Hindawi
Scientific Programming
Volume 2024, Article ID 7863381, 11 pages
https://doi.org/10.1155/2024/7863381

Research Article

Recognition of Hotspot Words for Disease Symptoms Incorporating Contextual Weight and Co-Occurrence Degree

Qingxue Liu,[1] **Lifang Wang**,[1] **Yuan Chang**,[2] **and Jixuan Zhang**[3]

[1]*School of Mechanical and Electrical Engineering, Kunming University, Kunming 650214, China*
[2]*School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China*
[3]*Juxian No. 2 Middle School, Juxian 276500, China*

Correspondence should be addressed to Qingxue Liu; hmxue2000@163.com

Received 23 January 2024; Revised 5 March 2024; Accepted 25 March 2024; Published 5 April 2024

Academic Editor: Pengwei Wang

Identifying hotspot words associated with disease symptoms is paramount for disease prevention and diagnosis. In this study, we propose a novel method for hotspot word recognition in disease symptoms, integrating contextual weights and co-occurrence information. First, we establish the MDERank model, which incorporates contextual weights. This model identifies words that align well with comprehensive weights, forming a collection of disease symptom words. Next, we construct a graph network for disease symptom words within each time period. Utilizing the graph attention network model, we incorporate word co-occurrence degree to identify potential hotspot words associated with disease symptoms. We conducted experiments using user-generated posts from the Dingxiangyuan Forum as our data source. The results demonstrate that our proposed method significantly improves the extraction quality of disease symptom words compared to other existing methods. Furthermore, the performance of our constructed recognition model for disease symptom hotspot words surpasses that of alternative models.

1. Introduction

Recently, the application and development of artificial intelligence in the medical field have gained extensive attention. Leveraging the analysis of large-scale medical data, we can extract disease symptom hotspot words to assist in disease-related endeavors. Disease symptom hotspot words refer to symptom-related terms that have garnered substantial attention and discussion within the medical field during specific time periods. Identifying these hotspot words can offer guidance to the medical community, enabling early detection, warnings, and proactive responses to potential disease risks. Consequently, the accurate identification of disease symptom hotspot words has emerged as a prominent research area.

Existing research primarily relies on electronic medical records and specialized medical datasets of patients for identification purposes [1]. However, this approach has several limitations. Electronic medical record data is typically obtained after patients' visits, resulting in identification occurring during the middle or later stages of a disease. Furthermore, the process of collecting and integrating medical data is time-consuming, making it challenging to promptly reflect on the most recent disease and medical conditions. Consequently, there is a delay in early warning and intervention efforts.

In contrast, utilizing data from online medical forums offers distinct advantages. These platforms capture users' conditions and requests for help in real-time, allowing for the timely identification of trending keywords related to potential disease symptoms. Patients tend to post various symptoms in various medical forums for consultation in the early stage of the disease. We can collect these symptoms and use them to predict upcoming epidemics. The doctor-patient communication information published in the medical forum has a strong timeliness and early warning effect. As a result, the identification process becomes more responsive and can facilitate early detection and intervention strategies.

Disease symptom words play a crucial role in recognizing and monitoring disease hotspots, as they reflect the signs and symptoms of a disease. Many researchers commonly employ keyword extraction techniques to obtain disease symptom words. However, when utilizing content from medical forums for disease symptom word extraction, challenges arise

due to the presence of noisy information and nonrelevant content.

Existing keyword extraction models like TF-IDF, YAKE [2], and TextRank [3] can assist in the extraction of disease symptom words. However, these models have limitations in considering contextual semantics, word co-occurrence degree, and the intrinsic properties of words. As a result, their targeting of disease descriptions for symptom word extraction is relatively weaker, leading to lower accuracy rates.

Certain researchers have explored the application of graph neural networks (GNNs) in modeling and recognizing disease symptom hotspot words. This approach effectively captures the intricate relationships between nodes and utilizes node information to extract context-aware features, thereby enhancing the precision and accuracy of disease symptom hotspot word recognition. In disease descriptions, there exists a certain level of association among disease symptom words, including co-occurrence and thematic relevance. Leveraging this association can facilitate a more effective determination of relationships and importance between nodes. However, the graph attention network (GAT) primarily relies on node connectivity, overlooking the significance and diversity of edge features. Consequently, this limitation restricts the model's expressive power to some extent.

To address the aforementioned issue and further enhance the quality of disease symptom hotspot identification, we propose a research framework that integrates contextual information for disease symptom word extraction and improves upon the GAT model for disease symptom hotspot word identification. The main objectives and contributions of this research framework are as follows:

(1) A novel method for disease symptom word extraction that incorporates contextual weights into the MDERank model is proposed. This method combines the contextual weights of words with their semantic relevance. By identifying words that align well with the integrated weights, we extract disease symptom words more effectively.
(2) We designed an improved GAT model for disease symptom hotspot word recognition, which incorporates word co-occurrence weights. By integrating co-occurrence degree into the edge features, this model enhances the learning of node representations, leading to improved accuracy in identifying disease symptom hotspot words.
(3) We construct a disease symptom word association graph and utilize the improved GAT model to implement the embedded representation of nodes. This approach enables the identification of disease symptom hotspot words through time series analysis. Experiment results demonstrate that our proposed method outperforms comparison methods in terms of the quality of disease symptom hotspot word recognition.

The rest of the paper is organized as follows: Section 2 provides an overview of existing research on the identification of hotspot words for disease symptoms. In Section 3, we present the proposed method for extracting disease symptom words. In Section 4, we detail the methodology to recognize hotspot words based on the extracted disease symptom words. Experiment is provided to evaluate the performance of the proposed method in Section 5. Finally, Section 6 concludes the paper and outlines future work.

2. Related Work

The open and communicative nature of online medical forums provides a valuable source of user-generated text, which presents an opportunity to extract disease symptom words and identify disease symptom hotspot words. To address this, researchers have explored various approaches, including the utilization of regression models.

Regression models involve building mathematical models that describe the relationship between known independent variables and corresponding dependent variable data. For instance, Yang et al. [4] constructed a word frequency matrix using TF-IDF and combined a Logistic growth model with a word frequency rate of change model to identify demand word attributes. Feng and Kong [5] employed weighted keyword word frequency analysis to calculate the integrated value of keywords, revealing research hotspots and change trends. These approaches demonstrate how regression models can be applied to extract meaningful information from user-generated text, aiding in the identification of disease symptom words and hotspot words.

In addition to the aforementioned approaches, other researchers have employed various techniques for analyzing and understanding disease-related information. For example, Zhong et al. [6] utilized machine learning and conducted bibliometric analysis. They employed co-occurrence relationships and clustering methods to identify the key causes of disease causation, specifically focusing on infectious disease research among liver transplant recipients [6]. Dong et al. [7] employed a topic modeling approach to analyze the semantic relationships among topics within a corpus. By comparing the distribution of topics between COVID-19 and other coronavirus infections, they aimed to explore the research hotspots in the field of disease infections [7].

Khan et al. [8] utilized convolutional neural networks and long short-term memory (LSTM) to extract multidimensional time-scale features. These features were then fed as inputs to the LSTM model, enabling the identification of potential disease outbreaks by learning representations from time-series data [8]. Zhang et al. [9] proposed a graph neural network incorporating attention. Their approach introduced an attention mechanism in the node feature representation and leveraged attention from neighbors during the embedding process to recognize node categories [9]. Chen et al. [10] employed BERT-BiLSTM to learn word embedding representations and contextual semantic relations. They then used graph convolutional networks (GCN) to utilize these representations as node features. The recognition results were obtained through a Softmax layer [10]. Peng et al. [11] utilized the self-attention mechanism of GATs to calculate node attention coefficients. By assigning different weights to nodes

Time: July 15, 2023
Title: Cervical pain with insufficient cerebral blood supply
Content: A 36-year-old female complained of frequent cervical pain and systemic soreness. Upon examination, she was found to have insufficient blood supply to the brain. Recently, she was found to have myopia of 120 degrees and occasionally fainted. May I ask where to check again?

Figure 1: An example of a disease description.

in the neighborhood, they achieved the recognition of node attributes [11]. These studies showcase a range of methodologies, including bibliometric analysis, topic modeling, deep learning, and graph neural networks, to analyze disease-related data and extract valuable insights.

To address the problems of noise, chromatic aberration, and detail distortion for enhancing low-light images using existing enhancement methods, Yang et al. [12] proposed an integrated learning approach (LightingNet) for low-light image enhancement. Similarly, he designed a powerful Vision Transformer-based Generative Adversarial Network (Transformer-GAN) for enhancing low-light images [13]. Guo et al. [14] proposed a deep dual-dynamic context-aware Poly (A) signal prediction model, called multiscale convolution with self-attention networks, to adaptively uncover the spatial–temporal contextual dependence information. He also presented a variational gated autoencoder-based feature extraction model to extract complex contextual features for inferring potential disease-miRNA associations [15]. Li et al. [16] devised an efficient gated convolutional recurrent network with residual learning to dynamically extract dependency patterns of raw genomic sequences in an efficient fusion strategy and successfully improve the predicting performance of the translation initiation sites.

Although the mentioned methods encompass various categories, they do not explicitly incorporate word co-occurrence weights in GATs. Recognizing this gap, we integrated word co-occurrence information into our disease symptom hotspot word recognition method through a correlation graph. This fusion of word co-occurrence with side features aims to enhance the overall quality of disease symptom hotspot word recognition. By incorporating word co-occurrence weights within the correlation graph, we can capture the relationships and associations between disease symptom words more effectively. This integration allows for a more comprehensive understanding of the context and connectivity among the words, thereby improving the accuracy and reliability of identifying disease symptom hotspot words.

The proposed method can accurately obtain hot symptom words from the disease help posts provided by patients, which is conducive to predicting the current popular diseases. Therefore, the research work of this paper can help the government or the hospital to make preparations for the prevention and treatment of the epidemic in advance with the help of the doctor-patient exchange information on the mutual benefit network in the early stage of the epidemic.

3. A Disease Symptom Word Extraction Model Incorporating Contextual Weights

Disease symptom words play a vital role in disease descriptions due to their high semantic relevance and importance. However, extracting these words from disease descriptions can be challenging due to the presence of noisy information, including emotional words and irrelevant topics. Such noise complicates the task of accurately identifying disease symptom words.

We extracted the hot words of disease symptoms from the help-seeking posts of disease diagnosis information published by patients on various doctor-patient communication websites such as Dingxiangyuan Forum. The following formal definitions are given for descriptive convenience.

Definition 1: Disease Description. A disease description is defined as a triplet, denoted as $d = (t, h, c)$, where t represents the date of the description, h represents the title, and c represents the content of the disease description.

Figure 1 illustrates a disease description example where disease symptom words, such as "cervical pain," "insufficient cerebral blood supply," "systemic soreness," "myopia," and "fainting," are identified. The hotspots of the disease symptoms are determined to be "cervical pain" and "insufficient cerebral blood supply." Table 1 presents the symbols and their meanings used in this paper.

MDERank [17] is a keyword extraction model that utilizes a pretraining approach. It obtains representations of masked documents and their originals by masking candidate words. These representations are then used to calculate similarities and rank the candidate words for keyword extraction. In comparison to other keyword extraction methods, MDERank maximizes the utilization of contextual semantic information and mitigates the issue of biased long keyword selection. However, MDERank does not account for information interference and the intrinsic properties of words themselves.

To address this, we propose incorporating contextual weight into MDERank to improve the extraction quality of disease symptom words. To achieve this, we introduce a disease symptom word extraction model called CW-MDERank (Contextual Weight MDERank). The model flowchart is depicted in Figure 2. CW-MDERank incorporates contextual weight, which considers the degree of association between a word and other words, as well as the word's importance

Table 1: Symbols and meanings.

Symbol	Description
d	Disease description
D	The set of disease descriptions
E_m	Masked vector for d
E_o	Original vector for d
MeanSemSim (w, d)	The mean value of semantic similarity between w and words in d
TF-IDF (w, d, D)	The value of TF-IDF for the word w in d
α	The weight balance factor
s_i	The ith disease symptom word
G	Disease symptom word association graph
$s_i \leftrightarrow s_j$.	s_i and s_j satisfy co-occurrence association
N_{co} (s_i, s_j)	The weight of the co-occurrence edge
$v_i^{(K)}$	Vector feature representation of the ith node in the Kth time step
$P^{(K)}$	The predicted value at Kth time step
p_i	The output result corresponding to the ith node
$L^{(K)}$	The loss at Kth time step

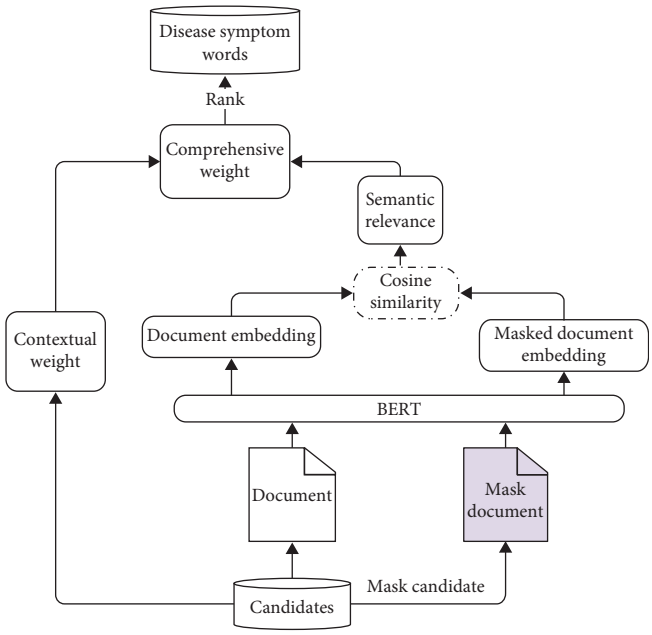

Figure 2: Framework of the CW-MDERank model.

within the dataset. By doing so, we aim to mitigate the issue of inappropriate extraction caused by noise interference in MDERank.

CW-MDERank builds upon the semantic similarity foundation of the MDERank model and enhances it by incorporating semantic relevance and contextual weights. By doing so, it provides a more comprehensive evaluation of the importance and semantic relevance of candidate words, ultimately improving the quality of disease symptom word extraction.

3.1. Semantic Relevance. w is a word contained in the disease description d, E_m is the masked description text vector generated by BERT after masking w, and E_o is the original description text vector generated by BERT. Both do similarity calculations, as shown in Equation (2).

$$E_m = \mathrm{BERT}(w, d),\ E_o = \mathrm{BERT}(d), \tag{1}$$

$$D(E_m, E_o) = \sqrt{\sum_{i=1}^{n} (E_m^i - E_m^i)^2}, \tag{2}$$

where n is the length of d. The higher the value of D (E_m, E_o), the less information is lost from the text after masking and the less important the words are. For this reason, the semantic relevance is defined as follows:

$$W_{\mathrm{sem}} = \frac{1}{D(E_m, E_o)}. \tag{3}$$

The larger the value, the more important the word is in the text.

3.2. Context Weight. D is a set of disease descriptions, w is a word contained in disease description d, MeanSemSim (w, d) is the mean of semantic similarity between w and words in d, and TF-IDF (w, d, D) is the TF-IDF value of the word w. The context weight is defined as follows:

$$W_{\mathrm{fre}} = \mathrm{MeanSemSim}(w, d) \times \mathrm{TF} - \mathrm{IDF}(d, w, D). \tag{4}$$

The comprehensive weight of a word in a disease description is determined based on the two feature weights mentioned previously. This weight value is calculated according to Equation (5). A higher weight value signifies a greater importance of the word within the disease description d.

$$\mathrm{CW}_{\mathrm{MDERank}(w)} = \alpha W_{\mathrm{sem}} + (1 - \alpha) W_{\mathrm{fre}}, \tag{5}$$

where α is the weight balance factor. The top K words ranked by their comprehensive weights are selected as the extracted disease symptom words from d, denoted as $S = \{s_1, s_2, \ldots, s_K\}$.

4. Incorporating Co-Occurrence for Recognizing Disease Symptom Hotspot Words

After extracting disease symptom words, many methods employ time series analysis and curve-fitting techniques to infer future trends and identify disease symptom hotspots. However, these methods often overlook the interword associations within the data. GNNs are a class of neural network models specifically designed to handle graph-structured data, enabling the learning and representation of relationships between nodes. Popular GNN models include GCN, GraphSAGE, and GATs [18]. Among these models, GAT stands out by utilizing graph attention mechanisms to learn relationships between nodes. It calculates attention coefficients between vertices and their neighboring nodes, aggregating node features. GAT is advantageous as it does not solely

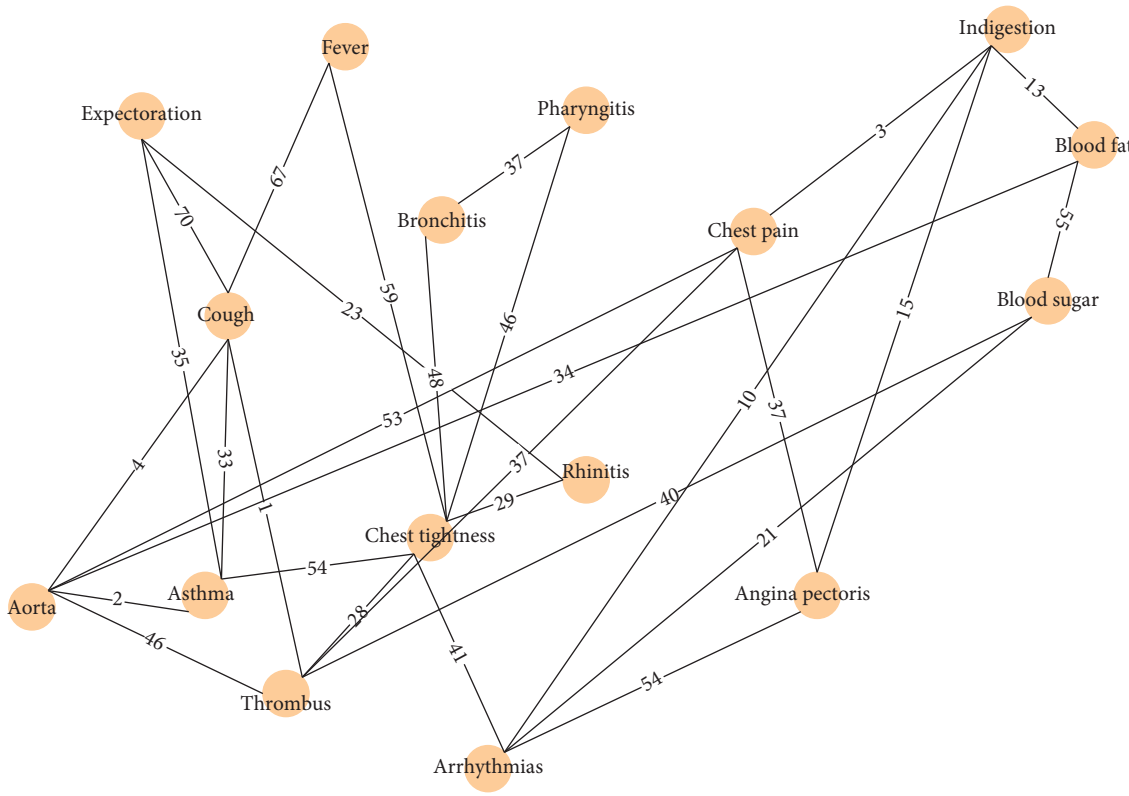

FIGURE 3: Disease symptom word association graph.

rely on graph structure information and effectively captures correlations between node features [19].

However, GAT solely considers the connectivity of edges and does not fully leverage edge features. In cases where two disease symptom words frequently co-occur in the text, there may exist some relationship or correlation, such as semantic association or contextual dependency. Consequently, the GAT model may overlook important information, leading to a potential loss of information and a decrease in recognition accuracy.

To address these limitations, we propose a disease symptom hotspot word recognition model called CO-GAT (co-occurrence-based GAT). CO-GAT incorporates the co-occurrence degree and accounts for the time series data. To facilitate better understanding, the following definitions are provided:

Definition 2: Co-Occurrence Correlation. Two disease symptom words, s_i and s_j, are considered to be a co-occurrence association if they exist in the same disease description d. They are denoted as $s_i \leftrightarrow s_j$.

Definition 3: Disease Symptom Word Association Graph. $G = \{V, E, W\}$ is a disease symptom word association graph, where

(1) $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes, and node v_i denotes the disease symptom word s_i;

(2) $E = \{e_1, e_2, \ldots, e_m\}$ is the set of edges, the s_i and s_j corresponding to v_i and v_j in $e = (v_i, v_j)$ satisfy $s_i \leftrightarrow s_j$;

(3) $W = \{w_{ij}\}$ is the set of edge weights, for $\forall e = (v_i, v_j) \in E$, $w_{ij} = N_{co}(s_i, s_j)$. Where $N_{co}(s_i, s_j)$ represents the weight of the co-occurrence edge, which denotes the number of times s_i and s_j appear together in the same description d.

Figure 3 illustrates a portion of the association graph constructed using disease symptom words extracted from disease descriptions posted on the Dingxiangyuan Forum platform. In this graph, each node represents a specific disease symptom word, while the edges represent the co-occurrence relationships between these words. The numerical values assigned to the edges indicate the corresponding edge weights, providing a measure of the strength or significance of the co-occurrence relationship.

Within the GAT model, node vectors are initially generated through random initialization and iteratively updated during the training process using self-attention mechanisms. However, it is crucial to incorporate the semantic information of words as an additional feature. To tackle this, the CO-GAT model integrates Word2Vec to generate word embeddings for each word. These word embeddings are subsequently employed as the initial embedding vectors for the nodes. This initialization step facilitates the creation of more expressive node representations, ultimately enhancing the learning and modeling of node relationships. The initial time step node embedding vectors can be characterized as follows:

$$v_i^{(0)} = \text{Word2Vec}(s_i), v_i \in V, \tag{6}$$

where $v_i \in R^F$ is the initial node embedding representation of the corresponding disease symptom word s_i.

In the CO-GAT model, we introduce the co-occurrence degree of nodes as an additional component within the edge features. This incorporation of the co-occurrence degree serves as the initial edge embedding representation, contributing to the improved learning and modeling of relationships between nodes.

$$u_{ij}^{(0)} = Ow_{ij}, w_{ij} \in W. \tag{7}$$

In this process, a trainable transfer matrix denoted as $O \in R^F$ is utilized to map the co-occurrence degree of disease symptom word to high-dimensional feature vectors. Then, the node embedding representations and edge embedding representations of the current time step are inputted into the model. To replace the fixed normalization operation in graph convolution, the attention mechanism is applied, assigning attention to the set of nodes $N_i^{(0)}$ in the neighborhood of node v_i to learn the weights between the nodes. The self-attention weights, as in Equation (8), are constructed as the importance of node v_j for node v_i. where $W_{ij} \in R^{F' \times F}$ are the parameters to be trained, and b_{ij} is the bias matrix.

$$e_{ij} = \text{sigmoid}\left(W_{ij} \cdot v_i^{(0)} + W_{ij} \cdot v_j^{(0)} + W_{ij} \cdot u_{ij}^{(0)} + b_{ij}\right). \tag{8}$$

To facilitate the comparison of e_{ij} between different neighboring nodes v_j of node v_i, we normalize it using the softmax function, as shown in Equation (9). This value represents the importance coefficient of node v_j to node v_i.

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{n=1}^{|N_i^{(0)}|} \exp(e_{in})}. \tag{9}$$

Finally, the node features are aggregated and updated at the current time step to obtain the output of the node features for the next time step, as depicted in Equation (10).

$$v_i^{(1)} = \sum_{n=1}^{|N_i^{(0)}|} \alpha_{in} \cdot v_n^{(0)}. \tag{10}$$

The output of each time step is used as the input for the next time step, then the output of the node at step K is as follows:

$$v_i^{(K)} = \sum_{n=1}^{|N_i^{(K-1)}|} \alpha_{in} \cdot v_n^{(K-1)}. \tag{11}$$

The set of node vectors is obtained as $V^{(K)} = \{v_1^{(K)}, \ldots, v_i^{(K)}, \ldots, v_n^{(K)}\}$, with n being the number of nodes, which is passed through the fully connected layer to obtain the prediction of the nodes, as shown in Equation (12).

$$P^{(K)} = \text{sigmoid}\left(WV^{(K)} + b\right). \tag{12}$$

In the given equation, $P = \{p_1, .., p_i, \ldots, p_N\}$, p_i represents the prediction result indicating whether node v_i is a hotspot word for disease symptoms, W is a trainable parameter, and b is a bias matrix.

The model is optimized using the cross-entropy loss function, and the loss function for the Kth time step is defined as Equation (13):

$$L^{(K)} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i)], \tag{13}$$

where y is the true label, and p is the predicted value.

Finally, the loss function is constructed by taking the sum of the losses of all time steps as the final loss of the model, denoted as $L = L^{(1)} + \ldots + L^{(K)} + \ldots + L^{(T)}$, which is used to minimize the difference between the model's predicted results and the ground truth results, where T represents the total number of time steps.

5. Experimentation

5.1. Data Set and Parameter Settings. The data source for our study consisted of disease descriptions posted by users in the Clinical Internal Medicine section of the Dingxiangyuan Forum, covering the period from January 2019 to March 2023. To ensure data quality, we excluded disease descriptions that were deemed too short, resulting in a total of 89,644 valid disease descriptions. Python was employed for data preprocessing tasks, such as segmentation and removal of stop-words.

To identify disease symptom words and disease symptom hotspot words, we employed the method of expert annotation. Experts with authoritative status in relevant fields were invited to carry out the annotation work, ensuring the accuracy and credibility of the results.

To evaluate the recognition effect, we divided the dataset into 19 subsets based on quarters. We then employed the Monte Carlo cross-validation method, taking 10 consecutive subsets at a time. The first seven subsets were used as the training set, while the last three subsets served as the testing set. A sliding window of four subsets was utilized, and these subsets are denoted as DS1–DS3.

The configuration of the experimental machine is Intel 12,900 k, 128 GB RAM, ubuntu18.02 operating system, GPU is RTX3090 *2, 24 GB video memory, and the programing language is Python3.8 + pytorch1.8. The optimal parameter Settings of the adopted model or method are shown in Table 2.

We record the results of different settings during the experiment by observing the curves of training loss and validation performance and perform comparative analysis to select the optimal model parameters. The choice of batch size depends on multiple factors, such as hardware resources and dataset size. An appropriate batch size can affect the training speed and memory utilization efficiency of the

Table 2: Experiment parameters.

Parameter name	Parameter value
Batch size	32
Train epoch	10
Word2Vec dim	300
Learning rate	5e-5

model. A larger batch size can increase the training speed, but it may also lead to insufficient memory. Considering the size of the dataset, we set the batch size to 32 in this study to ensure that more data is used to update the model parameters at each training step.

For the choice of train epoch, it indicates the number of times the model traverses the entire training dataset. The selection of the number of training epochs in this study is based on the early stopping method, where we observe the performance of the model on the training set. When the performance on the training set tends to stabilize, we stop training, and in this case, we choose this value as 10.

Regarding Word2Vec, it determines the dimensionality of the vector representation of each word. Generally, a larger dimension can capture more semantic information, but it also requires more training data and computational resources. In this study, we choose a moderate dimensionality of 300 to maintain sufficient semantic information while avoiding excessive computational resources and training data requirements.

As for the learning rate, it is a parameter that controls the step size of model weight updates, which greatly affects the convergence speed and stability of the model. We initially set the learning rate value to 0.001 and adopt learning rate scheduling technique to gradually adjust the learning rate. The learning rate is gradually reduced according to a certain rule, decreasing by a certain proportion at the end of each epoch. We observe the loss of the model on the training set and finally set this value to 5e-5.

5.2. Evaluation Indicators. The evaluation metrics utilized for assessing the recognition quality of disease symptom hotspot words were *Recall*, *Precision*, and *F1* score. In the context of these metrics, TP, FP, and FN represent positive samples predicted to be in the positive category, negative samples predicted to be in the positive category, and positive samples predicted to be in the negative category, respectively.

(1) *Recall*: *Recall* is calculated as the ratio of correctly predicted disease symptom hotspot words to the total number of actual disease symptom hotspot words. A higher *Recall* value indicates a greater proportion of disease symptom hotspot words that are correctly predicted, indicating a higher quality of recognition. The equation for *Recall* is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{14}$$

(2) *Precision*: *Precision* is determined by the number of correctly predicted disease symptom hotspot words divided by the total number of predicted disease symptom hotspot words. A higher *Precision* value signifies a larger proportion of correctly predicted disease symptom hotspot words and reflects a higher quality of recognition. The *Precision* equation is given by the following:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{15}$$

(3) *F1* score: *F1* Score serves as a composite metric for evaluating the performance of the model; the larger the value, the more robust the model is as follows:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{16}$$

5.3. Experimental Results and Analysis

5.3.1. Quality Evaluation of CW-MDERank. We selected YAKE, TextRank, KeyBERT [20], and MDERank [17] as the comparison models. YAKE and TextRank are statistically based methods, while KeyBERT and MDERank are neural network models based on BERT. We extracted Symptom words from the disease descriptions using different models, and each evaluation index is shown in Table 3.

Among these models, TextRank exhibited the poorest performance in terms of accuracy, *recall*, and *F1* score, resulting in the lowest quality of disease symptom words. This can be attributed to TextRank's inability to capture semantic information and its heavy reliance on high-frequency words. YAKE, on the other hand, demonstrated significantly better extraction quality than TextRank. YAKE considers the positional information of the words in the text and models the positional weights to access the words importance. However, both are limited in their ability to the semantic metrics.

The KeyBERT model, which employs BERT embeddings and cosine similarity, outperformed TextRank and YAKE in extracting symptom words. KeyBERT effectively utilizes contextual semantic information from the text, leading to improved extraction quality. Building upon KeyBERT, MDERank addresses the mismatch between words and document representations by converting from phrase-text level to text–text level for similarity computation. As a result, MDERank achieved higher metric scores compared to the previous three models, indicating superior performance. Overall, the comparison models demonstrated varying levels of performance in terms of accuracy, *recall*, and *F1* score, with MDERank exhibiting the highest scores, followed by KeyBERT, YAKE, and TextRank.

Among all the models, CW-MDERank achieved the highest *Recall* and *Precision* scores, surpassing MDERank with an average improvement of 2.4% and 3.57%, respectively. This

Table 3: Performance comparison of different disease symptom feature word extraction models.

Data set	Metrics	TextRank	YAKE	KeyBERT	MDERank	CW-MDERank
DS1	*Recall*	0.488	0.543	0.575	0.599	0.628
	Precision	0.455	0.52	0.535	0.552	0.589
	F1 score	0.471	0.531	0.554	0.575	0.608
DS2	*Recall*	0.491	0.559	0.599	0.608	0.639
	Precision	0.452	0.5	0.534	0.571	0.614
	F1 score	0.471	0.528	0.565	0.589	0.626
DS3	*Recall*	0.495	0.577	0.616	0.619	0.631
	Precision	0.435	0.477	0.488	0.554	0.581
	F1 score	0.463	0.522	0.545	0.585	0.605

notable improvement can be attributed to the incorporation of contextual weights in CW-MDERank, which helps address noise and irregularities in disease descriptions. This model not only leverages contextual semantic information but also enhances the attribute features of individual words. Consequently, CW-MDERank achieves superior recognition quality for disease symptom words.

Across different datasets, CW-MDERank consistently outperformed other models in terms of metric scores. It effectively considers factors such as word similarity and semantic relevance, resulting in the highest quality of disease symptom word recognition.

5.3.2. Ablation Experiment. Our proposed model, CWC-GAT (contextual weights co-occurrence GAT), primarily consists of two main components: MDERank (CW-MDERank) and GAT (CO-GAT). These components are fuzed together within the CWC-GAT framework to achieve the desired model performance.

GAT-MDERank: This variant of the CWC-GAT model removes the incorporation of contextual weights in CW-MDERank and the co-occurrence degree in CO-GAT. Instead, it consists of MDERank and the GAT. MDERank is responsible for generating disease symptom words, while the GAT performs disease symptom hotspot word recognition.

GAT-CW-MDERank: In this variant, the incorporation of the co-occurrence degree in CO-GAT is removed. It primarily consists of CW-MDERank and the GAT. CW-MDERank is responsible for extracting disease symptom words, while the GAT focuses on disease symptom hotspot word recognition.

CWC-GAT: Our proposed model, CWC-GAT, encompasses both CW-MDERank and CO-GAT. CW-MDERank is responsible for extracting disease symptom words from disease descriptions, while CO-GAT identifies disease symptom hotspot words from the extracted symptom words.

The recognition results of different datasets show that CWC-GAT improves synchronously in all metrics compared to GAT-CW-MDERank and GAT-MDERank. After the introduction of the contextual weighting of words, GAT-CW-MDERank improves in *Recall*, *Precision*, and *F1*, with an average improvement of 5.33%, 4.6%, and 4.94%, respectively. Comparing GAT-CW-MDERank, CWC-GAT improved the *Recall*, *Precision*, and *F1* scores by 3%, 2.47%, and 2.71% on average. It indicates that incorporating co-occurrence degree into node relations can effectively improve the quality of disease symptom hotspot word recognition.

5.3.3. Comparison of CWC-GAT with Other Methods. For the purpose of verifying the advancement of our proposed method, we selected identification methods published in domestic journals within the last 3 years for comparison. The identification quality index is presented in Table 4, and it includes the following methods:

(1) FP-tree [21]: This method utilizes an improved version of the FP-tree algorithm to extract recurring words as candidate hotspot words. These candidates are then expanded into multivariate pointwise mutual information (PMI) based on binary PMI. The method introduces temporal features of hotspot words by incorporating time pointwise mutual information (TPMI). Finally, neighbor entropy is employed to determine candidate boundaries and screen out the final hotspot words.

(2) I-BERT [22]: The I-BERT method segments the text description based on composite keywords. It obtains word vector representations using the BERT model and represents each composite word as a collection of lexical meanings after segmentation. Density clustering is performed, and the retained centers are concatenated to obtain the centers of the keyword collection, which are considered hotspot words.

(3) L-ATTN [23]: In this method, hotspot words are extracted using an improved Latent Dirichlet Allocation (LDA) model. Additionally, a recognition model based on the attention mechanism and LSTM network is proposed to predict the popularity and long-term trends of hotspot words. This information is then utilized to recognize disease symptom hotspot words.

(4) BBGANS [24]: BBGANS encodes syntactic features, such as contextual features and intersentence dependencies, using BioBERT. The method generates fusion representations by incorporating contextual and syntactic features with the help of GATs. Finally, the softmax function is employed to compute the values and obtain the results of disease symptom hotspot words.

Table 4: Performance comparison of CWC-GAT and other methods.

Data set	Metrics	FP-tree	I-BERT	L-ATTN	BBGANS	CWC-GAT
DS1	*Recall*	0.411	0.478	0.538	0.591	0.643
	Precision	0.356	0.395	0.443	0.479	0.532
	F1 score	0.382	0.433	0.486	0.529	0.582
DS2	*Recall*	0.388	0.451	0.526	0.589	0.624
	Precision	0.349	0.388	0.428	0.466	0.514
	F1 score	0.367	0.417	0.472	0.520	0.564
DS3	*Recall*	0.395	0.456	0.533	0.608	0.636
	Precision	0.365	0.382	0.434	0.473	0.535
	F1 score	0.379	0.416	0.478	0.532	0.581

Table 5: Results of ablation experiment.

Data set	Metrics	GAT-MDERank	GAT-CW-MDERank	CWC-GAT
DS1	*Recall*	0.556	0.616	0.643
	Precision	0.465	0.507	0.532
	F1 score	0.506	0.556	0.582
DS2	*Recall*	0.548	0.596	0.624
	Precision	0.456	0.488	0.514
	F1 score	0.498	0.537	0.564
DS3	*Recall*	0.549	0.601	0.636
	Precision	0.448	0.512	0.535
	F1 score	0.493	0.553	0.581

Table 5 provides a performance comparison of each model across different datasets. It can be observed that models utilizing the attention mechanism outperform FP-tree and I-BERT. This finding indicates that the introduction of the attention mechanism in the hotspot word recognition task effectively enhances the performance of the models.

Moreover, our proposed CWC-GAT model consistently outperforms previous models in terms of evaluation metrics. When compared to FP-tree, I-BERT, L-ATTN, and BBGANS, CWC-GAT exhibits significant improvements in the *Recall* metric, with average enhancements of 23.63%, 17.3%, 10.2%, and 3.83%, respectively. This indicates that CWC-GAT excels in identifying a greater number of true disease symptom hotspot words. In *Precision*, CWC-GAT demonstrates average improvements of 17.03%, 13.87%, 9.2%, and 5.43%, respectively, indicating a higher proportion of correctly identified disease symptom hotspot words compared to alternative methods. Moreover, in terms of the *F1* score, CWC-GAT exhibits average improvements of 19.96%, 15.39%, 9.69%, and 4.85%, respectively, indicating better overall recognition performance of the model.

Among the compared methods, FP-tree exhibits the lowest recognition quality. It overlooks polysemy and word disambiguation, solely relying on frequency statistics within different time windows for recognizing disease symptom hotspot words. Consequently, its quality is relatively poor. I-BERT employs the BERT model to generate vector representations and conducts clustering based on semantic representations for hotspot word recognition. However, it is more sensitive to noise and outlier points.

The L-ATTN method incorporates the attention mechanism into the LSTM network, assigning varying weights to individual words. BBGANS learns vector representations of contextual features, models node relationships, and assigns weights based on their importance. However, both methods lack consideration of node association relationships and fail to obtain richer features for attention mechanism calculation [25]. Consequently, their combined mean values for the three metrics are lower compared to our proposed CWC-GAT method.

6. Conclusion

To enhance the recognition quality of disease symptom hotspot words, we have developed a method that incorporates both contextual weights and co-occurrence degrees. Our approach begins with the MDERank model, which utilizes contextual weights to extract disease symptom words and effectively remove text information noise interference. Building upon this, we have constructed an improved GAT model that incorporates co-occurrence degree. By integrating the interword co-occurrence degree into the edge features, we enhance the representation of internode relationships, resulting in higher-quality recognition of disease symptom hotspot words. Experimental results have demonstrated the superiority of our method compared to the comparison method, as

indicated by improved *precision*, *recall*, and other evaluation metrics.

Future work will focus on optimizing the calculation method of contextual weights for words in disease descriptions. We also aim to further enhance the GAT model to improve the recognition quality of disease symptom hotspot words. Additionally, we plan to model and cluster potential disease symptom hotspot words for analysis, enabling us to predict the emergence of potential diseases.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the Special Basic Cooperative Research Programs of Yunnan Provincial Undergraduate Universities' Association (Nos. 202301BA070001-003, 202001BA070001-197, and 202001BA070001-173), the Foundation of Yunnan Province Science and Technology Department (Nos. 202305AO350007 and 202305AP350017), and Kunming University Foundation (No. YJL2205).

References

[1] T. Sarwar, S. Seifollahi, J. Chan et al., "The secondary use of electronic health records for data mining: data characteristics and challenges," *ACM Computing Surveys*, vol. 55, no. 2, pp. 1–40, 2022.

[2] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.

[3] W. Guo, Z. Wang, and F. Han, "Multifeature Fusion keyword extraction algorithm based on TextRank," *IEEE Access*, vol. 10, pp. 71805–71813, 2022.

[4] S. Yang, H. Wu, and S. Li, "Research on the attribute sorting of demand words in online medical communities: taking the Dingxiangyuan Forum as an example," *Intelligence Exploration*, vol. 1, no. 2, pp. 1–10, 2022.

[5] G. Feng and Y. Kong, "Research on disciplinary hotspots based on time-weighted keyword frequency analysis," *Journal of the China Society for Scientific and Technical Information*, vol. 39, no. 1, pp. 100–110, 2020.

[6] H. Zhong, C. Liu, Y. Dai et al., "A bibliometric analysis of infectious diseases in patients with liver transplantation in the last decade," *Annals of Translational Medicine*, vol. 9, no. 22, Article ID 1646, 2021.

[7] M. Dong, X. Cao, M. Liang, L. Li, G. Liu, and H. Liang, "Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modeling," Medrxiv, 2020.

[8] S. D. Khan, L. Alarabi, and S. Basalamah, "Toward smart lockdown: a novel approach for COVID-19 hotspots prediction using a deep hybrid neural network," *Computers*, vol. 9, no. 4, Article ID 99, 2020.

[9] J. Zhang, M. Li, K. Gao, S. Meng, and C. Zhou, "Word and graph attention networks for semi-supervised classification," *Knowledge and Information Systems*, vol. 63, no. 11, pp. 2841–2859, 2021.

[10] Z. Chen, R. Qi, and S. Li, "BiLSTM-based with word-weight attention for Chinese named entity recognition," in *2022 IEEE 13th International Conference on Software Engineering and Service Science*, pp. 150–154, IEEE, Beijing, China, 2022.

[11] M. Peng, B. Cao, J. Chen, J. Liu, and B. Li, "SC-GAT: web services classification based on graph attention network," in *CollaborateCom 2020: 16th EAI International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pp. 513–529, Springer, Shanghai, China, 2021.

[12] S. Yang, D. Zhou, J. Cao, and Y. Guo, "LightingNet: an integrated learning method for low-light image enhancement," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 29–42, 2023.

[13] S. Yang, D. Zhou, J. Cao, and Y. Guo, "Rethinking low-light enhancement via transformer-GAN," *IEEE Signal Processing Letters*, vol. 29, pp. 1082–1086, 2022.

[14] Y. Guo, D. Zhou, P. Li, C. Li, and J. Cao, "Context-aware poly (A) signal prediction model via deep spatial-temporal neural networks," in *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, IEEE, 2022.

[15] Y. Guo, D. Zhou, X. Ruan, and J. Cao, "Variational gated autoencoder-based feature extraction model for inferring disease-miRNA associations based on multiview features," *Neural Networks*, vol. 165, pp. 491–505, 2023.

[16] W. Li, Y. Guo, B. Wang, and B. Yang, "Learning spatiotemporal embedding with gated convolutional recurrent networks for translation initiation site prediction," *Pattern Recognition*, vol. 136, Article ID 109234, 2023.

[17] L. Zhang, Q. Chen, W. Wang et al., "MDERank: a masked document embedding rank approach for unsupervised keyphrase extraction," in *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 396–409, Association for Computational Linguistics, Linguistics, Dublin, Ireland, 2022, 2022.

[18] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2020.

[19] Q. Hu, J. Shen, K. Wang, J. Du, and Y. Du, "A web service clustering method based on topic enhanced Gibbs sampling algorithm for the Dirichlet Multinomial Mixture model and service collaboration graph," *Information Sciences*, vol. 586, pp. 239–260, 2022.

[20] B. Issa, M. B. Jasser, H. N. Chua, and M. Hamzah, "A comparative study on embedding models for keyword extraction using KeyBERT method," in *2023 IEEE 13th International Conference on System Engineering and Technology*, pp. 40–45, IEEE, Shah Alam, Malaysia, 2023.

[21] Y. Wang and J. Xu, "Hot new word discovery applied for detection of network hot news," *Journal of Computer Applications*, vol. 40, no. 12, pp. 3513–3519, 2020.

[22] B. Liu, Z. Lv, N. Zhu, D. Chang, and M. Lu, "Hot keyword extraction of scitech periodicals based on the improved BERT model," *KSII Transactions on Internet & Information Systems*, vol. 16, no. 6, pp. 1–18, 2022.

[23] L. Fu and F. Zhao, "Prediction of hot topics of agricultural public opinion based on attention mechanism LSTM model," *International Journal of Agricultural and Environmental Information Systems*, vol. 12, no. 4, pp. 1–16, 2021.

[24] X. Zheng, H. Du, X. Luo, F. Tong, W. Song, and D. Zhao, "BioByGANS: biomedical named entity recognition by fusing contextual and syntactic features through graph attention network in a node classification framework," *BMC Bioinformatics*, vol. 23, no. 1, pp. 1–19, 2022.

[25] Q. Hu, H. Qi, W. Huang, and M. Liu, "A method to recommend cloud manufacturing service based on the spectral clustering and improved slope one algorithm," *Journal of Cloud Computing*, vol. 12, no. 1, Article ID 115, 2023.