares the relationships between "imageological examination", "age", and "clinical manifestation" with different objects, and all relations are traversed to determine the object that can be used to construct a triple with the subject "pancreatic cancer". Finally, the two triplets "pancreatic cancer–imageological examination–ultrasonic examination" and "pancreatic cancer–clinical manifestation–pancreatic mass" are found.

For all training sets, the likelihood functions of entities and relations are optimized for each sentence $\mathbf{x}$. The optimizer utilizes the Adam [50] loss function to maximize $K$ by optimizing $p_\theta(s \mid \mathbf{x}_i)$ and $p_\theta(o \mid s, \mathbf{x}_i)$, which dynamically reduces the learning rate based on the number of times while increasing the model's efficiency and effectiveness. The indicator $K$ is written as follows:

$$K = \sum_{i=1}^{|D|} \left[ \sum_{s \in T_i} \log p_\theta(s \mid \mathbf{x}_i) + \sum_{r \in T_r} \log p_\theta(o \mid s, \mathbf{x}_i) \right] \tag{15}$$

where $|D|$ represents the cardinality of the training set, $T_i$ represents all subjects in the sentence, $T_r$ represents all relationships corresponding to the head entity, $p_\theta(s \mid \mathbf{x}_i)$ is defined in Equation (4), and $p_\theta(o \mid s, \mathbf{x}_i)$ is defined in Equation (14).

In the above, $K$ is the key indicator that determines when the training process of the proposed model terminates. Specifically, if $K$ is continuously updated until it reaches a steady state, then the training process is terminated. The training algorithm used for the proposed model is provided in Algorithm 1.

**Algorithm 1:** Training Algorithm of the Proposed Model

**Input:** Training dataset $D$, training epochs $N$
**Output:** Optimal parameter $K$
**Pre-trained**: Use RoBERTa to obtain the encoded feature vector **h** **Initialize**: Transform Chinese sentences into vector representations and initialize the model parameters
**for** $i = 0$ *to* $|D|$, **do**
  Select the $i$th sentence $\boldsymbol{x}^i$=$\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ from the training set $D$;
  **for** $j = 0$ *to* $N$, **do**
    Obtain the position information $\mathbf{E}_P$ using Equations (1) and (2) through the embedding module of RoBERTa;
    Obtain the probabilities of the start positions $\mathbf{s}^{start}$ and end positions $\mathbf{s}^{end}$ of the subjects using Equation (3);
    Obtain the likelihood function $p_\theta(s \mid \mathbf{x})$ using Equation (4) based on $\mathbf{s}^{start}$ and $\mathbf{s}^{end}$ in subject extraction;
    Obtain $\hat{\mathbf{h}}$ using Equation (5) in conditional layer normalization;
    Caculate the matrix $\mathbf{H}$ based on $\hat{\mathbf{h}}$ using Equation (6);
    Obtain $\mathbf{O}$ from Equations (10)–(12) in the talking-head attention layer;
    Obtain the probabilities of the start positions $\mathbf{r}^{start}$ and the end positions $\mathbf{r}^{end}$ of the relations using Equation (13);
    Obtain the likelihood function $p_\theta(o \mid s, \mathbf{x})$ using Equation (14) based on $\mathbf{r}^{start}$ and $\mathbf{r}^{end}$ in relation–object extraction;
    Update the $K$ parameter using Equation (15);
  **end**
**end**
Return $K$;

## 4. Performance Analysis

In this section, our experiments are mainly introduced from three aspects. First, the data sources and components of the Baidu2019 and CHIP2020 datasets are explained. Second, the experimental setup is described in detail, including the implementation details, evaluation metrics, and experimental environment settings. Third, the superiority of the proposed model is verified by comparison with baseline models.

### *4.1. Datasets*

The experiments were carried out on the Baidu2019 datasets and CHIP2020 datasets; the detailed description of the structures of these two datasets is as follows:

- Baidu2019 [51] contains sentences extracted from Baidu Baike and Baidu News Feeds; it is the largest Chinese information extraction dataset on the basis of schema, including more than 190,000 real-world Chinese sentences, more than 400,000 triplets, and 50 types of prespecified relations. To improve dataset availability, the dataset is divided into a training set and testing set using a certain proportion.
- CHIP2020 [52] is a Chinese medical dataset collected by the National Language Processing Laboratory of Zhengzhou University and the Key Laboratory of Computational Linguistics of the Ministry of Education of Peking University. It contains more than 17,000 Chinese medical sentences, more than 50,000 triplets, and 43 types of prespecified relations. The CHIP2020 dataset consists of diseases, symptoms, imaging tests, and other medical information. Moreover, the dataset is divided into a training set and testing set to enhance dataset standardization by establishing an official method.

The statistics of the two datasets are listed in Table 1 to further illustrate the characteristics of the Baidu2019 and CHIP2020 datasets. In addition, the datasets each include a training set and testing set that can be divided into three categories, as shown in Table 2,

that is, Normal, Entity Pair Overlap (EPO), and Single Entity Overlap (SEO) [4,38]. Below, the results of different categories are analyzed in detail.

**Table 1.** Statistics of the two datasets.

| Statistics | Baidu2019 | | CHIP2020 | |
|---|---|---|---|---|
| | **Train** | **Test** | **Train** | **Test** |
| sentence | 172,983 | 21,626 | 14,339 | 3585 |
| riplets | 363,895 | 45,558 | 43,660 | 10,626 |
| relations | | 50 | | 43 |

**Table 2.** Statistics of different categories of triplets in the datasets.

| Category | Baidu2019 | | CHIP2020 | |
|---|---|---|---|---|
| | **Train** | **Test** | **Train** | **Test** |
| Normal | 80,310 | 9984 | 5724 | 1496 |
| EPO | 19,049 | 2385 | 6937 | 1655 |
| SEO | 73,596 | 9257 | 1678 | 434 |

### *4.2. Experimental Setup*

First, a number of the parameters in the proposed model are introduced. Specifically, the batch size was set to 8, the learning rate was set to $1 \times 10^{-5}$, and the maximum length of the input sentence was set to 128. In addition, the number of heads for talking-head attention was 48, the number of transformer blocks was 12, and the size of the hidden state was 768. A stopping mechanism that ends the training process was adopted in the experiments. To measure the accuracy of the experimental results, the precision (pre), recall (rec), and F1-score are considered as the scoring functions, and can be written as shown below:

$$pre = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}} \times 100\% \tag{16}$$

$$rec = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \times 100\% \tag{17}$$

$$F1 = \frac{2 \times pre \times rec}{pre + rec} \times 100\% \tag{18}$$

where **TP** represents the number of correctly predicted triplets, **FP** represents the number of incorrectly predicted triplets, and **FN** represents the unpredicted triplets; *pre* explains the ratio of correctly predicted triplets to all predicted triplets, *rec* explains the ratio of correctly predicted triplets to all triplets in the datasets, and $F1$ provides a comprehensive evaluation of the results of precision and recall.

Next, the experimental environment is described concretely. As shown in Table 3, all models were implemented in TensorFlow-gpu 2.2.0 and trained on an Ubuntu 18.04 system with 32 GB of memory and an Nvidia 2080 GPU.

**Table 3.** Experimental environment settings.

| Item | Environment |
|---|---|
| Operating system | Ubuntu 18.04.5 LTS |
| CPU | i7-8700 @3.20 GHz |
| GPU | NVIDIA GeForce RTX 2080Ti |
| Memory | 31 G |
| Python version | 3.7 |
| TensorFlow [53] version | TensorFlow-gpu 2.2.0 |
| Transformers [54] version | 3.1.0 |

### *4.3. Results and Discussion*

In this section, the results of our comparative experiments and ablation experiments are analyzed and discussed. Overlapping experiments were constructed on different types of sentences and the performance of the proposed model was compared with baselines. In addition, the results of parametric experiments on inference methods were applied to verify the effectiveness of the talking-head attention mechanism. Finally, the experiments selected different sentences in order to judge the accuracy of the proposed model in a case study.

#### 4.3.1. Comparative Experiment with Existing Research Works

In the experiments, several baseline methods were considered for comparisons:

- MultiR [55]: a multi-instance learning algorithm combining a sentence-level extraction model with a simple corpus-level module, which alleviates the problem of noise caused by labeling.
- CoType [56]: an extraction model that jointly utilizes text features and type labels when carrying out entity and relationship extraction, which considers the problem of overlapping.
- Multi-head selection [35]: a neural model that identifies multiple relations for each entity to perform relation extraction; it can simultaneously train an entity recognition module and relationship extraction module.
- Casrel [44]: a joint model designed as a novel cascade binary tagging framework derived from a principled problem formulation.
- ETL-span [38]: a specific label scheme that decomposes entity recognition and relationship extraction into several labeling problems to extract multiple triplets.

In order to more comprehensively verify the effectiveness of the proposed model, experiments were conducted on the Baidu2019 and CHIP2020 datasets by comparing it with baseline models. Table 4 shows a comparison of precision, recall, and F1 between the proposed model and the baseline methods on the Baidu2019 and the CHIP2020 datasets.

**Table 4.** Comparisons with different methods on the Baidu2019 and the CHIP2020 datasets.

| Method | Baidu2019 | | | CHIP2020 | | |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F1** | **Precision** | **Recall** | **F1** |
| MultiR [55] | 0.634 | 0.389 | 0.482 | 0.261 | 0.378 | 0.312 |
| CoType [56] | 0.729 | 0.703 | 0.716 | 0.344 | 0.497 | 0.41 |
| Multi-head attention [35] | 0.764 | 0.712 | 0.737 | 0.412 | 0.572 | 0.471 |
| Casrel [44] | 0.800 | 0.720 | 0.758 | 0.42 | 0.581 | 0.48 |
| ETL-Span [38] | 0.779 | 0.801 | 0.790 | 0.41 | 0.633 | 0.494 |
| Ours | 0.801 | 0.838 | 0.809 | 0.566 | 0.767 | 0.64 |

Table 4 shows the results on the Baidu2019 and CHIP2020 datasets. For the Baidu2019 dataset, it can be observed from Table 4 that the proposed model outperforms the best baseline models by 1.9% in triplet extraction. This improvement can be explained by the employment of the cascade decoder, which can accurately capture multiple relations. In addition, the proposed model achieves a 5.1% improvement in F1-score over the Casrel model on the Baidu2019 dataset. Unlike the Casrel model, the proposed model exploits RoBERTa to capture the semantic features in sentences and utilizes a talking-head attention mechanism to obtain more effective attention. The results show that the proposed model performs well on the task of feature extraction on Chinese datasets. Considering the results on the CHIP2020 dataset, it can be observed from Table 4 that the proposed model overwhelmingly outperforms all the baselines in terms of all evaluation metrics; in particular, it achieves a 14.6% improvement in F1-score over the ETL-Span model on the CHIP2020 dataset. Moreover, the pre-trained RoBERTa and talking-head attention are utilized to effectively extract single triplets and overlapping triplets from the Chinese medical dataset.

In addition, the results on these datasets show that there is a significant gap between the general field and the medical field when extracting triplets, as the proposed model has more difficulty dealing with overlapping triplets in the field of medicine. More precisely, as shown in Table 2, the Baidu2019 dataset mainly consists of the Normal and SEO classes, while the CHIP2020 dataset mainly includes the Normal and EPO classes. This inconsistent distribution of categories between the two datasets leads to better performance on the Baidu2019 dataset. Nonetheless, the proposed model achieves a smaller gap between the Baidu2019 and CHIP2020 datasets than the baseline models, which demonstrates its superior effectiveness on the task of extracting overlapping triplets in medical contexts.

#### 4.3.2. Ablation Experiments

Using the Baidu2019 and CHIP2020 datasets, our ablation experiments focused on the contribution of the RoBERTa encoder, CLN layer, and THA layer. Each time, a module in the RoBERTa encoder, CLN layer, or THA layer was removed to obtain the effect of that module on the proposed model. First, it can be seen from Table 5 that if the RoBERTa encoder is removed on the Baidu2019 and CHIP2020 datasets, the F1 score is reduced by 3.9% and 3%, respectively. These results verify that the RoBERTa encoder can effectively extract Chinese sentence features. Removing the CLN layer has have an apparent degradation effect on the F1-score as well, which indicates that combining the position information with the encoder feature is beneficial to the process of extracting triplets. When comparing with the models with and without the THA layer, it is clear that the THA layer provides a remarkable improvement in the F1-score, which demonstrates that talking-head attention can effectively improve the accuracy of overlapping triplet extraction.

**Table 5.** Results of the ablation experiments.

| Method | Baidu2019 | CHIP2020 |
|---|---|---|
| | F1 | F1 |
| Ours | 0.809 | 0.64 |
| -RoBERT | 0.77 | 0.61 |
| -CLN | 0.778 | 0.62 |
| -THA | 0.782 | 0.61 |

#### 4.3.3. Analysis of Overlapping Triplets

To verify the ability of the proposed model to alleviate the problem of overlapping triplets, experiments were conducted on three categories of sentences and its performance was compared with the baseline models.

The results of the comparison between the proposed model and the baseline models on three categories of sentences are shown in Figure 3a–c. The results show that the proposed model achieves the best results on the Normal class, EPO class, and SEO class. Compared with the Casrel model, the F1-value improves by 10.9% and 11.9% on the EPO class of Baidu2019 and CHIP2020, respectively. Moreover, the F1 value improves by 6% and 16.2% for the SEO class. Indeed, among three categories of overlapping classes, the EPO and SEO classes are relatively complex collections of triplets. In contrast, the proposed model achieves consistently outstanding performance, especially for the EPO class, which shows that talking-head attention can alleviate problems on the EPO class by enhancing the relevance of features.

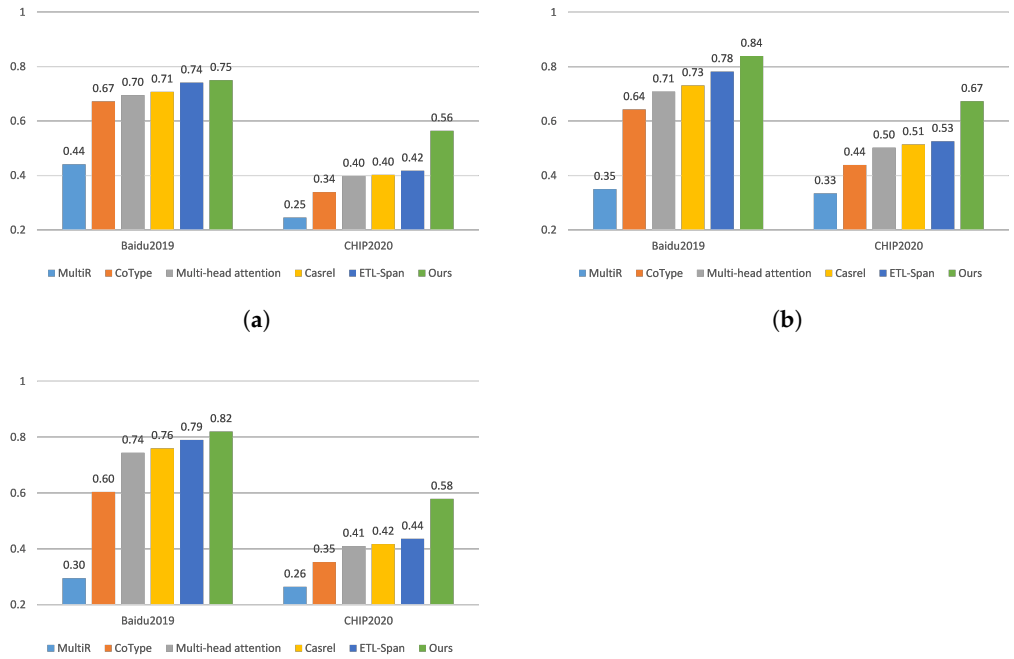



**Figure 3.** F1-score when extracting relational triplets from sentences on the different classes: (**a**) Normal class; (**b**) EPO class; (**c**) SEO class.

#### 4.3.4. Inference Method

To test the talking-head attention mechanism [49], experiments were structured comparing different heads to verify the reliability of the mechanism. Figure 4a,b shows the results of different heads on the Baidu2019 and CHIP2020 datasets.

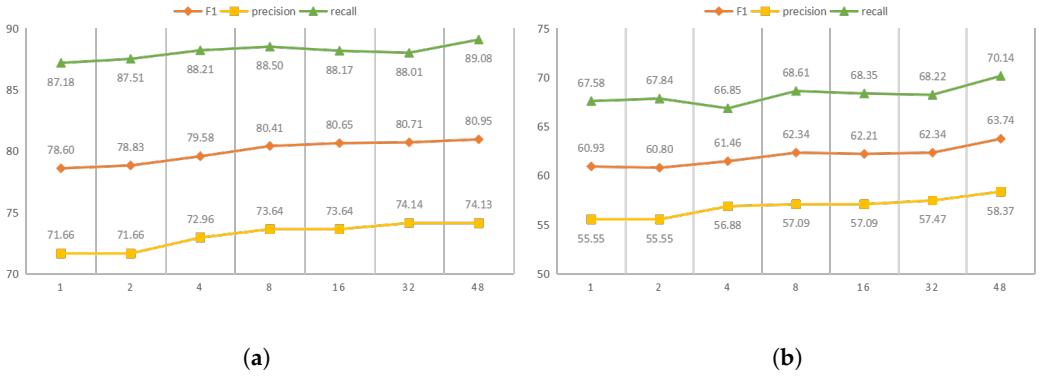



**Figure 4.** The performance on different datasets with different heads: (**a**) Baidu2019 dataset and (**b**) CHIP2020 dataset.

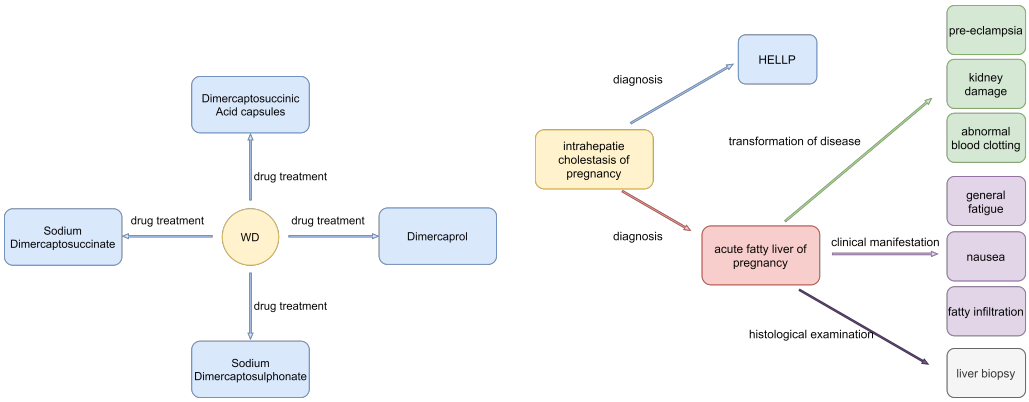
It can be seen that talking-head attention efficaciously adjusts the trade-off between precision and recall with different choices of heads. It can be seen that when increasing the number of head from 1 to 48, the F1-score significantly increases by 2.4% and 2.8% on the Baidu2019 and CHIP2020 datasets, respectively. Furthermore, the proposed model works more effectively on both the Baidu2019 and CHIP2020 datasets as the number of heads increases. Due to the limitations of the experimental environment, 48 was chosen as the maximum number of heads.

#### 4.3.5. Case Study

To specifically and intuitively observe the ability of the proposed model in overlapping extraction, triplets were extracted from complex sentences selected from the CHIP2020 dataset and its performance was analyzed. For ease of understanding, the English annotations of the Chinese sentences selected from CHIP2020 dataset are shown in Table 6.

**Table 6.** Results of the case study on different sentences.

| Sentence | Text | Our Model |
|---|---|---|
| case_1 | Trientine is also a complexing agent, which can promote the excretion of copper. It is sometimes used as a first-line drug in WD patients with neurological symptoms. It is effective in all types of patients, and the general dose is 40–50 mg/(kg·d). Other copper drugs: Dimercaprol (because of side effects have been less), Sodium Dimercaptosuccinate, Dimercaptosuccinic Acid capsules and Sodium Dimercaptosulphonate and other heavy metal chelate agents. | see Figure 5a |
| case_2 | Intrahepatie cholestasis of pregnancy(HELLP) is acute fatty liver of pregnancy. The patient developed the classic symptoms of general fatigue, nausea, pre-eclampsia, abnormal blood clotting and kidney damage. Liver biopsy showed fatty infiltration, but biopsy is rarely performed during diagnosis. | see Figure 5b |




(**a**) (**b**)

**Figure 5.** The triplets extracted by the proposed model in the case study: (**a**) case_1 and (**b**) case_2.

The first sentence of the selected Chinese sentences is shown in Table 6, which indicates all triplets of the entity 'WD'. This sentence is classified as the SEO class, as the triplets of the sentence contain repeated entities. It can be be seen from Figure 5a that most of triplets were extracted accurately. More specifically, the proposed model correctly identified one disease and four drugs, and only missed on one drug. Moreover, it can be seen that the relationship between the entity 'trientine' and the entity 'WD' was not extracted because of the specific position of the entity 'trientine' in the sentence. In summary, the proposed model has more accurate extraction effects on the SEO class. On the second sentence shown in Table 6, classified as being from the EPO class, the results are shown in Figure 5b. This sentence contains more triplets, and has more than ten entities. Specifically, the proposed model correctly identified six diseases, three symptoms, and one examination, although it had one symptom is wrong. Furthermore, the wrong symptom ("fatty infiltration") from the proposed model was extracted because the relationship between the symptom "fatty infiltration" and the disease "acute fatty liver during pregnancy" was misidentified. In the end, the results of the case study fully prove the excellent performance of the proposed model on Chinese medical sentences.

### *4.4. Engineering Applications*

Intelligent medical treatment is a prominent future development trend in internet-based medical treatment. With the rapid development of information technology, more and more intelligent medical systems have been constructed to assist hospitals in diagnosing patients and even predicting diseases in advance. A health monitoring system is regarded as a kind of intelligent medical application scenario in the medical field, and is an important development for realizing medical data sharing and fusion. Due to the diversity and complexity of medical texts, existing research on the structuring of electronic medical records has mostly paid attention to exploiting deep learning models for completion intelligent medical tasks, such as medical entity recognition, medical relationship extraction, medical entity linking, medical entity alignment, etc. The structured electronic medical records are then applied for health monitoring and disease prediction. Medical entity recognition and medical relationship extraction are essential parts of the medical text structuring task. However, extracting the wrong triplets can have an enormous negative impact on the accuracy and universality of subsequent applications. In this regard, a more accurate deep learning model is proposed in this paper to complete the task of entity relation extraction. As shown in Figure 6, a data-driven approach is employed to structure electronic medical record data. More specifically, electronic medical records from hospitals and health indicators from devices are exploited to predict patients' condition.

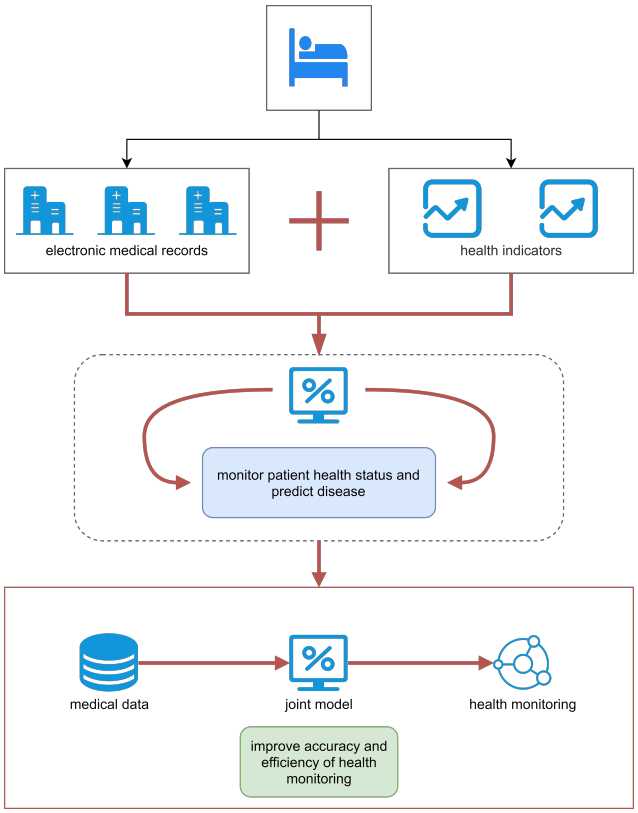



**Figure 6.** An application diagram of the proposed model in health monitoring.

As mentioned above, in order to further improve the accuracy of health monitoring, our next work will focus on training more types of data, including technical medical terms, unstructured crawler data, etc. In addition, as the deep learning model has visible shortcomings, such as the lack of robustness, the lack of annotated data, the few-shot learning method could be incorporated into the proposed model to reduce the dependence of the model on labeled data. In addition, this method would improve the efficiency of extraction process while saving labor cost. The proposed model could have great significance for the construction of intelligent medical system.

## 5. Conclusions and Future Work

This article highlights the advantages of artificial intelligence technology for health monitoring. Specifically, we propose a novel model for joint extraction of entities and relationships to improve the accuracy of health monitoring. The proposed model transforms joint extraction into a binary tagging problem. We introduce RoBERTa to fully extract sentence features. Furthermore, we exploit conditional layer normalization in the decoder to combine entities with relationships. Talking-head attention is applied to strengthen the interaction between entity recognition and relation extraction. Thus, the proposed model can simultaneously extract different triplets from sentences and alleviate the problem of overlapping triplets. We conducted complex experiments on two Chinese datasets to demonstrate the effectiveness of the proposed model. Our experimental results on the Baidu2019 and CHIP2020 datasets show that the proposed model outperforms baseline models. Ablation experiments were used to demonstrate the importance of each module. In summary, the experiments show that the proposed model can effectively extract overlapping triplets and has better performance than existing methods.

In the future, different technologies can be further explored to extract information efficiently in Chinese sentences. First, the accuracy of the proposed model on the SEO class needs to be improved in order to increase the accuracy of the healthcare monitoring system. In order to solve the problem of low efficiency on SEO class identification, one option is to investigate different attention fusion methods. A second issue is that this model does not effectively identify medical texts, as special medical sentences are complex. Hence, for special medical sentences, medical dictionaries could be combined with medical sentences to enhance the model's performance on extracting overlapping triplets in medical contexts.

**Author Contributions:** Conceptualization, B.S. and L.Z. (Lijuan Zhang); methodology, B.S.; software, B.S. and J.W.; validation, B.S. and L.Z. (Lijuan Zhang); formal analysis, B.S. and N.X.; investigation, B.S. and R.F.; resources, L.Z. (Lei Zhang) and J.H.; data curation, B.S.; writing—original draft preparation, B.S. and R.F.; writing—review and editing, L.Z. (Lijuan Zhang); visualization, B.S. and A.V.; supervision, L.Z. (Lijuan Zhang) and J.H.; project administration, L.Z. (Lei Zhang) and J.H.; funding acquisition, L.Z. (Lei Zhang). All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partially funded by the Zhejiang Province Key Research and Development Project (2020C03071, 2021C03145).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## References

1. Elbattah, M.; Arnaud, É; Gignon, M.; Dequen, G. The role of text analytics in healthcare: A review of recent developments and applications. *Healthinf* **2021**, *5*, 825–832.
2. Bose, P.; Srinivasan, S.; Sleeman, W.C., IV; Palta, J.; Kapoor, R.; Ghosh, P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Appl. Sci.* **2021**, *11*, 8319. [CrossRef]
3. Arnaud, É.; Elbattah, M.; Gignon, M.; Dequen, G. Deep learning to predict hospitalization at triage: Integration of structured data and unstructured text. In Proceedings of the IEEE International Conference on Big Data, Atlanta, GA, USA, 10–13 December 2020; pp. 4836–4841.
4. Lai, T.; Cheng, L.; Wang, D.; Ye, H.; Zhang, W. Rman: Relational multi-head attention neural network for joint extraction of entities and relations. *Appl. Intell.* **2022**, *52*, 3132–3142. [CrossRef]
5. Huang, W.; Zhang, J.; Ji, D. Correction: A transition-based neural framework for chinese information extraction. *PLoS ONE* **2021**, *16*, e0250519. [CrossRef] [PubMed]
6. Liu, X.; Liu, Y.; Wu, H.; Guan, Q. A tag based joint extraction model for chinese medical text. *Comput. Biol. Chem.* **2021**, *93*, 107508. [CrossRef] [PubMed]

7. Takanobu, R.; Zhang, T.; Liu, J.; Huang, M. A hierarchical framework for relation extraction with reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 7072–7079.
8. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z. Pre-training with whole word masking for chinese bert. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [CrossRef]
9. Khalifa, M.; Shaalan, K. Character convolutions for arabic named entity recognition with long short-term memory networks. *Comput. Speech Lang.* **2019**, *58*, 335–346. [CrossRef]
10. Khalifa, M.; Shaalan, K. Improving the performance of dictionary-based approaches in protein name recognition. *J. Biomed. Inform.* **2004**, *37*, 461–470.
11. Jiang, M.; Chen, Y.; Liu, M.; Rosenbloom, S.T.; Mani, S.; Denny, J.C.; Xu, H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 601–606. [CrossRef]
12. Lei, J.; Tang, B.; Lu, X.; Gao, K.; Jiang, M.; Xu, H. A comprehensive study of named entity recognition in chinese clinical text. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 808–814. [CrossRef]
13. Ponomareva, N.; Rosso, P.; Pla, F.; Molina, A. Conditional random fields vs. hidden markov models in a biomedical named entity recognition task. In Proceedings of the Recent Advances in Natural Language Processing (RANLP), Valencia, Spain, 27–29 September 2007; pp. 479–483.
14. Sherstinsky, A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
15. Devlin, J.; Chang, M.W.; Lee, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
16. Liu, Z.; Yang, M.; Wang, X.; Chen, Q.; Tang, B.; Wang, Z.; Xu, H. Entity recognition from clinical texts via recurrent neural network. *BMC Med. Inform. Decis. Mak.* **2017**, *17*, 67. [CrossRef] [PubMed]
17. Gridach, M. Character-level neural network for biomedical named entity recognition. *J. Biomed. Inform.* **2017**, *70*, 85–91. [CrossRef]
18. Zhang, Y.; Yang, J. Chinese ner using lattice lstm. *Assoc. Comput. Linguist.* **2018**, *1*, 1554–1564.
19. Zhao, S.; Cai, Z.; Chen, H.; Wang, Y.; Liu, F.; Liu, A. Adversarial training based lattice lstm for chinese clinical named entity recognition. *J. Biomed. Inform.* **2019**, *99*, 103290. [CrossRef]
20. Li, X.; Zhang, H.; Zhou, X.H. Chinese clinical named entity recognition with variant neural structures based on bert methods. *J. Biomed. Inform.* **2020**, *107*, 103422. [CrossRef]
21. Gao, W.; Zheng, X.; Zhao, S. Named entity recognition method of chinese emr based on bert-bilstm-crf. *J. Physics Conf. Ser. (JPCS)* **2021**, *1848*, 012083. [CrossRef]
22. Kong, J.; Zhang, L.; Jiang, M.; Liu, T. Incorporating multi-level cnn and attention mechanism for chinese clinical named entity recognition. *J. Biomed. Inform.* **2021**, *116*, 103737. [CrossRef]
23. Wang, J.; Xu, W.; Fu, X.; Xu, G.; Wu, Y. Astral: Adversarial trained lstm-cnn for named entity recognition. *Knowl.-Based Syst.* **2020**, *197*, 105842. [CrossRef]
24. Li, F.; Zhang, M.; Fu, G.; Ji, D. A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinform.* **2017**, *18*, 198. [CrossRef]
25. Zhang, Y.; Lin, H.; Yang, Z.; Wang, J.; Zhang, S.; Sun, Y.; Yang, L. A hybrid model based on neural networks for biomedical relation extraction. *J. Biomed. Inform.* **2018**, *81*, 83–92. [CrossRef]
26. Tian, Y.; Chen, G.; Song, Y.; Wan, X. Dependency-driven relation extraction with attentive graph convolutional networks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Virtual Meeting, 1–6 August 2021; pp. 4458–4471.
27. Wang, D.; Tiwari, P.; Garg, S.; Zhu, H.; Bruza, P. Structural block driven enhanced convolutional neural representation for relation extraction. *Appl. Soft Comput.* **2020**, *86*, 105913. [CrossRef]
28. Zhang, X.; Zhang, Y.; Zhang, Q.; Ren, Y.; Qiu, T.; Ma, J.; Sun, Q. Extracting comprehensive clinical information for breast cancer using deep learning methods. *Int. J. Med. Inform.* **2019**, *132*, 103985. [CrossRef]
29. Xu, S.; Sun, S.; Zhang, Z.; Xu, F.; Liu, J. Bert gated multi-window attention network for relation extraction. *Neurocomputing* **2022**, *492*, 516–529. [CrossRef]
30. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef] [PubMed]
31. Miwa, M.; Bansal, M. End-to-end relation extraction using lstms on sequences and tree structures. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Chicago, IL, USA, 7–12 August 2016.
32. Katiyar, A.; Cardie, C. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 917–928.
33. Gu, J.; Lu, Z.; Li, H.; Li, V.O. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv* **2016**, arXiv:1603.06393.
34. Zeng, X.; Zeng, D.; He, S.; Liu, K.; Zhao, J. Extracting relational facts by an end-to-end neural model with copy mechanism. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 506–514.

35. Bekoulis, G.; Deleu, J.; Demeester, T.; Develder, C. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Syst. Appl.* **2018**, *114*, 34–45. [CrossRef]
36. Bekoulis, G.; Deleu, J.; Demeester, T.; Develder, C. Adversarial training for multi-context joint entity and relation extraction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 1 January 2018; pp. 2830–2836.
37. Huang, W.; Cheng, X.; Wang, T.; Chu, W. Bert-based multi-head selection for joint entity-relation extraction. In Proceedings of the Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, 9–14 October 2019.
38. Yu, B.; Zhang, Z.; Shu, X.; Wang, Y.; Liu, T.; Wang, B.; Li, S. Joint extraction of entities and relations based on a novel decomposition strategy. *arXiv* **2019**, arXiv:1909.04273.
39. Dixit, K.; Al-Onaizan, Y. Span-level model for relation extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5308–5314.
40. Eberts, M.; Ulges, A. Span-based joint entity and relation extraction with transformer pre-training. *arXiv* **2019**, arXiv:1909.07755.
41. Luo, L.; Yang, Z.; Cao, M.; Wang, L.; Zhang, Y.; Lin, H. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *J. Biomed. Inform.* **2020**, *103*, 103384. [CrossRef]
42. Hong, Y.; Liu, Y.; Yang, S.; Zhang, K.; Wen, A.; Hu, J. Improving graph convolutional networks based on relation-aware attention for end-to-end relation extraction. *IEEE Access* **2020**, *8*, 51315–51323. [CrossRef]
43. Lai, Q.; Zhou, Z.; Liu, S. Joint entity-relation extraction via improved graph attention networks. *Symmetry* **2020**, *12*, 1746. [CrossRef]
44. Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; Chang, Y. A novel cascade binary tagging framework for relational triple extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 5–10 July 2020; pp. 1476–1488.
45. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv* **2019**, arXiv:1910.03771.
46. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
47. Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; Liu, R. Plug and play language models: A simple approach to controlled text generation. *arXiv* **2019**, arXiv:1912.02164.
48. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
49. Shazeer, N.; Lan, Z.; Cheng, Y.; Ding, N.; Hou, L. Talking-heads attention. *arXiv* **2020**, arXiv:2003.02436.
50. Bock, S.; Weiß, M. A proof of local convergence for the adam optimizer. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
51. In Proceedings of the 14th China Conference on Knowledge Graph and Semantic Computing, Hangzhou, China, 24–27 August 2019. Available online: https://sigkg.cn/ccks2019/ (accessed on 5 May 2023).
52. In Proceedings of the 6th China Health Information Processing Conference, Online, 28–29 November 2020. Available online: http://cips-chip.org.cn/2020/ (accessed on 5 May 2023).
53. Abadi, M.; Agarwal, A.; Barham, P. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: https://www.tensorflow.org (accessed on 5 May 2023).
54. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brooklyn, NY, USA, 16–20 November 2020; pp. 38–45.
55. Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; Weld, D.S. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 541–550.
56. Ren, X.; Wu, Z.; He, W.; Qu, M.; Voss, C.R.; Ji, H.; Abdelzaher, T.F.; Han, J. Cotype: Joint extraction of typed entities and relations with knowledge bases. In Proceedings of the 26th International Conference on World Wide Web, Geneva, Switzerland, 3 April 2017; pp. 1015–1024.