Article

Valuable Knowledge Mining: Deep Analysis of Heart Disease and Psychological Causes Based on Large-Scale Medical Data

Ling Wang [1], Minglei Shan [1], Tie Hua Zhou [1,*] and Keun Ho Ryu [2,3]

1   Department of Computer Science and Technology, School of Computer Science, Northeast Electric Power University, Jilin 132013, China; smile2867ling@neepu.edu.cn (L.W.); 2202000666@neepu.edu.cn (M.S.)
2   Data Science Laboratory, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam; khryu@tdtu.edu.vn
3   Biomedical Engineering Institute, Chiang Mai University, Chiang Mai 50200, Thailand
*   Correspondence: thzhou@neepu.edu.cn; Tel.: +86-186-8631-3355

Abstract: The task of accurately identifying medical entities and extracting entity relationships from large-scale medical text data has become a hot topic in recent years, aiming to mine potential rules and knowledge. How to conduct in-depth context analysis from biomedical texts, such as medical procedures, diseases, therapeutic drugs, and disease characteristics, and identify valuable knowledge in the medical field is our main research content. Through the process of knowledge mining, a deeper understanding of the complex relationships between various factors in diseases can be gained, which holds significant guiding implications for clinical research. An approach based on context semantic analysis is proposed to realize medical entity recognition and entity relationship extraction. In addition, we build a medical knowledge base related to coronary heart disease and combine the NCBI disease dataset and the medical lexicon dataset extracted from the text as the test data of the experiment. Experimental results show that this model can effectively identify entities in medical text data; the WBC model achieved an F1 score of 89.2% in the experiment, while the CSR model achieved an F1 score of 83.4%, and the result is better than other methods.

Keywords: knowledge extraction; contextual analysis; semantic analysis; data mining; NLP

1. Introduction

The medical named entity mining method is a hot issue in the fields of literature information acquisition and medical knowledge discovery acquisition [1]. Medical named entities have a variety of types, especially for disease-related named entities having a key role in medical research [2], including disease etiology analysis, disease association exploration, clinical diagnosis and treatment opinion discovery, disease prediction, etc. [3]. The main research task of medical text processing relies on medical domain NER (named entity recognition) and ERE (entity and relation extraction) [4–8] for medical text mining.

There are many complications in the current medical NER and ERE that affect the effectiveness of this operation. For example, some names of medical entities are composed of a combination of Greek letters and various affixes, and these names are difficult to identify using only word or phrase forms. Also, many disease names and drug names are composed of descriptive words for body parts, which can lead to confusion with other types of entities when identifying disease names or drug entities. In addition, a disease may be described in different ways, with different names and name abbreviations, a situation that exists in abundance in medical text.

To be able to extract semantics effectively from a large amount of medical text, Mikol et al. provided a negative sampling word embedding method [9], which was able to achieve better word-embedding effects and simplify the traditional structure to speed up the processing of a large amount of texts. As a semantic feature extraction process method,

the LSTM model can combine multiple methods to optimize parameters and improve the effectiveness of NER tasks.

Considering the complex contextual relationships in the ERE task, pre-trained models can provide advanced processing methods and excellent performance in such tasks. In Émilien Arnaud's study [10], BERT was employed to learn contextual semantic relationships from a large corpus of texts. The trained model was then compared and analyzed with CamemBERT, mBERT, and FlauBERT, yielding outstanding results.

Therefore, we mainly study coronary heart disease, aiming to mine the related and valuable medical knowledge based on context semantic analysis, and our proposed model has a higher F1 score and accuracy rate than other methods.

The rest of the paper is organized as follows: Section 1 is the introduction to the background of current medical knowledge discovery research; Section 2 is the introduction to the current state of research; Section 3 presents the motivation for the research; Section 4 is the detailed description of the calculation process for the WBC and CSR models; Section 5 describes the experimental analysis process; Section 6 is the analysis of mined multivariate relationships; Section 7 is the research summary.

2. Related Work

Medical entity mining refers to objects such as names, terms, attributes, etc. There are three main types of techniques for large-scale entity extraction, such as feature word vector calculation, rule-based contextual approaches, and complex-model approaches. Most of the rule and lexicon-based matching approaches rely on the lexicon of the entity's corresponding domain [11–14]. By relying on the complete lexicon or semantic rules of the relevant domain, this approach can achieve better results. The workload is immense, as all the rules need to be fully given in the lexicon construction process, as do the entity concepts of the related domains. And there are large limitations in multi-domain applications and extensions. In recent years, entity recognition approaches incorporating machine learning have been widely used for such textual tasks.

The complex-model approach to solve the NER problem is to dynamically label the keywords in the text, such as the SVM [15] or sequential labeling models [16,17], which can combine the labeling information to calculate the contextual relevance and hidden relationships existing among the labels. These approaches have good scalability compared to lexicon-based approaches. Moreover, deep learning models are mainly used for solving amount of texts feature exaction and label problems. Collobert et al. [18] proposed a CNN/FNN-CRF model for learning semantic features contained in text from the unlabeled text. This approach has better performance and time efficiency than other models. Other deep learning models also have good performance, such as CNN, attention mechanism, etc. [19,20].

Previous models were able to obtain better results in the case of general-purpose domains, but medical texts contain a large amount of medical specialized terms and medical abbreviations in the text of professional domains (e.g., medical text processing, the text of electronic medical records, etc.). There is a lack of sufficient amounts of annotated datasets and knowledge bases in these domains, and the NER models of general-purpose domains do not work well when applied to professional domains. In recent years, named entity recognition for the medical domain has become a hot topic [21].

The effectiveness of complex-model approaches depends on data representation of the text, and different features can be obtained from different representations [22]. Mikolov et al. [23] study jump graphs that can compute word vectors of words in large datasets. Since the disease-named entities contain cold words, the model removes the implicit layer for efficiency, and the words in the input layer share a mapping layer. This approach is an improvement of the noise-contrast estimation [24], which speeds up the calculation and maintains word vector quality, simplifying the model.

Relation extraction refers to the determination of the relationship between the two entities based on semantic relation after the entity in the text is obtained through

a NER task. Since relationships are usually pre-defined, relationship extraction can be considered a multi-classification task.

The initial ERE tasks mainly rely on template matching, which manually summarizes the rules of the entity relationship in the text and then matches the rules in the new text. The generalization ability of this method is weak in the face of different syntax structures and writing methods in different corpora; accurate identification will be difficult, and it is only suitable for small-scale data. The latter method integrates machine learning, maps the entity feature vector to high-dimensional space, and carries out relation extraction for feature classification. Commonly used models include support vector machine and conditional random field.

Currently, common deep learning-based methods use deep learning to automatically construct features and avoid the limitations of manual feature selection. Common models include RNN and CNN. Liu et al. [25] used the CNN model in an ERE task to input the word vector in the text into the model and construct the feature relation for relation extraction. Zhang et al. [26] used the bidirectional RNN model to solve the problem of CNN not being able to learn the temporal features and to obtain the long-distance feature information in the text. Dai et al. [27] used a rule-based template and Bert-BiLSTM-CRF to identify hazardous material accident report documents with structured text in Chinese, carried out a named entity recognition task, which solved the problem of the colloquial description of professional words in the report, and verified the effectiveness of the method in the dataset. Panoutsopoulos et al. [28] built a model on Python's spaCy library and trained it on a manually annotated text corpus, which improved the problem of domain ambiguity and consistency among annotators. Sun et al. [29] introduced a new neural network based on morphology and syntactic grammar into a model to solve complex language structure problems encountered in entity recognition tasks in different languages, and verified multiple grammar rules in four Nordic languages, proving the improved effect. LFC Cunha et al. [30] created an intelligent document browsing tool through entity link and record link technology, trained the algorithm model through several annotated corpora, and compared with other architectures to prove the effect of the model. Sboev A et al. [31] used the complex named entity recognition method to label pharmaceutical-related entities and proposed a corpus of drug reviews in the network. The accuracy of the corpus was evaluated through the deep learning model, and the experimental effect was higher than average.

The NER task and ERE task are executed sequentially, and some researchers reuse the parameters in the two processes to achieve a coherent task. Wei et al. [32] proposed two types of task sequencing models, first identifying entities, then identifying another entity that may form the entity relationship according to the relationship, and finally outputting the triplet relationship. Miwa et al. [33] used LSTM and RNNs to extract the syntactic and sequence information in the text, so as to carry out the two tasks together.

3. Motivation

The number of medical documents available on the web is increasing, and it has become a hot issue to perform text processing tasks from a large number of medical documents, to obtain medical named entities and the relationship of entities through the tasks, and to use them to mine the causes of disease pathogenesis, clinical diagnosis, and to perform disease prediction.

The medical knowledge extraction task is one of the medical text processing tasks, as shown in Figure 1. In the diseases spanning different departments of medicine, two or more kinds of medical knowledge are needed for diagnosis and treatment research, and there will be some difficulties in the process of implementation. In different diseases, there are different names for the same disease or symptom, and the same abbreviation may have different meanings. In order to solve this problem in the process of entity recognition, we put forward the model of conditional random field and BiLSTM combined word embedding, which can solve the above problems and ensure accuracy and effectiveness.

In order to solve the problem of the accuracy of relation extraction being low due to the long distance between entities in the process of entity relation extraction, we put forward a model that uses context location relation and attention mechanism to ensure the effect of entity relation extraction. After entity relationship extraction, medical knowledge in related medical fields can be obtained, and high-quality medical knowledge extraction can be completed.
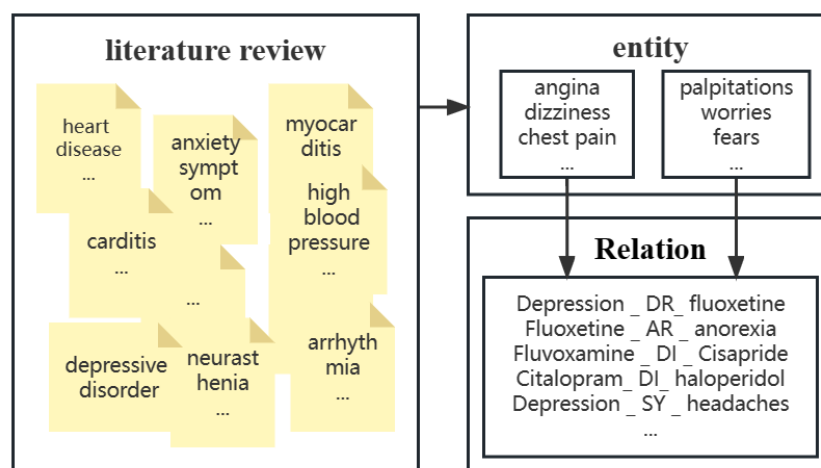

Figure 1. Knowledge extraction scheme.

4. Materials and Methods

4.1. Dataset

The testing dataset mainly uses the NCBI disease corpus dataset [34], which contains 793 abstracts, 6892 sentences, and 790 disease concepts manually annotated by medical professionals. Additionally, we constructed a medical knowledge base consisting of 3167 phrases mined from 2037 works of medical literature related to coronary heart disease, with the literature source being PubMed. Moreover, combined with Wikipedia's medical category, our constructed medical dictionary contained 335,769 valuable phrase words. Another part of the dataset used for entity relationship verification used the i2b2 relation to extract text, and the original dataset contained 871 annotated texts.

The textual data of the experimental data on emotional heart disease came from authoritative medical literature websites such as NCBI and Lancet. The literature was retrieved using keywords such as 'emotional heart disease', 'psychosomatic heart disease', 'cardiomyopathy of emotional origin', 'stress-induced heart condition', and 'neurogenic heart disease'. Clinical research medical literature related to emotional heart disease was obtained. After downloading, the data underwent preprocessing, converting the text format from PDF to TXT. Apart from the core contents such as the main text and abstract, the converted text also includes some information irrelevant to the calculation process, such as source information, author information, journal information, and references. This information is not involved in the text mining calculation process, so it is removed. The total number of texts is 2037, containing the core contents such as the main text and abstract. The entity relationships in the emotional heart disease-related textual dataset are automatically labeled based on the SemMedDB knowledge base, thereby obtaining accurate and authoritative annotated text.

4.2. WBC Model

There are three layers of our proposed WBC model, as shown in Figure 2: word embedding, BiLSTM, and conditional random field:
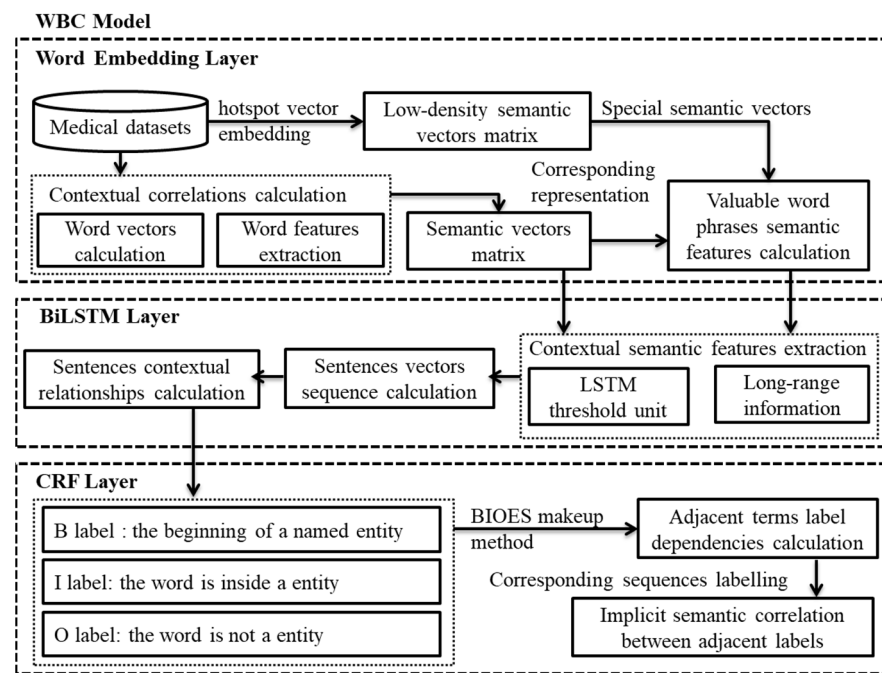
Figure 2. WBC flowchart.

4.2.1. Word-Embedding Layer

In the word-embedding layer, we avoid a single hotspot vector that leads to a sparse matrix, which in turn does not represent the relevance between text contexts well. Our model further embeds the hotspot vector into a low-density semantic vector. After getting the input, we calculate the word vectors and then label the word features according to contextual correlations and corresponding representation to calculate the semantic vector, which makes the word vectors have different representations in different named entities according to its special semantic vectors.

The embedding process is as follows: in order to calculate accurate semantic features for valuable word phrases, we construct look-up tables based on the knowledge base by using a large-scale amount of biomedical literature and a professional dictionary. In this study, we also considered the unlabeled Wikipedia dataset for training through the word2vec [24], with 200 dimensions of each character vector, which can support a more diverse semantic representation.

4.2.2. BiLSTM Layer

The second layer is BiLSTM, which learns contextual semantic relations in the texts through the word vector matrix. The long-range information is captured by the threshold unit of the LSTM, and a larger range of contextual semantic features can be obtained through the long-range information. This layer of the model can learn the contextual relationships of the sentences before and after the target sentence. The model takes a literary text as an input sequence and outputs a vector $h = (h_1, h_2, h_3, \ldots, h_n)$ that is a sequence of each step of the input sequence. LSTM solves the long-term dependency problem by combining storage units, controlling the ratio of input, and forgetting the current storage unit, and forgetting gates are used to forget the previous control; the state of the unit is forgotten.

There are *n* vectors, and each vector *et* represents the vector where the character vector v_{c_t} and the corresponding feature v_{d_t} are combined in the vector matrix. The parameter h_t is represented as the current hidden state by calculating x_t (input vector) and h_{t-1} (last state vector) for each position *t*, as shown in the following formula [35].

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{2}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{4}$$

$$h_t = o_t \odot \tanh(c_t) \tag{5}$$

However, one-way LSTM will have problems. In the process of state calculation, subsequent state information will be ignored; we use BiLSTM to solve the one-way LSTM limitation, as it will calculate a forward representation $\overrightarrow{h_t}$, and a backward representation $\overleftarrow{h_t}$, from which the contextual state of each input sequence is obtained, and hidden state h_t can be calculated as follows [35]:

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \tag{6}$$

4.2.3. CRF Layer

The CRF layer is used to analyze label dependencies between adjacent terms. Here, we use the BIOES markup method for sequence labeling of text. B is the beginning of the entity; I is the middle part of the entity; O is the non-entity used to mark the part that has nothing to do with the entity; E is the end of the entity; S is the single-character entity, that is, the entity represented by a single word. The sequence labeling task can achieve the conversion from text sequence to tag sequence by labeling each position in the given input sequence with a corresponding label. In the process of sequence labeling, there is an implicit semantic correlation between adjacent labels. After labeling, it is necessary to determine whether the labels in adjacent positions are reasonable. For example, the I tag represents the middle part of the entity, so it must come after the B tag that represents the beginning of the entity. If the S tag and the B tag appear next to each other after a B tag, this situation is not reasonable.

The objective function is used for calculating the occurred probability for each tag, which shows that the larger the value it has, the greater the probability of selecting the tag it has at this location.

Sequence Y uses the state labels at different times to predict the most likely sequence of label T, where θ denotes the set of parameters in this layer, all of which can be obtained by the following log–likelihood ratio. T is the sequence of sequence labels, and p is the conditional probability of T given the set of sentences and parameters.

$$L(\theta) = \sum\nolimits_{(S,\ T)} \log_p(T \mid Y,\ \theta) \tag{7}$$

$S_\theta(B, T)$ denotes the score of a given input sentence to obtain a labeled sequence T, obtained from the transfer matrix A based on the output B of the previous layer. The normalization result of the score S could be considered as conditional probability p. $B_{s_t,t}$ is the probability that the current position character gets tag s_t, and the other parameter is the probability that the contents of the previous location adjacent to the character has label s_{t-1}.

$$b_t = \tanh h_t \tag{8}$$

$$S_\theta(B, T) = \sum (B_{s_t,t} + A_{s_{t-1},s_t}) \tag{9}$$

In the end, since abbreviations in medicine can be identical to abbreviations in other fields, this situation can lead to errors in the recognition of named entities. To address the occurrence, we use Ab3p [25] to identify abbreviations in medical named entities; Ab3P is

a biomedical text abbreviation recognition tool that has excellent results in abbreviation recognition, with an F1 score of 0.9 for its recognition effect on the Medstract corpus.

4.3. CSR Model

The method proposed in this paper takes into account the contextual semantic relation between entities, including the contextual relative position relation between entities and the semantic supplement of the entity relation. The model is divided into input layer, embedding layer, attention computing layer, feature coding layer, and output layer, which are described in Figure 3 as follows:
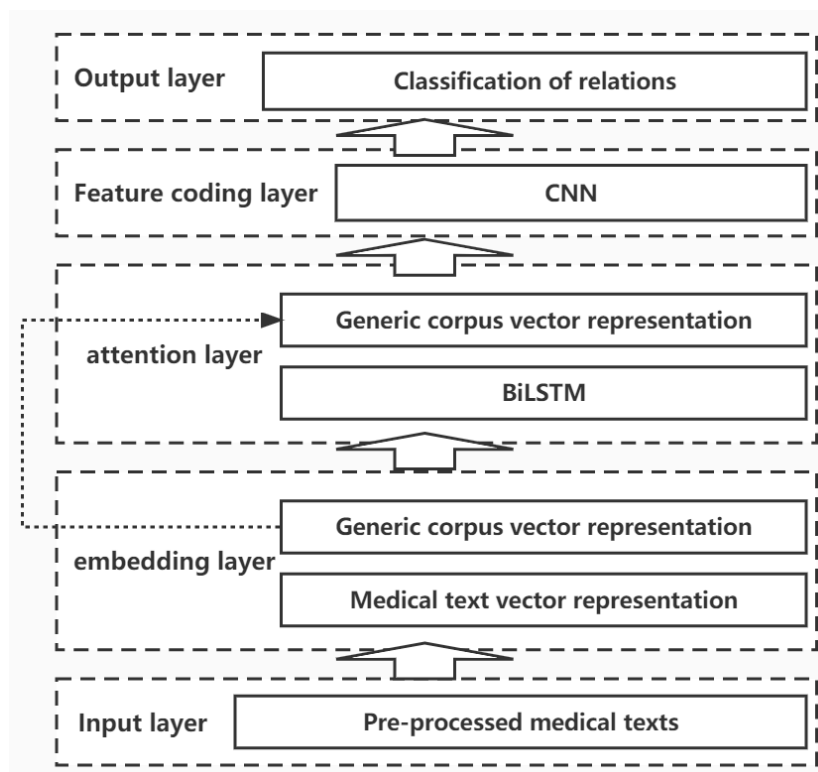

Figure 3. CSR flowchart.

The input layer is the pre-processed medical text. Multiple entities may appear in one sentence in a text, and multiple entities may participate in multiple relationships. In order to obtain the relationship between each entity pair without omission, the entity type will be used as the constraint to calculate the relationship between the entity pair. Since many disease-related entities are very long, that is, an entity is composed of multiple words, such a situation will interfere with the semantic computation of the entity. Therefore, the entity will here be replaced by the corresponding entity type in the medical text, and the entity type will also be added into the related computation of the entity.

The embedding layer will convert the corpus obtained from the input layer into a vector representation, which can be expressed in two ways. One is splicing the word vector obtained from the corpus training in the text with the corresponding position vector. The position vector formula here is as follows [36]:

$$POS_{(p,2r)} = \sin\left(\frac{p}{10000^{\frac{2r}{d}}}\right) \tag{10}$$

$$POS_{(p,2r+1)} = \cos\left(\frac{p}{10000^{\frac{2r}{d}}}\right) \tag{11}$$

Here, POS is the matrix of position vector; p is the absolute place of the word; d is the dimension of the place vector; r is the position of the position vector; the cosine formula is used to count the odd position; the sine formula is used to calculate the even position. Here, the embedding vector of the word in the original sentence sequence is represented as $X = \{x_1, x_2, x_3, \ldots, x_n\}$, where x_i represents the vector representation of the ith word of the sentence, and n is the length of the sentence. The other is the word vector obtained from the training on a general corpus, expressed as $Y = \{y_1, y_2, y_3, \ldots, y_n\}$, which is used for the extraction and calculation of the features of relative words as a semantic supplement.

The attention layer calculates the degree of semantic correlation between each word in a sentence and two entities, and calculates the common words related to the entity pairs. First of all, the word vector and position vector obtained from the medical text are spliced together to input the encoded sentence information in BiLSTM. $h_t^{(f)}$ and $h_t^{(b)}$ respectively represent the forward and backward outputs of BiLSTM at time t, and $h_t = [h_t^{(f)}, h_t^{(b)}]$ is the final output. The calculation formula of $h_t^{(f)}$ is as follows:

$$h_t^{(f)} = \text{Bilstm}(h_{t-1}, x_t) \tag{12}$$

And the same thing happens with $h_t^{(b)}$. After h is obtained, the degree of correlation between words and entities in the sentence is calculated through the attention mechanism. Here, cosine similarity is used to calculate the degree of correlation between words and entities. The calculation formula is as follows:

$$Re_t^{(1)} = \frac{h_t^T h_e^{(1)}}{|h_t| \times \left|h_e^{(1)}\right|} \times \frac{|h_t|}{\left|h_e^{(1)}\right|} \tag{13}$$

$$\omega_t = Re_t^{(1)} \times Re_t^{(2)} \tag{14}$$

In the formula, $h_e^{(1)}$ represents the hidden layer output of the first entity; $\left|h_e^{(1)}\right|$ represents the value of its modulus; $Re_t^{(1)}$ represents the degree of correlation between the term at the t position and the entity; and ω_t represents the final weight of the term.

Through the above calculation, we can obtain the weight of the word by combining the context position and semantic information, and the weight can represent the contribution of the word to the relationship between the two entities. Here, words with a large weight are generally common words, which can be represented by the word vector of the general domain corpus. Therefore, the vector of the general corpus is introduced into the calculation to expand the semantics. In the original sentence, the weight is expressed as $\omega = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_n\}$, and the eigen values are calculated using the following calculation methods:

$$I = Y \odot \omega \tag{15}$$

Formula $\odot$ expresses the word vector and the corresponding weights; Y is the calculated general domain corpus word vector; I is the input feature coding layer.

The feature coding step is used to calculate and extract the word features of the entity relationship. Here, the I obtained above is input to CNN, and then the output is processed with maximum pooling to obtain the features most relevant to the entity relationship. Here, the filter weight of the I-th channel of one of the Windows is set as W_i, and the size of the convolution window is m × k. The formula of the convolution layer is as follows:

$$Out_j^i = \varphi\left(\text{sum}\left(I^{j:j+m} \cdot W_i\right) + a_i\right) \tag{16}$$

$$v^i = \max_{1 \leq j < (n-m-1)} \left(Out_j^i\right) \tag{17}$$

$1 \leq j < n - m - 1$ represents the product of position elements corresponding to vectors; a is the convolution deviation; $\varphi(x)$ is the activation function; Out_j^i is the output of the convolution layer; and v^i is the feature vector obtained after maximum pooling. Here, a hidden layer is added between the feature layer and the output layer for regularization processing to speed up training.

The output layer takes the above regularized hidden layer α as a feature to represent the input fully connected layer for classification processing. The result obtained is the probability that the output o belongs to the r relation type, and the formula is as follows:

$$P(o = r) = \phi(W_\alpha \cdot \alpha + a) \tag{18}$$

The W_α in the formula is the weight matrix; a is the bias; and the activation function is $\phi(x)$, taking the maximum probability as the resulting category.

5. Results

We conducted experimental analyses on both the WBC model and the CSR model. In the experimental section of the WBC model, we first conducted comparative experiments with reference to the introduced comparison models. We then compared the use of dictionaries and the use of the Ab3p module in the WBC model. Finally, the parameters of the hidden layer and sub-embedding layer in the WBC model were determined. In the experimental section of the CSR model, we first analyzed the experimental effects of different category relationships in the I2b2 dataset and confirmed that the model can achieve excellent entity relationship mining results. We then conducted comparative analysis experiments with the aforementioned dataset, a 10-fold cross-validation with 80% training set and 20% testing set, introducing the models used for comparison in the experimental section. Lastly, the CSR model was used for relationship mining in the emotional heart disease data, and conclusions were drawn.

5.1. Experiment about WBC Model

In our study, the 60 dimension of the hidden layer after calculating testing is the best option for effectively optimizing the complexity of network layers; the results are better, and the computational complexity is lower for our testing datasets. Other requirement parameter settings are as follows in Table 1:

Table 1. Bi-LSTM parameter setting.

Parameter	Setting	Description
Word_ dimension	200	Token embedding dimension
Word_LSTM_dim	110	Token size in hidden layer
Word_bidirectional	TRUE	Using Bi-LSTM
Word Embedding	TRUE	Using word embedding
CRF	TRUE	Using CRF
Ab3P	TRUE	Using Ab3P

Comparative Experiment of the WBC Model: the text continues here (Figure 2 and Table 2). We tested our WBC model with four other NER models, CRF-Mesh [37], C-Bi-LSTM-CRF [37], CTAKES [38], DNER [39]; in terms of accuracy, recall, and F1-score [13], our proposed model had better results on all evaluation criteria, as shown in Figure 4.

Table 2. Top 10 examples of diverse relationships with high levels of support.

Pluralistic Relation	Support Degree
Stress cardiomyopathy–depression–palpitation–sleep disorders–TCAs	0.3165
Takotsubo cardiomyopathy–anxiety–hormonal changes–chest pain–lorazepam	0.3038
Takotsubo cardiomyopathy–depression–anxiety–tachycardia–aspirin	0.2970
Takotsubo cardiomyopathy–psychological stress–Dyspnea–Biological differences–diazine pyridine	0.2775
Stress cardiomyopathy–anxiety–insomnia–palpitation–metoprolol	0.2511
Stress cardiomyopathy–anxiety–elevated blood pressure–loss of appetite–SSRIs	0.2396
Takotsubo cardiomyopathy–heart failure–arrhythmia–vasoconstriction–ACE inhibitors	0.2006
Stress cardiomyopathy–anxiety–elevated blood sugar–tachycardia–clopidogrel	0.1869
Broken heart syndrome–hypertension–headache–atherosclerosis–nifedipine	0.1788
Broken heart syndrome–hyperlipidemia–arrhythmia–thrombosis–warfarin	0.1628

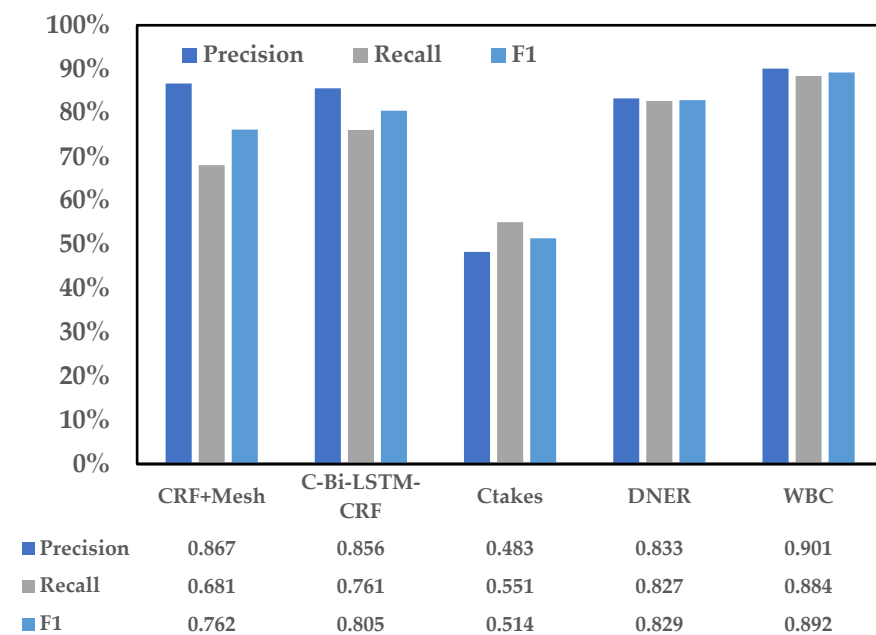|	CRF+Mesh	C-Bi-LSTM-CRF	Ctakes	DNER	WBC
Precision	0.867	0.856	0.483	0.833	0.901
Recall	0.681	0.761	0.551	0.827	0.884
F1	0.762	0.805	0.514	0.829	0.892

Figure 4. Precision, recall, and F1 score comparison.

Ab3p Module Effect Experiment: To verify the influence of the abbreviations of disease names or drug names in medical naming entities on recognition results, we did experiments to compare different cases of Ab3p parameters respectively, and the final results, as shown in Figure 5, indicate that the recognition of abbreviations in medical naming entities is necessary, and the same abbreviations may have different meanings, and medical abbreviations may have different meanings in other fields.
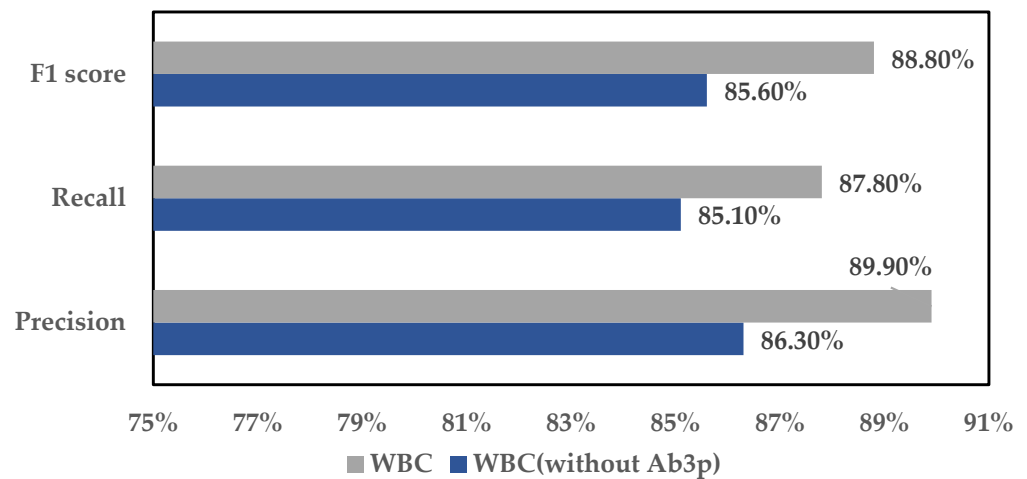
Figure 5. Ab3p comparison results.

Dictionary Impact on Experiment: The entity in the medicine named entities classified dictionary was recognized if there was a dictionary of participating entities' classification respectively under different conditions of contrast experiments, and the final result is shown in Figure 6; in a dictionary of entity categories labeled with cases, medicine named entity recognition will have a better effect.
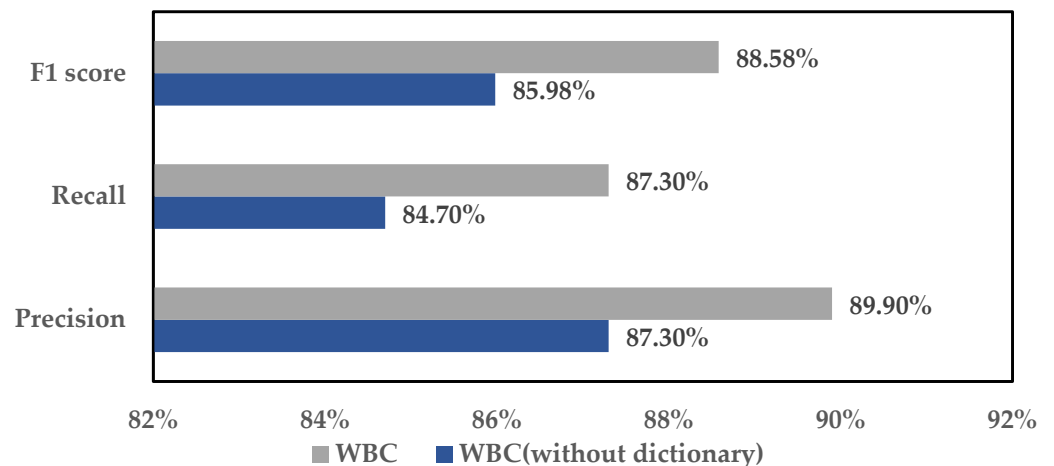

Figure 6. Dictionary comparison results.

Hidden Layer Parameter Calibration: In the BiLSTM layer network structure, the higher the hidden tier dimension, the higher the calculation of complexity. To optimize the enumerative complexity of this layer, we tested the different levels of computational efficiency caused by different hidden layer dimensions on the training dataset, ranging from 50 to 200, and tested the time complexity of different dimensions with a step size of 10. The final results of different dimensions of hidden layers are shown in Figure 7. When the dimension is 60, the score of F1 is 0.780, and when the dimension is 110, the score of F1 is 0.814. According to the different results, the time efficiency of calculation is weighed, and dimension 110 is finally selected, taking into account both the effect and efficiency of the calculation.
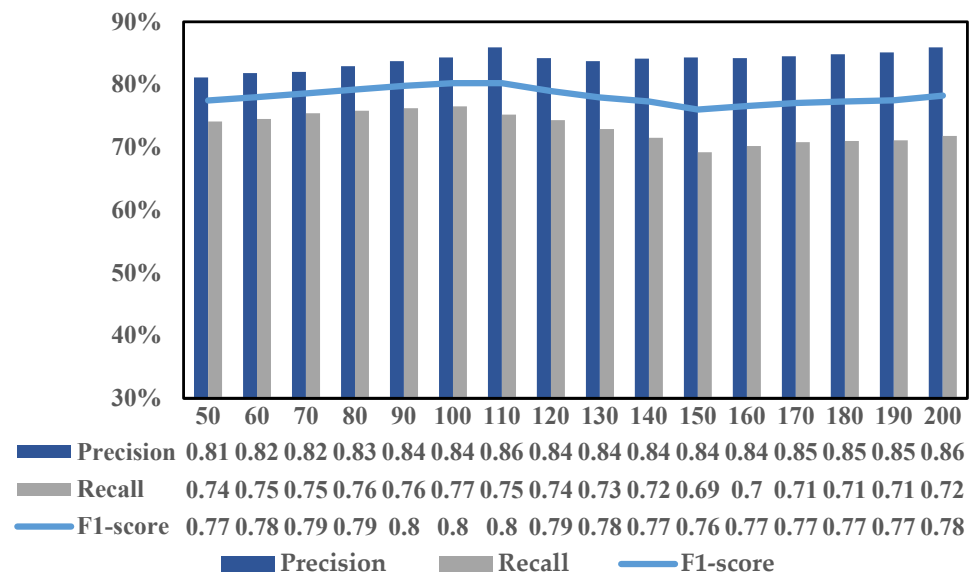
	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
Precision	0.81	0.82	0.82	0.83	0.84	0.84	0.86	0.84	0.84	0.84	0.84	0.84	0.85	0.85	0.85	0.86
Recall	0.74	0.75	0.75	0.76	0.76	0.77	0.75	0.74	0.73	0.72	0.69	0.7	0.71	0.71	0.71	0.72
F1-score	0.77	0.78	0.79	0.79	0.8	0.8	0.8	0.79	0.78	0.77	0.76	0.77	0.77	0.77	0.77	0.78

Figure 7. The different dimensions of the hidden layer comparison.

Calibration of Parameters in Word-Embedding Layer: In the word-embedding stage, different dimensions in the word-embedding process also lead to unused effects and time complexity. To find the optimal dimensionality for the trade-off between time efficiency and results, we experimented with different dimensions from 50 to 200 with a step size of 10 and finally obtained the following Figure 8 results. This dimension will be applied to different word-embedding tasks in the future.
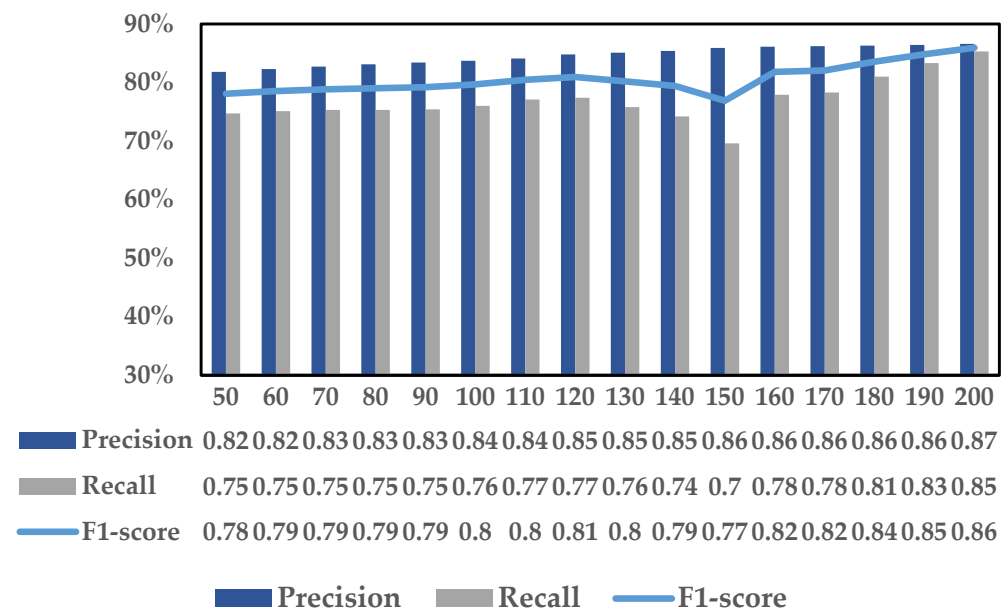

	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200
Precision	0.82	0.82	0.83	0.83	0.83	0.84	0.84	0.85	0.85	0.85	0.86	0.86	0.86	0.86	0.86	0.87
Recall	0.75	0.75	0.75	0.75	0.75	0.76	0.77	0.77	0.76	0.74	0.7	0.78	0.78	0.81	0.83	0.85
F1-score	0.78	0.79	0.79	0.79	0.79	0.8	0.8	0.81	0.8	0.79	0.77	0.82	0.82	0.84	0.85	0.86

Figure 8. The different dimensions of the word-embedding comparison.

5.2. Experiment with CSR Model

The entity relationship extraction performance of the model is verified by multiple comparison experiments with the i2b2 dataset. The I2b2 dataset contains three major relationship categories, namely, the concern between disease problems and treatment methods, the concern between disease problems and examination, and the relationship between disease problems and other disease problems. The small categories under each category are TrIP (treatment improves disease problem); TrWP (treatment worsens disease

problem); TrCP (treatment causes disease problem); TrAP (treatment manages disease problem); TrNAP (treatment is not managed due to disease problem); TeRP (examination reveals disease problem); TeCP (treatment investigates disease problem); PIP (disease problem) indicates a disease problem.

Accuracy of Relationship Categories: First, an ablation experiment is performed on the model. The baseline of the model is created by splicing the word vector trained by the corpus with the position vector as input and directly inputting the output of BiLSTM into CNN to extract features. The experimental results are as follows in Figure 9:
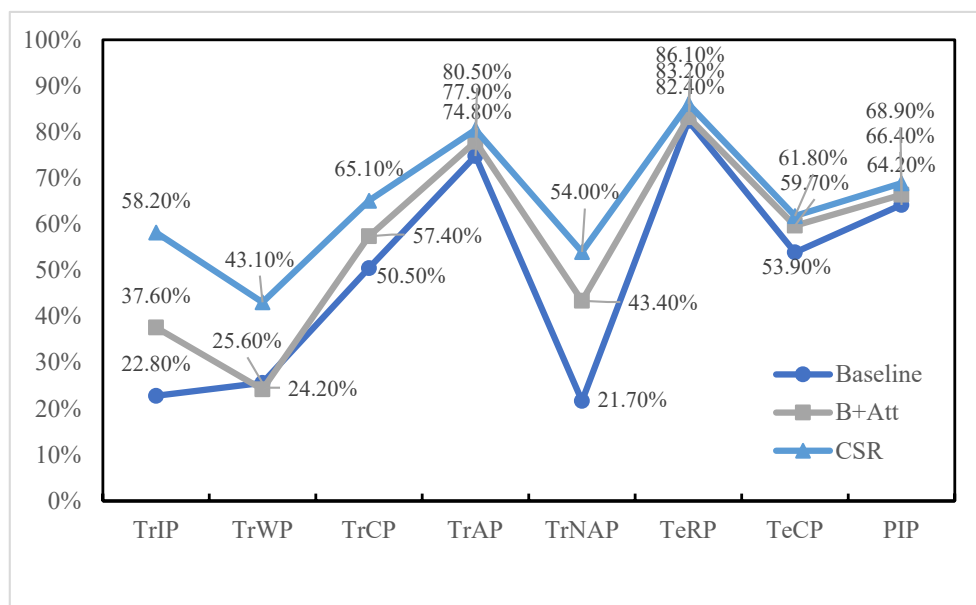
Figure 9. Ablation experiment.

It can be seen that, after the attention computation procedure is added to reduce the noise of the position vector in the vector, the F1 value of the model is improved. The reason for the improvement is that there is the noise of the position vector in the exportation result of the hidden tier of the BiLSTM model, which will influence the extraction result effect of semantically relevant features in the subsequent calculation. After the attention computation procedure is added, only the semantic recognition of the word vector will be achieved, which will not be affected by the position vector noise, so that the subsequent calculation effect will be improved. After the semantic expansion, the effect will be improved again. The general domain corpus introduced here can make the semantic recognition of common words more accurate. After the selection of the attention computation procedure, the common words related to the entity relationship will obtain a higher weight in the calculation, so as to facilitate the judgment of the entity relationship.

Comparative Experimental Study on the CSR Model: Here, several existing entity relationship extraction methods are selected for comparison, including CRNN [40], SVM [41], ConvLSTM-Att [37], CNN [41], and Seg-CNN [38]. The comparison results are shown in Figure 10:
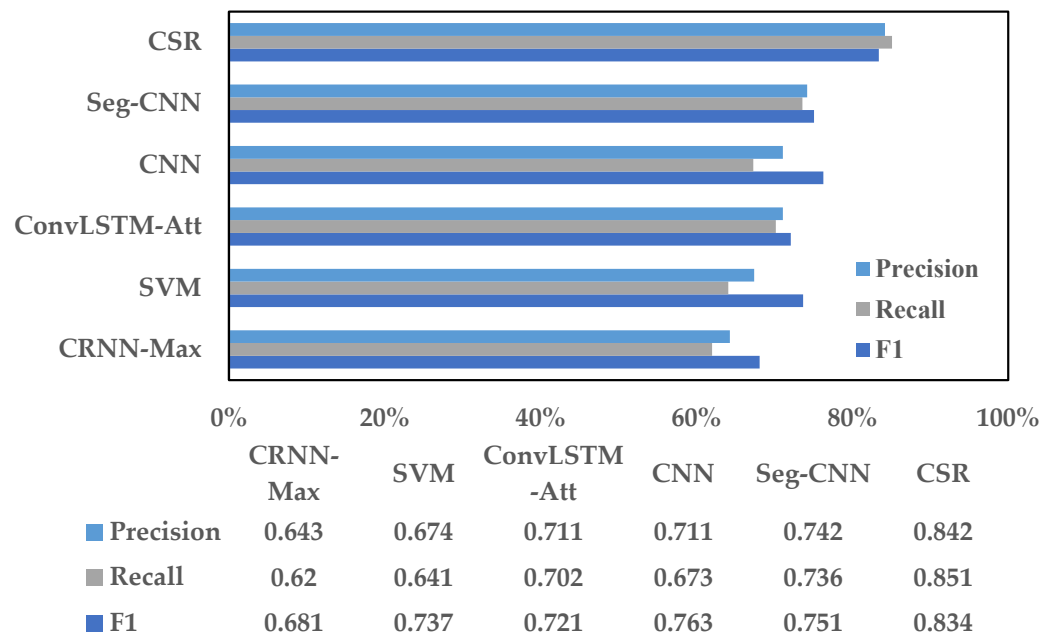
	CRNN-Max	SVM	ConvLSTM-Att	CNN	Seg-CNN	CSR
■ Precision	0.643	0.674	0.711	0.711	0.742	0.842
■ Recall	0.62	0.641	0.702	0.673	0.736	0.851
■ F1	0.681	0.737	0.721	0.763	0.751	0.834

Figure 10. Experiment of contrast.

It can be seen from the comparison test results that the method proposed in this paper adds the context position relation and semantic extension between entities in the calculation process of entity relation extraction, which makes the calculation effect achieve the optimal effect.

Accuracy of Emotion-Related Heart Disease Dataset Mining: We also tested the literature on coronary heart disease and mental illness obtained from medical websites. In this paper, the relationship between entities is divided into several types: drug–disease, drug–symptom, disease–disease, disease–symptom, and disease–site. In this paper, the entities extracted above are marked in the form of entity pairs in a uniform ratio in the text data, and the resulting entity relationship is evaluated after calculation. The accuracy rate obtained is shown in Figure 11.
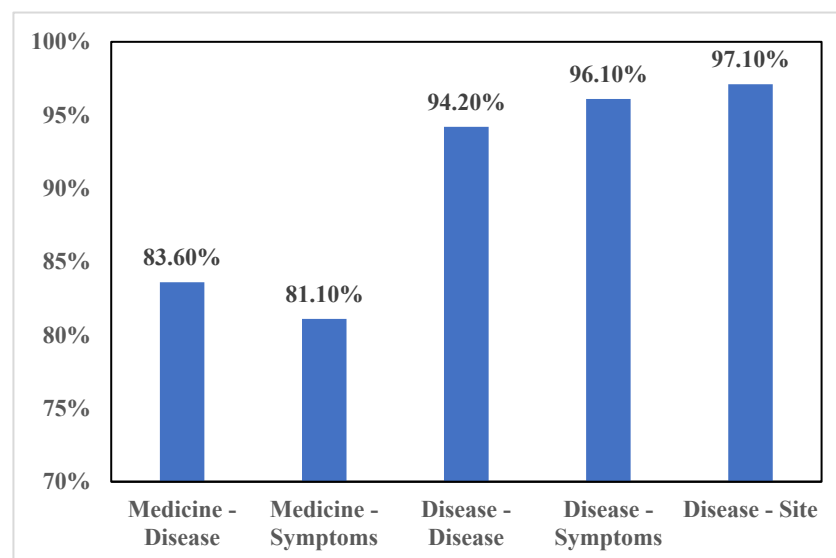

Figure 11. Precision of each type of entity relationship.

It can be seen from the results that the recognition effect of all relationship categories is excellent, among which the accuracy of disease–part is the highest. The reason is that the

entities of disease and part in the medical text are close to each other, and the regularity is strong, so that the model can learn the features well. The reason for the poor recognition of drugs and symptoms is that there is a large text span between the two entities of drugs and symptoms, and there will be many other different types of entities in the introduction of a drug, and the context is relatively special.

After the entity relationship is mined from all the medical texts at the intersection of coronary heart disease and mental disease, the ternary relationship between coronary heart disease and mental disease, such as symptoms, therapeutic drugs, and other factors, is obtained; for example, "depression-dr-fluoxetine" is a ternary relationship between depression with more occurrences and fluoxetine. More occurrences indicate that the use of fluoxetine in depression-related therapeutic drugs is relatively common in a large number of the works of literature collected so far. For example, "anxiety_CD_cardiovascular disease" is also representative of the relationship between diseases. It represents the concurrent relationship between anxiety disorder, hypertension, and cardiovascular disease in mental diseases. The frequency of occurrence in the final process of sorting out the relationship results is also very high, representing that these two diseases are typical concurrent diseases of anxiety disorder. In the relationship results obtained from the experimental results, there are a large number of emotional factors in the related disease representations of mental illness. For example, "depression_SY_anxiety" indicates that the disease representations of depression contain emotional factors of anxiety.

The association analysis of the extracted binary relationships revealed the multivariate relationships among various factors of the diseases. The following are the top 10 multivariate relationships ranked by support.

6. Diseases Pathogenesis Study

The following content represents the knowledge expression of the relationship between emotional factors and the treatment of cardiovascular diseases, derived through the mining of multivariate relationships, as shown in Figure 12.
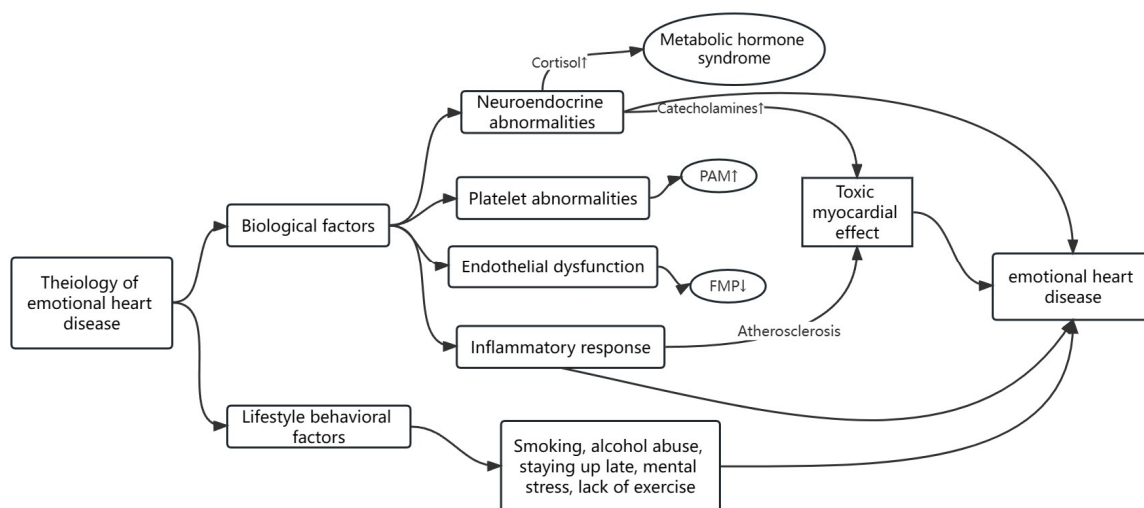

Figure 12. Schematic representation of the pathogenic mechanisms underlying emotional heart disease.

6.1. Biological Factors

Through the analysis of binary relationships extracted from disease entities, this study identified several key areas for analysis of the pathogenesis of emotional heart disease, including lifestyle factors, such as daily habits and lifestyle practices, and biological factors, such as various biological mechanisms and biochemical reactions within the organism.

(1) Inflammatory response (chest pain, headaches, etc.): inflammatory response is one of the important factors linking psychological disorders and emotional heart disease. Patients with psychological problems or under mental stress have elevated levels of inflammatory markers in their bodies. There is a correlation between elevated levels of inflammatory cytokines and the presence of emotional issues (depressive symptoms, anxiety symptoms, etc.) in patients, with significantly elevated levels of inflammatory markers such as C-reactive protein (CRP), pro-inflammatory cytokines (IL1β, IL2, IL6), and tumor necrosis factor-α (TNF-α) in patients with emotional problems. Compared to healthy individuals, the concentration of kyn trp^{-1} in patients' blood increases from 36.3 ± 13.26 µmol L^{-1} to 28.1 ± 5.15 µmol L^{-1}, while the concentration of tryptophan decreases from 8.51 ± 4.11 µmol L^{-1} to 5.84 ± 1.30 µmol L^{-1}. Based on the changes in kyn trp^{-1} and tryptophan indices, it is evident that cellular immune response has been activated, resulting in an increased rate of tryptophan degradation. This demonstrates that negative psychological factors such as stress and negative emotions can activate the body's stress pathways, leading to an inflammatory response, thereby causing heart problems in patients through symptoms such as arterial atherosclerosis.

(2) Endothelial dysfunction, manifested as hypertension and tachycardia, is a fundamental factor in acute coronary syndrome, which is a heart issue. Flow-mediated dilation (FMD) is used to quantify endothelial function. FMD refers to the metabolic waste produced by muscle contraction entering the bloodstream through the arteries, which is sensed by endothelial cells that then release signaling molecules. The FMD value of emotional heart disease patients, as analyzed, was $4.36 \pm 0.75\%$ when the value below 5%, while that of non-emotional heart disease patients was $7.46 \pm 0.89\%$. A healthy value should be greater than 10%. The FMD index indicates that emotional heart disease patients have endothelial dysfunction, indicating that psychological issues resulting in emotional disturbances (such as depressive and anxiety symptoms) play a role in the pathogenesis of emotional heart disease.

(3) Platelet abnormalities (thrombosis, etc.): platelets, endothelial components, and coagulation factors interact with each other, playing an important role in the process of thrombus formation. In the arteries of patients with atherosclerosis, serotonin mediates platelet aggregation by binding to 5-hydroxytryptamine (5-HT). In a healthy state, the serotonin uptake rate of platelets is between 50% and 80%, with a serotonin content of about 0.09–0.27 ng/108 platelets. The platelet release rate is usually between 20% and 70%. Patients with emotional symptoms (depression, anxiety, etc.) have abnormal levels of platelet serotonin, with a decrease in platelet serotonin transporter levels of 17.6%, and an increase in platelet serotonin receptor concentration of about 20.6%. Serotonin is an endogenous substance that mainly participates in the onset of depression, and it binds to 5-hydroxytryptamine (5-HT) receptors on platelets, promoting platelet function and affecting the process of blood coagulation.

(4) Abnormal neurotransmitters (palpitation, elevated blood glucose, etc.): there is an association between abnormal neurotransmitters and psychological as well as cardiac diseases. Higher concentrations of catecholamines (adrenaline and noradrenaline)—products of sympathetic adrenomedullary activation—have been observed in cardiac patients. Activation of the sympathetic adrenomedullary system leads to vasoconstriction, hypertension, increased heart rate, and platelet activation in cardiac patients. Analysis suggests that levels of adrenaline and cortisol in the blood of cardiac patients with emotional distress are elevated, possibly due to autonomic nervous system changes that increase their mortality rate. HPA (hypothalamic–pituitary–adrenal axis)-related abnormalities lead to higher than normal range indicators of adrenaline and dopamine (adrenaline: 107–412 pg/mL (M); 62–363 pg/mL (F), dopamine: 10–178 pg/mL (M); 10–150 pg/mL (F)), ultimately causing other clinical conditions such as metabolic disorders like obesity, hypertension, impaired glucose tolerance, hypertriglyceridemia, and hypercholesterolemia, which directly lead to adverse development of cardiovascular conditions.

(5) Heart rate variability (HRV): the normal range of HRV values in individuals can vary depending on factors such as age, gender, physical health status, and activity level. Generally, higher HRV values indicate better cardiac stability and stronger autonomic nervous system function. HRV is significantly lower in emotional heart disease patients compared to non-emotional heart disease patients (90 ± 35 vs. 117 ± 26 ms), and reduced HRV is an important factor in the onset and exacerbation of emotional heart disease.

6.2. Lifestyle Behavioral Factors

In the association between emotional heart disease and related factors, we also identified lifestyle factors such as sleep deprivation and dietary imbalance. These characteristics are often caused by other diseases such as anxiety disorders. However, they can also refer to unhealthy lifestyle habits, such as staying up late and under-eating. These unhealthy habits and lifestyle factors increase the risk of developing emotional heart disease.

6.3. Therapeutic Drugs Factors

For the treatment and prevention of diseases, in addition to in-depth research on the triggering factors of diseases, drug research can also be used to seek more effective treatment methods. In terms of therapeutic drugs, there are several drugs that are commonly used for patients with emotional heart disease based on the exploration of multiple relationships obtained. Tricyclic antidepressants (TCAs) work by increasing serotonin (not adrenaline) and dopamine levels in the brain. This type of drug is effective in relieving emotional symptoms, but its efficacy is limited. Antagonism of muscarinic acetylcholine receptors, histamine receptors, and α-1 adrenergic receptors can cause anticholinergic, cardiovascular, and nervous system side effects. Cardiovascular side effects are primarily tachycardia, arrhythmia, etc., and the dose of the drug needs to be arranged according to the specific condition of the patient. SSRIs: they are safer and more effective for patients with emotional heart disease because they have low or no affinity for histamine muscarinic and α receptors. Their relative safety and better acceptability make them a first-line drug for patients with emotional heart disease. Nifedipine: a calcium channel blocker, it can reduce the burden and oxygen consumption of the heart, regulate cardiovascular system function, and stabilize heart rate. It can effectively improve the symptoms of myocardial ischemia in the treatment of emotional heart disease, and the specific dosage needs to be customized according to individual conditions.

In summary, both lifestyle and biological factors have a directive role in the development and progression of emotional heart disease, making them key contributing factors to its occurrence. Prevention and treatment of emotional heart disease require consideration and intervention from multiple perspectives, including maintaining emotional stability, psychological health, and healthy lifestyle habits to reduce the risk of developing the illness.

7. Conclusions

Our proposed WBC model incorporates word embedding, BiLSTM, and conditional random fields to process words in text data to identify medical named entities, and finally Ab3p to identify abbreviations in medical named entities for correction. We experimentally validated the model using the NCBI medical annotation corpus, our constructed medical knowledge base, and Wikipedia medical category datasets, and our model showed a significant improvement in F1 score compared to several other methods; the F1 score of the WBC model is 89.2%. In the CSR model, the contextual semantic relationship between entities is considered, and the location information of entities is incorporated into the calculation process, while the semantic meaning of general domain vocabulary lacking in professional fields is supplemented. Finally, the attention mechanism is used to de-noise the location information and obtain the relationship between entities. In the verification experiment of dataset i2b2, good results are obtained, and the F1 score of the CSR model is 83.4%.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A publicly available dataset was analyzed in this study. The link to the NCBI disease corpus dataset is https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/ (accessed on 18 September 2023).

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

References

1. Névéol, A.; Li, J.; Lu, Z. Linking Multiple Disease-Related Resources through UMLS. In Proceedings of the ACM SIGHIT International Health Informatics Symposium, Miami, FL, USA, 28–30 January 2012; pp. 767–772.
2. Doğan, R.I.; Leaman, R.; Lu, Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inform.* **2014**, *47*, 1–10. [CrossRef]
3. Leaman, R.; Islamaj Doğan, R.; Lu, Z. DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics* **2013**, *29*, 2909–2917. [CrossRef] [PubMed]
4. Meystre, S.M.; Savova, G.K.; Kipper-Schuler, K.C. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb. Med. Inform.* **2008**, *17*, 128–144.
5. Eltyeb, S.; Salim, N. Chemical named entities recognition: A review on approaches and applications. *J. Cheminform.* **2014**, *6*, 17. [CrossRef]
6. Goulart, R.R.V.; Strube de Lima, V.L.; Xavier, C.C. A systematic review of named entity recognition in biomedical texts. *J. Braz. Comput. Soc.* **2011**, *17*, 103–116. [CrossRef]
7. Meystre, S.M.; Friedlin, F.J.; South, B.R. Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Med. Res. Methodol.* **2010**, *10*, 70. [CrossRef] [PubMed]
8. Rzhetsky, A.; Seringhaus, M.; Gerstein, M. Seeking a new biology through text mining. *Cell* **2008**, *134*, 9–13. [CrossRef]
9. Mikolov, T.; Sutskever, I.; Chen, K. Distributed representations of words and phrases and their compositionality. *Neural Inf. Process. Syst.* **2013**, *26*, 1–9.
10. Arnaud, É.; Elbattah, M.; Gignon, M.; Dequen, G. Learning Embeddings from Free-text Triage Notes using Pretrained Transformer Models. *HEALTHINF* **2022**, *5*, 835–841.
11. Wang, X.; Zhang, Y.; Li, Q. Distantly supervised biomedical named entity recognition with dictionary expansion. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 496–503.
12. Xu, K.; Yang, Z.; Kang, P. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. *Comput. Biol. Med.* **2019**, *108*, 122–132. [CrossRef]
13. Mu, X.; Wang, W.; Xu, A. Incorporating token-level dictionary feature into neural model for named entity recognition. *Neurocomputing* **2020**, *375*, 43–50.
14. Shang, J.; Liu, L.; Ren, X. Learning named entity tagger using domain-specific dictionary. *arXiv* **2018**, arXiv:1809.03599.
15. Fan, R.; Wang, L.; Yan, J. Deep learning-based named entity recognition and knowledge graph construction for geological hazards. *ISPRS Int. J. Geo-Inf.* **2019**, *9*, 15. [CrossRef]
16. Li, Y.; Shetty, P.; Liu, L. BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition. *arXiv* **2021**, arXiv:2105.12848.
17. Greenberg, N.; Bansal, T.; Verga, P. Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2824–2829.
18. Collobert, R.; Weston, J.; Bottou, L. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
19. van de Kerkhof, J. *Convolutional Neural Networks for Named Entity Recognition in Images of Documents*; Aalto University: Espoo, Finland, 2016.
20. Chiu, J.P.C.; Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]

21. Xin, J.; Lin, Y.; Liu, Z. Improving neural fine-grained entity typing with knowledge attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; p. 1.

22. De Magistris, G.; Russo, S.; Roma, P. An explainable fake news detector based on named entity recognition and stance classification applied to COVID-19. *Information* **2022**, *13*, 137. [CrossRef]

23. Mikolov, T.; Chen, K.; Corrado, G. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.

24. Gutmann, M.; Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; Proceedings of the JMLR Workshop and Conference Proceedings. pp. 297–304.

25. Liu, C.; Sun, W.; Chao, W. Convolution neural network for relation extraction. In *International Conference on Advanced Data Mining and Applications*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 231–242.

26. Zhang, Y.; Qi, P.; Manning, C.D. Graph convolution over pruned dependency trees improves relation extraction. *arXiv* **2018**, arXiv:1809.10185.

27. Dai, H.; Zhu, M.; Yuan, G.; Niu, Y.; Shi, H.; Chen, B. Entity recognition for Chinese hazardous chemical accident data based on rules and a pre-trained model. *Appl. Sci.* **2022**, *13*, 375. [CrossRef]

28. Panoutsopoulos, H.; Brewster, C.; Espejo-Garcia, B. Developing a Model for the Automated Identification and Extraction of Agricultural Terms from Unstructured Text. *Chem. Proc.* **2022**, *10*, 94.

29. Sun, M.; Yang, Q.; Wang, H.; Pasquine, M.; Hameed, I.A. Learning the Morphological and Syntactic Grammars for Named Entity Recognition. *Information* **2022**, *13*, 49. [CrossRef]

30. Cunha, L.F.C.; Ramalho, J.C. NER in Archival Finding Aids: Extended. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 42–65. [CrossRef]

31. Sboev, A.; Sboeva, S.; Moloshnikov, I.; Gryaznov, A.; Rybka, R.; Naumov, A.; Selivanov, A.; Rylkov, G.; Ilyin, V. Analysis of the Full-Size Russian Corpus of Internet Drug Reviews with Complex NER Labeling Using Deep Learning Neural Networks and Language Models. *Appl. Sci.* **2022**, *12*, 491. [CrossRef]

32. Wei, Z.; Su, J.; Wang, Y. A novel cascade binary tagging framework for relational triple extration. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 1476–1488.

33. 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, Berlin, Germany, 7–12 August 2016; pp. 1105–1106.

34. NCBI. Available online: https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/ (accessed on 18 September 2023).

35. Gauch, M.; Kratzert, F.; Klotz, D.; Nearing, G.; Lin, J.; Hochreiter, S. Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 2045–2062. [CrossRef]

36. Chen, K.; Wang, R.; Utiyama, M.; Sumita, E. Context-aware positional representation for self-attention networks. *Neurocomputing* **2021**, *451*, 46–56. [CrossRef]

37. Xu, K.; Zhou, Z.; Hao, T.; Liu, W. A bidirectional LSTM and conditional random fields approach to medical named entity recognition. In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Cairo, Egypt, 9–11 September 2017; Springer: Cham, Switzerland, 2018; pp. 355–365.

38. Savova, G.K.; Masanz, J.J.; Ogren, P.V.; Zheng, J.; Sohn, S.; Kipper-Schuler, K.C.; Chute, C.G. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 507–513. [CrossRef]

39. Wei, Q.; Chen, T.; Xu, R.; He, Y.; Gui, L. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *Database* **2016**, *2016*, baw140. [CrossRef]

40. Goldberg, Y.; Levy, O. Word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv* **2014**, arXiv:1402.3722.

41. Sohn, S.; Comeau, D.C.; Kim, W. Abbreviation definition identification based on automatic precision estimates. *BMC Bioinform.* **2008**, *9*, 402. [CrossRef]