Counterfactual Supporting Facts Extraction for Explainable Medical Record Based Diagnosis with Graph Network

Haoran Wu[1,2], **Wei Chen**[1], **Shuang Xu**[1] and **Bo Xu**[1,2]
[1]Institute of Automation, Chinese Academy of Sciences,
Beijing, 100190, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences,
Beijing, 100049, China
{wuhaoran2018, w.chen, shuang.xu, xubo}@ia.ac.cn

Abstract

Providing a reliable explanation for clinical diagnosis based on the Electronic Medical Record (EMR) is fundamental to the application of Artificial Intelligence in the medical field. Current methods mostly treat the EMR as a text sequence and provide explanations based on a precise medical knowledge base, which is disease-specific and difficult to obtain for experts in reality. Therefore, we propose a counterfactual multi-granularity graph supporting facts extraction (CMGE) method to extract supporting facts from the irregular EMR itself without external knowledge bases in this paper. Specifically, we first structure the sequence of the EMR into a hierarchical graph network and then obtain the causal relationship between multi-granularity features and diagnosis results through counterfactual intervention on the graph. Features having the strongest causal connection with the results provide interpretive support for the diagnosis. Experimental results on real Chinese EMRs of the lymphedema demonstrate that our method can diagnose four types of EMRs correctly, and can provide accurate supporting facts for the results. More importantly, the results on different diseases demonstrate the robustness of our approach, which represents the potential application in the medical field[1].

1 Introduction

Electronic Medical Record (EMR) based diagnosis has attracted extensive attention due to its comprehensive historical information and clinical descriptions with the development of natural language processing and medical informatics (Yang et al., 2018; Choi et al., 2018; Liu et al., 2019; Dong et al., 2020; Ma et al., 2020b). The application of deep learning in medicine requires adequate medical explanations for the result. Specific to the diagnosis of EMR, the model needs to provide the text description supporting the diagnosis results.

① **Age:** 49
② **Gender:** woman
③ **Document:** In December 2011, the patient underwent surgical treatment for endometrial cervical cancer in a local hospital, and received radiotherapy 12 times after surgery. Regular reexamination, there are no signs of tumor recurrence…(患者于2011年12月因子宫内膜宫颈癌于当地医院行手术治疗，术后行放射治疗12次。定期复查，未见肿瘤复发征象……)
④ **Diagnosis:** Secondary lymphedema of right lower limb, after endometrial cancer surgery(右下肢继发性淋巴水肿,子宫内膜癌术后)
⑤ **Supporting Facts:** received radiotherapy 12 times after surgery(术后行放射治疗12次)

Figure 1: An example of EMR. We consolidated the various parts of the EMR into a single document as input and our goal is to extract supporting facts at the granularity of the clause.

As shown in Figure 1, an irregular EMR is a document of disease-related information, including symptoms, history of the disease, preliminary examination results, and so on, which is disordered and sparse with meaningless noisy text. Existing methods provide explanation through medical entities (Yuan et al., 2020), text spans (Mullenbach et al., 2018) and the weights of external knowledge (Ma et al., 2018). The entity is critical to the diagnosis (Sha and Wang, 2017; Girardi et al., 2018), but for the medical explanation, it cannot provide specific information of symptoms (such as positive or negative). And the form of the span is too fragmented and lacks readability. Therefore, the clause as a more informative and readable representation is needed to be combined above the level of entities.

Most of the previous methods provide reliable explanations for diagnosis by calculating the similarity with an external medical knowledge base (ICD[2] and CCS[3]) (Xu et al., 2019, 2020). KAME

[1]The code is available at https://github.com/CKRE/CMGE

[2]https://www.cdc.gov/nchs/icd/icd10cm.htm
[3]https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1942–1955
June 6–11, 2021.

(Ma et al., 2018) uses the weights of the nodes in the introduced knowledge graph to provide explanations. Depending on the hierarchical relations in the database, GMAN (Yuan et al., 2020) builds a disease hierarchy graph and a causal graph to find critical entities. However, a trusted medical knowledge base requires a mass of expertise in different fields to build, and it may be incomplete or erroneous in practical clinical applications. So far, how to extract supporting facts from the EMR itself without an external medical knowledge base is still a problem.

Counterfactual reasoning provides a link between what could have happened when inputs had been changed (Verma et al., 2020). Doctors usually make a judgment based on several related symptoms during diagnosing a disease. In this regard we can consider a question: will a doctor make a misdiagnosis without one of the critical symptoms? The result is clear. In a counterfactual way, if we gradually weaken the features until the diagnosis changes dramatically, then this feature can be considered as a supporting fact.

Based on this consensus, we propose a counterfactual multi-granularity graph supporting facts extraction (CMGE) method for the irregular EMR in this paper. First, we model the EMR as a hierarchical graph structure, which contains sentences, clauses, and entities. Specifically, sentences are used to model the temporal relationship, clauses provide a complete descriptive explanation, and entities provide symptom support as others. On this basis, we use a graph attention network to aggregate all information from different granularities. Then, we can do a counterfactual intervention to obtain the causal relation between feature and diagnosis. Specifically, we train a learnable soft-mask matrix to mask the feature of nodes or edges in the graph while keeping the diagnosis unchanged, and the remaining features are the supporting facts of the diagnosis. Counterfactual reasoning on the graph requires enhancing the medical features contained in the text of different granularity, so we use clustering labels[4] to cluster clauses and entities. The experimental results demonstrate the effectiveness of our method. The contributions of this paper are summarized as follows:

- We propose a multi-granularity structured modeling method based on the hierarchical graph network that decomposes the EMR into sentences, clauses, and entities, and use clustering labels to enhance the expression of medical features.
- We adapt counterfactual intervention to extract critical supporting facts from the EMR during diagnosis. Importantly, our method is disease-independent and does not require a precise external medical knowledge base, so that it is suitable for a wide range of applications.
- The evaluation conducted on the real EMR dataset shows that our method can correctly diagnose the types of lymphedema. Keyword coverage and human evaluation show that the counterfactual reasoning method has better extraction accuracy and robustness compared to two existing methods reimplemented by ourselves.

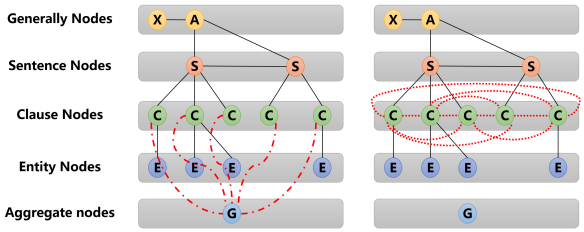

Figure 2: This figure shows the hierarchical connection structure between multi-grained nodes. The black edges in the graph represent the tree structure connection between the four types of nodes in the EMR. For the red edges, the left part shows the connection between the clause nodes and the graph aggregate nodes, and the right part shows the fully connected form between clause nodes.

2 Proposed Method

Given an irregular EMR in the form of free text $X = [x_1, x_2, \cdots, x_L]$ with L words, the task for us is to extract supporting facts that can be used to explain the diagnosis result without relying on external knowledge while performing diagnosis. The supporting facts can be entities or clauses of text.

2.1 Multi-Granularity Graph Construction

The medical features in the EMR are sparse and medical entities are insufficient to provide sufficient explanation for diagnosis. Therefore, we do multi-granularity segmentation for EMRs, which

[4]Notice that this label is disease-free and can be initially labeled without expert knowledge by crowdsourcing annotation.

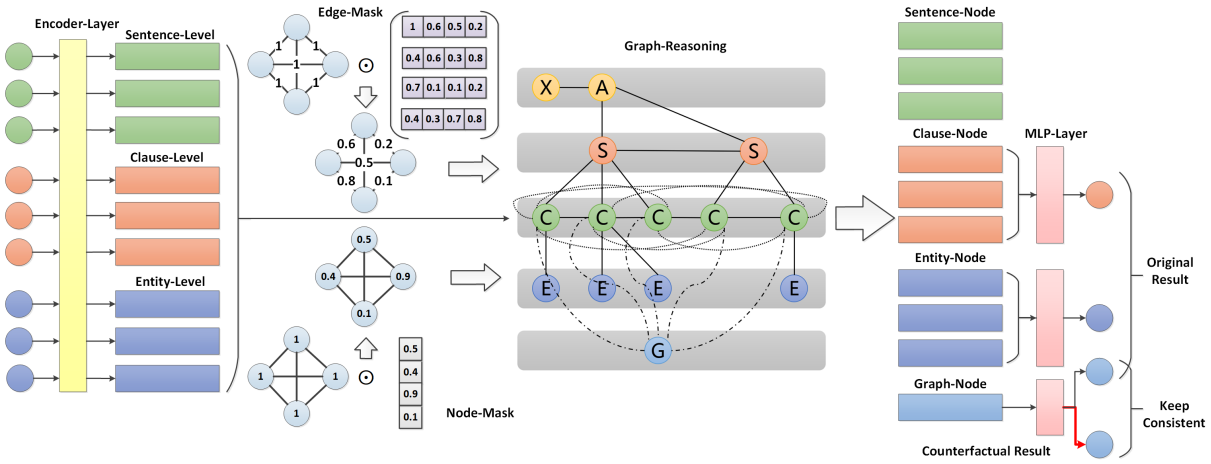

Figure 3: An overview of counterfactual multi-granularity graph supporting facts extraction network. To show the soft-mask process clearly, we assume that the features of both nodes and edges in the graph are 1, and $\odot$ denotes element-wise multiplication between graph features and mask matrix. All the edges in the graph are bidirectional. For clear reading, we only mark the monodirectional mask value for the bidirectional edges in the Edge-Mask.

enhances the symptom features of entities and explanation of diagnosis, while maintaining the integrity of the text. An EMR can be divided by periods into sentences, which can be further divided into clauses by commas or semicolons as a more granular segmentation. In order to keep the symptom features of entities, we do Named Entity Recognition[5] and number extraction for each clause[6]. In addition, we add two general nodes representing the gender and age of the patient respectively.

After segmentation, as shown in Figure 2, we can build a hierarchical tree structure. The nodes at each level represent the text of sentences, clauses, and entities respectively. Specifically, for each EMR, we connect the two general nodes, sentence nodes sequentially. Then, we connect the clause node to the sentence node to which it belongs and the entity nodes disassembled from it. In particular, a fully-connected relationship is established between all the clause nodes, which overcomes the defect that Graph Attention Network (GAT) can only aggregate the information from adjacent nodes when the network is shallow and expands the receptive field of each sub-sentence node to the whole EMR. Then, all clause nodes are connected to an aggregate node which is used to do the diagnosis. All the edges in the graph are bidirectional to make the information between nodes flow better.

[5]https://github.com/daiyizheng123/Bert-BiLSTM-CRF-pytorch

[6]We recommend Stanza (Qi et al., 2020; Zhang et al., 2020) for English EMR. https://github.com/stanfordnlp/stanza

2.2 Clustering labels

In the original EMR, all tokens have the same weight, so noisy text will degrade the performance of diagnosis and explanation. To improve the accuracy of symptom presentation, clustering-labels are used to cluster clauses and entities into corresponding medical classifications. Specifically, the clause is divided into 33 classes and the entity is divide into 10 classes, which is a scientific classification method in medicine derived from the textbook "Diagnostics" (Xuehong Wan, 2013). These labels are disease-free and can be labeled without expert knowledge by crowdsourcing annotation. We manually annotated the corresponding labels for the entire dataset on our own platform. And we have trained a BERT (Devlin et al., 2019) based text classifier on 30% of the data, which can achieve the annotation accuracy of 80.76% on clauses and 97.13% on entities on the remaining data. This shows that our method can easily annotate large-scale data. With these labels, we can gather the same types of features together in the feature space, thereby enhancing the model's overall attention to important types of features. Please refer to Appendix B.2 for more details.

2.3 Input Encoder

After building the multi-granularity graph for a medical record, each node in the graph contains a sequence $X_{node} = [x_1, x_2, \cdots, x_n]$ with n words, which is tokenized by the tokenizer of BERT (Devlin et al., 2019). In order to maintain the con-

sistency of the results of different granularity encoding, we use one bi-directional RNN (Schuster and Paliwal, 1997) with GRU (Cho et al., 2014) to cover the sequence of sentences, clauses, entities and general information into hidden state sequence respectively $\boldsymbol{H}_m = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n)$:

$$\boldsymbol{h}_t = \boldsymbol{BiGRU}(\boldsymbol{h}_{t-1}, \boldsymbol{e}(x_t)) \quad (1)$$

where $\boldsymbol{h}_t$ is the hidden state of the t-th token and $\boldsymbol{e}(x_t)$ is the embedding vectors with random initialization of x_i. Finally, we use the last hidden state of i-th text sequence as the feature $\boldsymbol{H}_i$ of $node_i$.

2.4 Graph Reasoning

Once we get the feature of the node, we use the Graph Attention Network (GAT) (Velickovic et al., 2018) to aggregate the information between different granularity. GAT can obtain the correlation score between nodes based on the attention mechanism, which is the key to the interpretability of our model. Specifically, GAT takes all the node features as input and calculates the attention coefficients α_{ij} by

$$e_{ij} = LeakyReLU(\boldsymbol{a}^T([\boldsymbol{W}\boldsymbol{H_i}; \boldsymbol{W}\boldsymbol{H_j}])) \quad (2)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})} \quad (3)$$

where $\boldsymbol{H_i}$ is the feature of node i, $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ is a learnable weight matrix for the linear projection, $\boldsymbol{a} \in \mathbb{R}^{2d}$ is a learnable weight vector used to transform the adjacent node feature representations to the edge score e_{ij} between the i-th and j-th nodes. Equation (4) means to do a softmax normalization between all the edge attention scores on the edges connected to node i. Then, we update the feature of each node by

$$\boldsymbol{H_i'} = LeakyReLU(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \boldsymbol{W}\boldsymbol{H_j}) \quad (4)$$

After graph reasoning, the representation $\boldsymbol{H}$ of each node has been updated with the granular information aggregated from adjacent nodes and can be used for subsequent tasks.

2.5 Multi-task Prediction

After obtaining the updated node features, we use them in three subtasks: (i) graph classification for automatic diagnosis; (ii) sub-sentence classification for clustering; and (iii) entity classification for clustering.

Taking entity node classification as an example, for each entity node, we use a two-layer MLP with the ReLU activation function to calculate the probability. For an entity node i, we can get

$$P_{entity,i} = MLP_{entity}(\boldsymbol{E_i}) \quad (5)$$

By the same way, we can obtain the probability P_{graph}, P_{clause}, P_{entity}. The same as the common multi-task learning, we joint all the losses together as:

$$\mathcal{L}_{joint} = \lambda_1 \mathcal{L}_{graph} + \lambda_2 \mathcal{L}_{clause} + \lambda_3 \mathcal{L}_{entity} \quad (6)$$

where λ_1, λ_2 and λ_3 are hyper-parameters, and all the loss are calculated by cross-entropy loss.

2.6 Counterfactual Reasoning on Graph

Providing supporting information while making the diagnosis is the key to applying Artificial Intelligence into the medical field. Inspired by (Ying et al., 2019), we add node-mask or edge-mask into GAT to obtain the counterfactual result after the training and eliminate the noise nodes while keeping the diagnostic results unchanged.

For edge-mask, we introduce a learnable matrix $\boldsymbol{M}$ with the same form as the adjacency matrix of the medical record graph. Each element m_{ij} in the matrix represents the degree of mask for message aggregation from node i to node j in the graph. With this method, the calculation of attention coefficients in the GAT has been changed to

$$\alpha_{ij} = \frac{\exp(e_{ij} m_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik} m_{ik})} \quad (7)$$

And for node-mask, similarly, we introduce a learnable parameter β_i for each node i in the graph. The parameter represents the degree of mask for the feature in the node. After node-mask, the calculation of e_{ij} and $\boldsymbol{H_i'}$ has been changed to

$$e_{ij} = LeakyReLU(a^T([\beta_i \boldsymbol{W}\boldsymbol{H_i}; \beta_j \boldsymbol{W}\boldsymbol{H_j}])) \quad (8)$$

$$\boldsymbol{H_i'} = LeakyReLU(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \beta_j \boldsymbol{W}\boldsymbol{H_j}) \quad (9)$$

In the training of counterfactual reasoning, we jointly optimize three loss functions to obtain accurate counterfactual results. To ensure that the model can make a correct diagnosis after the counterfactual intervention, we use the original model

type	train	test
Secondary Lymphedema	448	36
Primary Lymphedema	185	22
Chylous Reflux Lymphedema	19	21
Others	248	21
All	900	100

Table 1: The statistics of the datasets.

to obtain the fact result D_i and maximizes the probability of selecting the correct diagnosis in counterfactual reasoning. Besides, we minimize the sum of all elements in the mask matrix to ensure that all noise nodes are filtered as much as possible. Since there is an exponential level possibility of counterfactual intervention on the model through the node-mask or edge-mask, we minimize the information entropy of the mask matrix regarding which node to select to reduce the uncertainty of the result. Finally, the loss of counterfactual reasoning is as follows:

$$\begin{aligned}\mathcal{L}_c = &-\lambda_4 log P(D = D_i) + \lambda_5 sum(\boldsymbol{M}) \\ &-\lambda_6 \frac{1}{N} \sum_{m_i \in \boldsymbol{M}} m_i log(m_i) \\ &-\lambda_6 \frac{1}{N} \sum_{m_i \in \boldsymbol{M}} (1 - m_i) log(1 - m_i)\end{aligned} \tag{10}$$

where λ_4, λ_5 and λ_6 are hyper-parameters, N is the number of elements in the mask matrix $\boldsymbol{M}$, and all the elements in $\boldsymbol{M}$ are mapped to the $[0, 1]$ by sigmoid function. For node-mask, the training is similar.

After counterfactual reasoning, we extract the nodes or edges (each edge represents the two nodes connected) represented by the top-k elements in the mask matrix as supporting facts.

3 Experiment

3.1 Experimental Setting

Based on the cooperation with the hospitals, we conducted experiments with real EMR data. We selected the EMRs from the department of lymphedema and diagnose the disease of primary lymphedema (原发性淋巴水肿), secondary lymphedema (继发性淋巴水肿), chylous reflux lymphedema (乳糜返流性淋巴水肿) and others (其他). The reasons for us to choose this department are as follows: (I) Lymphedema is a sub-discipline in medicine, so the researches on it, whether in Medicine or Artificial Intelligence, is still limited. For example, ICD10 can not provide full medical supporting. (II) The pathogenesis and treatment methods of different types of lymphedema vary greatly, but their outward manifestations are similar. Therefore, there is an urgent need for a simple method of earlier diagnosis system of lymphedema. (III) Specialist doctors pay more attention to the diagnosis in sub-discipline disease and do not concern with the large-scale rough diagnosis.

Formally, there are 1000 EMRs used in our experiment, of which 900 are used for training and 100 are used for testing. The statistics of four types of diseases are shown in Table 1. The average length of all EMRs is 345 words in Chinese. And our model is implemented based on PyTorch (Paszke et al., 2019), and use Adam (Kingma and Ba, 2015) optimizer for training. Please refer to Appendix B.1 for datasets details and Appendix A.1 for implementation details.

3.2 Baseline

We designed two representative models to compare the ability to extract medical support facts under similar task conditions based on attention and variational inference:

Self-Attention This method represents most of the existing approaches and provides explanations through attention similarity. We use BiGRU to encode the EMR. With the sequence embedding, following (Choi et al., 2016), we use average pooling to obtain the overall representation for automatic diagnosis. For supporting fact extraction, following (Mullenbach et al., 2018), we calculate the self-attention weight of each token, and design a sliding window method to obtain the average attention scores of fixed length spans, among which having high scores are taken as the supporting facts.

PostKS This is another method based on variational inference we've designed in addition to attention. Inspired by the dialogue knowledge selection model PostKS (Lian et al., 2019), we convert the pivotal information extraction into a clause selection problem. This method uses the text result of the diagnosis(as shown in Figure 1) to calculate the correlation with the clause as posterior distribution through the attention mechanism, and then uses self-attention and average pooling between clauses to obtain correlation score as the prior distribution. During training, based on variational inference, the model uses posterior information to guide prior selection, so that makes the prior distribution and the

Model	Diagnosis			Clause			Entity		
	P/%	R/%	F1/%	P/%	R/%	F1/%	P/%	R/%	F1/%
Self-Attention	94.95	95.00	94.97	-	-	-	-	-	-
PostKS	95.13	97.00	96.06	-	-	-	-	-	-
CMGE$_{-c-e}$	96.17	96.00	96.08	1.91	3.72	2.52	14.36	9.05	11.10
CMGE$_{-e}$	97.40	96.00	96.69	81.26	81.80	81.53	15.22	32.18	20.67
CMGE$_{-c}$	97.19	97.11	97.15	25.75	1.55	2.92	95.33	95.12	95.22
CMGE	**99.04**	**99.00**	**99.02**	**82.49**	**82.53**	**82.51**	**96.43**	**96.38**	**96.40**

Table 2: The first two lines are the diagnostic performance of the compared model and the last line is ours. The middle three rows are ablation experiments. CMGE$_{-c-e}$ represents the model only use diagnosis label, CMGE$_{-e}$ represents the model without entity labels, and CMGE$_{-c}$ represents the model without clause labels.

posterior distribution consistent. Finally, during inference, we select the clauses with high prior attention scores among clauses as supporting facts. Please refer to Appendix A.2 for more details.

3.3 Evaluation Metrics

To measure the performance of our pivotal information extraction module, we built a simple diagnostic criterion from (Levine, 2017), which is a complete diagnosis and treatment guide for lymphedema written by medical experts. Based on this diagnosis criteria, we used a combination of automatic evaluation and human evaluation.

Automatic Evaluation The precision, recall, and F1 are used as the metrics to measure the diagnostic accuracy of the model, which is the basis for the practical application. Specifically, several key-phrases for the three types of lymphedema are manually identified respectively to represent diagnostic features, and they are the re-descriptions of diagnostic criteria in the guide using phrases from EMRs. We use hit@1/3/5 (Bordes et al., 2013) to measure the coverage rate of the extracted results to the key-phrases. These metrics represent whether one of the diagnostic features is included in the top-1/3/5 extracted results. Please refer to Appendix B.3 for more details.

Human Evaluation Since some of the implicit medical features cannot be covered by key-phrases, human evaluation is necessary. We used each model to extract the top 3 supporting facts respectively for all 100 EMR samples in the testset, and randomly shuffled the order of the results. Then we invited 3 evaluators with medical backgrounds and having read the guide to determine whether the results conform to medical knowledge. We focus on the comprehensiveness and trustworthiness of each model. Comprehensiveness is used to measure whether the model can provide more medical features, and trustworthiness is used to measure whether the extraction results are helpful for diagnosis. For each item, the evaluator is asked to score in $0 \sim 2$. The final indicator is the average of the three evaluators.

4 Results and Analysis

4.1 Diagnostic Result

The diagnostic results are shown in Table 2. From the results, we can see our model performs better than all the compared models and can achieve about 99% accuracy in the diagnosis of lymphedema, which exceeds the comparison models by 3%-5% in precision, recall, and F1. Based on our model, the categories of clauses and entities can be distinguished correctly, which demonstrates that the clustering information contained in the pseudo-labels is correctly learned by our multi-granularity model. This result indicates that the accuracy of our method in the diagnosis of lymphedema is in line with clinical requirements. Since our goal is to make the model really help doctors in clinical practice with reliable medical explanations, we will focus on the performance of the counterfactual extraction of the supporting facts for the diagnosis that follows. Please refer to Appendix A.4 for the effectiveness of our model in diagnosis on the benchmark data.

4.2 Counterfactual Extraction Result

Automatic Result Table 3 shows the automatic evaluation results of the supporting facts extraction. Since the identified keywords are difficult to accurately cover the features for diagnosis and models have different adaptability to various diseases, the performance is distinguishing on different diseases. Compared with other models, the counterfactual-

Model	Secondary Lymphedema			Primary Lymphedema			Chylous Reflux Lymphedema		
	hit@1	hit@3	hit@5	hit@1	hit@3	hit@5	hit@1	hit@3	hit@5
Self-Attention	5.45%	36.36%	50.91%	13.64%	45.45%	**54.55%**	4.76%	9.52%	33.33%
PostKS	9.09%	43.64%	60.00%	9.09%	**40.91%**	**54.55%**	0.00%	14.29%	19.05%
Node-Mask	25.45%	52.73%	69.09%	**22.73%**	31.82%	**54.55%**	9.52%	19.05%	23.81%
Edge-Mask	**36.36%**	**61.82%**	**70.91%**	**22.73%**	**40.91%**	50.00%	**61.90%**	**66.67%**	**76.19%**

Table 3: The automatic evaluation for the extraction of diagnostic supporting facts for three types of lymphedema.

Model	Secondary Lymphedema		Primary Lymphedema		Chylous Reflux Lymphedema		Others	
	C	T	C	T	C	T	C	T
Self-Attention	0.85	0.83	0.75	0.95	0.62	0.33	1.09	0.76
PostKS	0.64	0.96	0.66	0.82	0.67	0.48	0.57	0.42
Node-Mask	1.44	0.91	1.11	1.02	1	0.71	1.67	1.24
Edge-Mask	1.67	1.22	1.36	1.11	1.33	1.38	1.67	1.52

Table 4: The human evaluation for the extraction of diagnostic supporting facts for each type. In the table, C stands for comprehensiveness and T stands for trustworthiness.

based methods, especially the Edge-Mask method, has an advantage in accuracy and robustness on the whole. Hit@1 shows that the Edge-Mask can locate key facts more quickly than the comparison methods and hit@5 shows that it achieves over 70% accuracy on secondary lymphedema and chylous reflux lymphedema. In the comparison of different lymphedema, other methods have a greater performance degradation, and only the Edge-Mask maintains high accuracy in various diseases, indicating that the Edge-Mask method is highly robust to different diseases.

Human Result Table 4 shows the results of the human evaluation of the four categories of diagnosis. Compared with other methods, the counterfactual-based methods have great advantages in comprehensiveness, which indicates that our method can focus more on useful medical information and eliminate invalid noise in the EMR. The fourth category requires focus. This category includes all non-lymphedema medical records, and its diseases are diverse and complex. It can be seen that the method of counterfactual reasoning has strong performance in this type in terms of comprehensiveness and credibility, indicating that our method is truly independent of the type of disease and suitable for large-scale promotion.

4.3 Effectiveness of Clustering Labels

Table 2 shows the ablation experiment results for the clustering labels. For the experiment without corresponding labels, we used a classifier with random initialization parameters for classification, which can reflect the expectation of the ability to encode medical features of the model. The results show that both the clause label and entity label can improve the accuracy of diagnosis by about 1% on the basis of over 96% accuracy. Since we use the same encoder to encode the three granular texts of the sentence, clause and entity, the addition of clause labels also improves the accuracy of entity classification and vice versa. The result indicates that the introduction of cluster tags enhances the expression of medical information in the model and enables the model to better extract and utilize relevant medical knowledge from irregular text.

4.4 Advanced Analysis

Results in Primary Lymphedema Since the diagnosis of primary lymphedema is mainly diagnosed by excluding other types of lymphedema, the keywords we established are not standardized in the EMR, the performance of all models in Table 3 has a significant decline and only be used for comparison. And the performance in human evaluation is consistent with other diseases in Table 4.

Results in Chylous Reflux Lymphedema Except for Edge-Mask, the performance of the other methods on chylous reflux lymphedema has dropped significantly. Since this type of EMR only accounts for 4% of the dataset, the models based on frequency statistics are difficult to capture key features. And Edge-Mask, using counterfactual intervention to obtain causal relation, is disease-independent and can adapt to few data.

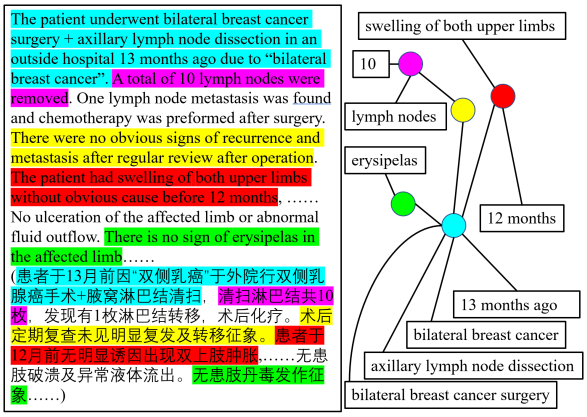

Figure 4: An example of a trustworthy supporting facts graph extracted by Edge-Mask. The keywords "cancer", "surgery" and "chemotherapy" in this result meet the diagnostic criteria for secondary lymphedema in Appendix B.3.

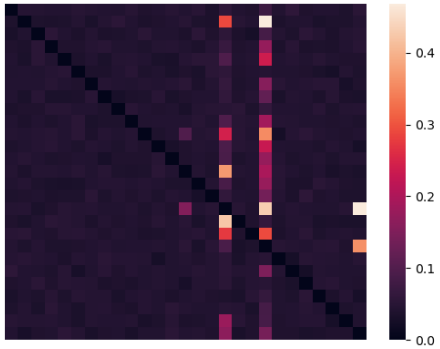

Figure 5: An example of visualization of Edge-Mask matrix. The same as the adjacency matrix, the rows and columns in the figure correspond to the nodes in the graph, and each grid in the figure represents a value in the Edge-Mask matrix.

Node-Mask and Edge-Mask Edge-Mask is included in Node-Mask. Masking the feature of a node will inevitably reduce the flow of information on all connected edges. So compared to Node-Mask, Edge-Mask is a fine-grained counterfactual intervention. For Node-Mask, the flow of multi-granularity information between nodes will be truncated. For example, when a clause node is masked, the entity features belonging to it are truncated together. Therefore, Node-Mask has a weaker performance than Edge-Mask.

4.5 Visual Presentation of Results

Figure 4 is an example randomly obtained from the test set. In this graph, each node represents a clause that contains the entities used to describe the symptoms of the disease and the edges represent the connection between them. All the aforementioned features constitute a hierarchical supporting graph to provide effective help for doctors' diagnosis. As we can see, our model successfully extracted the patient's history of cancer, surgery and chemotherapy, which can clearly indicate that the patient is suffering from secondary lymphedema. This shows that the supporting facts we extracted are effective. We provide a comparison of the extraction results of different models in Appendix A.3.

Figure 5 shows an example of the visualization of the Edge-Mask matrix. It can be seen that most of the edges have been masked, and only the edges from two key feature nodes have been preserved. This proves that our method can effectively filter noisy features and extract supporting facts.

5 Related Works

Explainable Diagnosis with EMR It is necessary to provide explainability for automatic diagnosis systems. CAML (Mullenbach et al., 2018) provides explanations with the spans having the high attention weights in the text sequence and (Feng et al., 2020) calculates a threshold for attention selection. AdaCare (Ma et al., 2020a) calculate the average importance weights in the overall dataset to obtain symptoms strongly associated with the diseases. These works focus on correlations based on attention and ignore causality between features and diagnosis.

Document Modeling with Graph Network Document modeling with graph network has been widely used in text classification (Yao et al., 2019), multi-hop reading comprehension (Cao et al., 2019) and abstract extraction (Wang et al., 2020). An EMR can also be considered as a document. There are two main ways to structure a document into a graph, based on the entity (Qiu et al., 2019) or based on the structure of the document (Zheng et al., 2020). (Tu et al., 2019) considers the integration of documents and entities as heterogeneous nodes in the graph network, and (Fang et al., 2019) propose a hierarchical model that combines document structure and entity structure. We used a multi-granularity hierarchical graph network to model the EMR documents.

Counterfactual Reasoning Providing explanations based on counterfactual reasoning has a long

history (Lewis, 1973; Woodward, 2005). In recent years, (Oberst and Sontag, 2019) introduces a kind of structural causal model to genera counterfactual trajectories in a synthetic environment of sepsis management. (Lin et al., 2020) presents a patient simulator to generate informative counterfactual response in the disease diagnosis. (Lenis et al., 2020) identifies salient regions of a medical image by measuring the effect of local counterfactual image-perturbations. We use counterfactual reasoning in EMRs to provide explanations for diagnosis.

6 Conclusion

In this paper, we propose a counterfactual multi-granularity graph supporting facts extraction (CMGE) method for the irregular EMR without an external medical knowledge base. Based on this model, we can correctly diagnose lymphedema. The proposed counterfactual-based approach can discover the causal relationship between symptoms and diagnosis. The results of supporting fact extraction show that our method has strong robustness and can maintain accuracy in various diseases and even in categories with few data resources. In the future, we will introduce multi-modal into the model such as radiology images to discover more medical knowledge from EMRs.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No.2018YFB1005104) and the Key Research Program of the Chinese Academy of Sciences (ZDBS-SSW-JSC006).

References

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. Hypercore: Hyperbolic and co-graph representation for automatic ICD coding. In *ACL*, pages 3105–3114. Association for Computational Linguistics.

Yu Cao, Meng Fang, and Dacheng Tao. 2019. BAG: bi-directional attention entity graph convolutional network for multi-hop reasoning question answering. In *NAACL-HLT (1)*, pages 357–362. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734. ACL.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. Doctor AI: predicting clinical events via recurrent neural networks. In *MLHC*, volume 56 of *JMLR Workshop and Conference Proceedings*, pages 301–318. JMLR.org.

Edward Choi, Cao Xiao, Walter F. Stewart, and Jimeng Sun. 2018. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In *NeurIPS*, pages 4552–4562.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. 2020. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *CoRR*, abs/2010.15728.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2019. Hierarchical graph network for multi-hop question answering. *CoRR*, abs/1911.03631.

Jinyue Feng, Chantal Shaib, and Frank Rudzicz. 2020. Explainable clinical decision support from text. In *EMNLP (1)*, pages 1478–1489. Association for Computational Linguistics.

Ivan Girardi, Pengfei Ji, An-phi Nguyen, Nora Hollenstein, Adam Ivankay, Lorenz Kuhn, Chiara Marchiori, and Ce Zhang. 2018. Patient risk assessment and warning symptom detection using deep attention-based neural networks. In *Louhi@EMNLP*, pages 139–148. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Dimitrios Lenis, David Major, Maria Wimmer, Astrid Berg, Gert Sluiter, and Katja Bühler. 2020. Domain aware medical image classifier interpretation by counterfactual impact analysis. In *MICCAI (1)*, volume 12261 of *Lecture Notes in Computer Science*, pages 315–325. Springer.

S. M. Levine. 2017. Lymphedema: Complete medical and surgical management. *Plastic & Reconstructive Surgery*, 139(3):772.

David Lewis. 1973. Counterfactuals, blackwells.

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *AAAI*, pages 8180–8187. AAAI Press.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *IJCAI*, pages 5081–5087. ijcai.org.

Junfan Lin, Ziliang Chen, Xiaodan Liang, Keze Wang, and Liang Lin. 2020. Learning reinforced agents with counterfactual simulation for medical automatic diagnosis. *CoRR*, abs/2003.06534.

Luchen Liu, Haoran Li, Zhiting Hu, Haoran Shi, Zichang Wang, Jian Tang, and Ming Zhang. 2019. Learning hierarchical representations of electronic health records for clinical outcome prediction. In *AMIA*. AMIA.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. KAME: knowledge-based attention model for diagnosis prediction in healthcare. In *CIKM*, pages 743–752. ACM.

Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020a. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *AAAI*, pages 825–832. AAAI Press.

Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. 2020b. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *AAAI*, pages 833–840. AAAI Press.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *NAACL-HLT*, pages 1101–1111. Association for Computational Linguistics.

Michael Oberst and David A. Sontag. 2019. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 4881–4890. PMLR.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *ACL (1)*, pages 6140–6150. Association for Computational Linguistics.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45(11):2673–2681.

Ying Sha and May D. Wang. 2017. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In *BCB*, pages 233–240. ACM.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *ACL (1)*, pages 2704–2713. Association for Computational Linguistics.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR (Poster)*. OpenReview.net.

Sahil Verma, John P. Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *CoRR*, abs/2010.10596.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *ACL*, pages 6209–6219. Association for Computational Linguistics.

James Woodward. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.

Keyang Xu, Mike Lam, Jingzhi Pang, Xin Gao, Charlotte Band, Piyush Mathur, Frank Papay, Ashish K. Khanna, Jacek B. Cywinski, Kamal Maheshwari, Pengtao Xie, and Eric P. Xing. 2019. Multimodal machine learning for automated icd coding. volume 106 of *Proceedings of Machine Learning Research*, pages 197–215, Ann Arbor, Michigan. PMLR.

Yuan Xu, Seungwon Lee, Elliot Martin, Adam G. D'souza, Chelsea T.A. Doktorchik, Jason Jiang, Sangmin Lee, Cathy A. Eastwood, Nowell Fine, Brenda Hemmelgarn, Kathryn Todd, and Hude Quan. 2020. Enhancing icd-code-based case definition for heart failure using electronic medical record data. *Journal of Cardiac Failure*, 26(7):610 – 617.

Xuefeng Lu Xuehong Wan. 2013. *Diagnostics*. Beijing: People's Health Publishing House.

Zhongliang Yang, Yongfeng Huang, Yiran Jiang, Yuxi Sun, Yu-Jin Zhang, and Pengcheng Luo. 2018. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Scientific reports*, 8(1):1–9.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *AAAI*, pages 7370–7377. AAAI Press.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, pages 9240–9251.

Quan Yuan, Jun Chen, Chao Lu, and Haifeng Huang. 2020. The graph-based mutual attentive network for automatic diagnosis. In *IJCAI*, pages 3393–3399. ijcai.org.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. 2020. Biomedical and clinical english model packages in the stanza python nlp library. *arXiv preprint arXiv:2007.14640*.

Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu. 2020. Document modeling with graph attention networks for multi-grained machine reading comprehension. In *ACL*, pages 6708–6718. Association for Computational Linguistics.

A Experimental Setting

A.1 Implementation Details

To implement our model, we use the tokenizer of BERT (Devlin et al., 2019) to obtain the tokens of the EMR text sequence. For BiGRU (Cho et al., 2014) encoder, the embedding dimension is 300 and the hidden dimension is 256 with two layers. In graph reasoning, we use 2 multi-heads GAT layers with 8 heads. The input dimension of GAT is 1024 and the output dimension is 128. For counterfactual reasoning, we fix the parameters of all the diagnostic models and only optimized the matrix of Edge-Mask or Node-Mask. The hyper-parameters can be set to any possible value based on the tuning. With the manual tuning for diagnostic accuracy, except for λ_5, all the hyper-parameters of the loss function are set to 1 in the experiment, and λ_5 is set to 0.1 for Node-Mask and 0.005 for Edge-Mask. We trained on the diagnostic model for 20 epochs and do counterfactual training on each sample for 200 epochs. Our model has a total of 16.7M parameters and can easily train and infer in Titan XP. Since we are not doing parallel processing, counterfactual reasoning is the most consuming, and it takes 7 seconds for each instance.

A.2 PostKS

We modified the PostKS (Lian et al., 2019) model to this task. In order to enhance the accuracy of supporting facts extraction, except the diagnosis label, we use some additional diagnostic descriptions related to the disease, which are shown in "Diagnosis" in Figure 1.

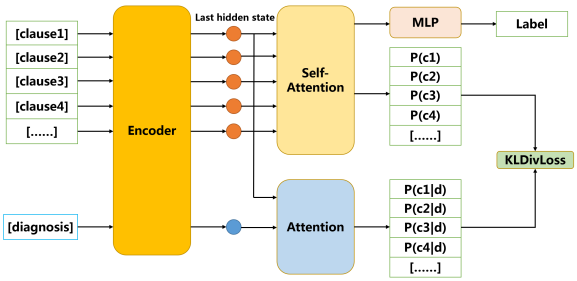

Figure 6: An overview of modified PostKS.

Figure 6 shows the overview of variational inference model. All the clauses and the diagnosis are encoded by BiGRU and we take the last hidden state $\boldsymbol{h}_n$ as the feature sequence $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_n]$ for the clauses and the feature $\boldsymbol{d}$ for diagnosis. Based on these, we can calculate the posterior distribution as:

$$p(\boldsymbol{c} = \boldsymbol{c}_i | \boldsymbol{C}, \boldsymbol{d}) = \frac{exp(\boldsymbol{c}_i \cdot \boldsymbol{d})}{\sum_{j=1}^{N} exp(\boldsymbol{c}_j \cdot \boldsymbol{d})} \quad (11)$$

where N is the number of clauses, c_i is the feature of the i-th clause, and for prior distribution, we calculate as:

$$p(\boldsymbol{c} = \boldsymbol{c}_i | \boldsymbol{c}_i, \boldsymbol{c}_k) = \frac{exp(\boldsymbol{c}_i \cdot \boldsymbol{c}_k)}{\sum_{j=1}^{N} exp(\boldsymbol{c}_j \cdot \boldsymbol{c}_k)} \quad (12)$$

Then use the average pooling to obtain the self-attention weight $p(\boldsymbol{c} = \boldsymbol{c}_i | \boldsymbol{C})$ of each clause and optimize:

$$\mathcal{L}_c = \lambda_d \mathcal{L}_{diagnosis} + \lambda_k \sum_{N}^{i=1} p(\boldsymbol{c} = \boldsymbol{c}_i | \boldsymbol{C}, \boldsymbol{d}) log \frac{p(\boldsymbol{c} = \boldsymbol{c}_i | \boldsymbol{C}, \boldsymbol{d})}{p(\boldsymbol{c} = \boldsymbol{c}_i | \boldsymbol{C})} \quad (13)$$

where λ_d, λ_k are hyper-parameters and $\mathcal{L}_{diagnosis}$ is the cross-entropy loss for diagnosis result.

A.3 Result Comparison

Table 7 shows two examples of supporting facts extraction results. For Secondary Lymphoma, it can be seen that except for PostKS, all other methods can find critical features. PostKS discovered the word "lymphedema" since it is highly correlated with the diagnosis text. The result indicates that

Model	AUC		F1		P@5
	Macro	Micro	Macro	Micro	
BiGRU (Mullenbach et al., 2018)	82.8	86.8	48.4	54.9	59.1
CNN (Mullenbach et al., 2018)	87.6	90.7	57.6	62.5	62.0
CAML (Mullenbach et al., 2018)	87.5	90.9	53.2	61.4	60.9
DR-CAML (Mullenbach et al., 2018)	88.4	91.6	57.6	63.3	61.8
MultiResCNN (Li and Yu, 2020)	89.9	92.8	60.6	67.0	64.1
HyperCore (Cao et al., 2020)	89.5	92.9	60.9	66.3	63.2
MultiResCNN*	89.3	92.2	59.3	66.2	62.8
BiGRU + MHG	88.9	92.5	57.8	66.6	64.1
MultiResCNN + MHG	**90.2**	**93.1**	**60.9**	**67.5**	**64.7**

Table 5: The experimental result on benchmark data. MultiResCNN* represents the result of MultiResCNN on our pre-processed hierarchical structured data. BiGRU, CNN, CAML and DR-CAML are the most frequently compared baselines for this task. MultiResCNN and HyperCore are currently strong and effective baselines.

the posterior information has worked, but it cannot provide an explanation for the diagnosis. The Edge-Mask discovered the "swelling after surgery immediately", which is the best support for the diagnosis of secondary lymphedema, indicating the effectiveness of it. For Chylous Reflux Lymphedema, only Self-Attention and Edge-Mask find critical information like "milky white liquid". Compared with Self-Attention, Edge-Mask has a more complete description of the supporting facts.

A.4 Evaluation on benchmark data

We didn't find any benchmarks on the task of directly extracting supporting facts from EMRs without other knowledge. To better prove the performance of our model, we have done experiments on the English EMR benchmark "MIMIC-III-50" (Mullenbach et al., 2018) for the task of assigning ICD codes to EMRs. This task assigns multiple codes to EMRs from 50 labels. Compared to our diagnosis of four types of EMRs, the difficulty is obvious.

The key module of our model in diagnosis is the multi-granularity hierarchical graph (MHG) document modeling method based on clauses and entities. In the experiments, we subsequently connect our multi-granularity hierarchical graph network module after BiGRU (Mullenbach et al., 2018) and MultiResCNN (Li and Yu, 2020) to further encode the EMRs. Since the clause categories are not labeled on this dataset, we only used the entity labels obtained by NER and do not constrain the clause node.

The result shown in Table 5 show that our module achieves effective performance improvements on all metrics based on MultiResCNN and BiGRU. With our module, BiGRU even surpasses MultiResCNN in some metrics, while they originally have a huge gap in performance. This experimental result proves the effectiveness of our model in diagnosis on the benchmark data.

B Data and Metrics Description

B.1 Data Collection

We collected data from the real historical electronic medical records (EMRs) of the department of lymphedema. It contains the patient's self-complaint, history of present illness, past illness, personal history, family history, physical examination, and specialist examination. In order to protect the privacy of patients, we have deleted all content related to personal information. For the experiment, we extracted three types of EMRs of primary lymphedema, secondary lymphedema, and chyle reflux lymphedema from all EMRs. In addition, we added 25% of the confounded EMRs which includes patients who were hospitalized in the department of lymphedema, but the final diagnosis was other diseases. The statistics of four diseases in the final dataset are shown in Table 1.

Although the EMR distinguishes information such as the history of present illness and past illness, since the content of each part is still irregular text, and most of the existing EMRs are not standardized, we treat the EMR as an unstructured text and connect all the pieces together. Since our EMRs contain a complete physical examination and life history, most of the symptomatic entities present are negative and unrelated to diagnosis, which introduces a lot of noise into diagnosis and explanation. This is also an important reason that we cannot use entities as supporting facts. We do not have permission from hospitals to publish the

Type	Label
Clause	null(无标签), nature of symptom(症状的性质),relation of symptom(症状之间的联系), cause of change in symptom(引起症状变化的因素), position of symptom(症状的部位), duration of symptom(症状持续时间), time of onset of symptom(症状出现的时间), degree of symptom(症状的程度), new symptoms(新出现的症状), change of symptom(症状的变化), cause of disease(起病原因), time of disease(患病时间), severity of disease(疾病的严重程度), method of examination(检查方法), result of examination(检查结果), method of treatment(治疗方法), location of treatment(治疗地点), cause of treatment(治疗原因), purpose of treatment(治疗目的), effect of treatment(治疗效果), doses of drug(药物剂量), sleep condition(睡眠情况), mental condition(精神情况), defecation and urination condition(大小便情况), weight condition(体重情况), appetite condition(食欲情况), negative information(阴性资料), description of number(数量描述), name of drug(药物名称), physical description(体力描述), time description(时间描述), general description(通用描述), others(其它),
Entity	position(部位), drug(药品), duration(时长), disease(疾病), hospital(医院), surgery(手术), number(数字), examination(检查), time(时间), others(其他)

Table 6: The detailed description of the cluster label.

Chinese EMR data since they are legally protected by the laws. So we can only provide two examples in Table 7 with extraction results of each model.

B.2 Clustering Label

The detailed description of the cluster label is shown in Table 6. They are derived from the textbook "Diagnostics" (Xuehong Wan, 2013). It's a scientific classification method in medicine. In our experimental data, we manually annotated the corresponding labels on our own platform. These labels are crude, disease-independent and there may be intersections between categories, because they are only used to cluster clauses or entities and do not require a high degree of accuracy. Therefore, we can easily annotate a small part of the dataset manually and train a text classifier to classify the remaining data based on BERT (Devlin et al., 2019). The experimental results show that the classifier we trained on 30% of the data can achieve the annotation accuracy of 80.76% on clauses and 97.13% on entities in the remaining data.

B.3 Diagnostic Criterion

In this section, we will briefly introduce the diagnostic criteria for three types of lymphedema. Different hospital or even departments have their own ways to describe the recognized diagnostic criteria, which will be reflected in the EMRs. So our diagnostic criteria are manually annotated by analyzing the EMRs and the diagnosis guide (Levine, 2017). They are the re-descriptions of diagnostic criteria in the guide using phrases from EMRs.

Secondary Lymphedema For secondary lymphedema, the most important diagnostic criterion is whether the patient's lymphatic vessels have been damaged. Therefore, if there are descriptions related to tumors, surgery, radiotherapy, etc. in the medical records, it is likely to be secondary lymphedema.

Primary Lymphedema For primary lymphedema, the main basis for diagnosis is whether the patient's lymphatic vessels have congenital dysplasia or edema. Since there are few descriptions of this basis in the medical records, we will also take "edema without an inducement many years ago(多年前无诱因出现水肿)" as the basis for correct extraction in the evaluation.

Chylous Reflux Lymphedema For chylous reflux lymphedema, the key to the diagnosis is whether the patient has chylous reflux. Therefore, if there are descriptions related to milky white fluid, effusion reflux, etc. in the medical record, it is roughly considered to be chylous reflux lymphedema.

EMR Sample	Extraction Result	
Age:46. **Gender**:woman. **Document**: The patient underwent radical mastectomy for right breast cancer at a local hospital 3 years ago. The regular postoperative review showed no signs of tumor recurrence. After the surgery, the right upper limb was swollen immediately, which is concave without pain, fever, paresthesia, and other symptoms, so the patient did not pay attention to it. The swelling gradually developed from the upper arm to the whole right upper arm, aggravated after activity, and decreased after rest. No symptoms of infection such as redness, swelling, heat, and pain of the affected limb, or fever of the whole body were observed. The patient was admitted to our hospital for further diagnosis and treatment. ETC in the outpatient department shows: no obvious abnormalities were found in the ultrasound of the upper limb vein; Right upper extremity magnetic resonance is consistent with lymphedema. The outpatient department is admitted with "lymphedema". The patient had good mental, appetite, sleep, urine, and feces since the onset of the disease, and had not lost weight recently.	patient had right breast cancer three years ago	Self-Attention
	no significant weight loss	
	lymphatic swollen; upper limb vein	
	can concavity	PostKS
	regular postoperative review	
	the outpatient department was admitted with "lymphedema"	
	the patient underwent radical mastectomy for right breast cancer at a local hospital 3 year ago	Node-Mask
	the swelling decreased after rest	
	there were no signs og tumor recurrence	
	swelling begins in the upper arm and progresses gradually throughout the right upper arm	Edge-Mask
	swelling of the right upper limb was present immediately after surgery	
Diagnosis: Secondary Lymphoma of Right Upper Limb	there were no signs of tumor recurrence	
Age: 15. **Gender**: man. **Document**: The patient developed multiple cystic vesicle-like structures in the right thigh 6 years ago without obvious inducement. After standing and walking for a long time, the lesions could be ruptured, leaving a milky white fluid. Since then, the patient appeared edema in the right thigh, hip, right waist, scrotal. The swelling gradually aggravated, and gradually developed from thigh to calf. The swelling was concave, first appearing in the thigh and then gradually descended to the lower leg. The swelling of the affected limb was significantly increased after standing and walking for a long time, and the swelling could be significantly alleviated after lying down and raising the affected limb. No change in skin color of the affected limb, no sensory and motor disturbance of affected limb, milky white or clear fluid may flow out after skin rupture. Self-report shows that there was no obvious relation between swelling and diet. The lower extremity vascular ultrasound examination showed no definite abnormality in a local hospital. For further diagnosis and treatment, he was admitted to our hospital.	milky white or clear liquid to flow out	Self-Attention
	no sensory and motor impairments were observed in affected limbs	
	sand and walk for long periods	
	came to our outpatient clinic for futher diagnosis and treatment	PostKS
	it can break down after standing and walking for a long time	
	the patient presented edema in right thigh, hip, right waist and scrotal	
	the swelling is getting worse	Node-Mask
	no skin color change of affected limb	
	treatment in a local hospital	
	and leave a milky white liquid	Edge-Mask
	after the skin ruptures may have the milky white or the clear liquid outflow	
Diagnosis: Chylous Reflux Lymphedema of Right Waist and Hip, Right Lower Limb, Scrotum	the swelling is concavity	

Table 7: The form of the EMR and the result of the extraction.