# Konwledge-Enabled Diagnosis Assistant Based on Obstetric EMRs and Knowledge Graph

**Kunli Zhang[1,2], Xu Zhao[1,2], Lei Zhuang[1], Qi Xie[1] and Hongying Zan[1,2]**
[1]School of Information Engineering, Zhengzhou University, Zhengzhou, China
[2]Peng Cheng Laboratory, Shenzhen, China
{ieklzhang,ielzhuang,ieqxie,iehyzan}@zzu.edu.cn
zhaox917@163.com

## Abstract

The obstetric **E**lectronic **M**edical **R**ecord (EMR) contains a large amount of medical data and health information. It plays a vital role in improving the quality of the diagnosis assistant service. In this paper, we treat the diagnosis assistant as a multi-label classification task and propose a **K**nowledge-**E**nabled **D**iagnosis **A**ssistant (KEDA) model for the obstetric diagnosis assistant. We utilize the numerical information in EMRs and the external knowledge from Chinese Obstetric Knowledge Graph (COKG) to enhance the text representation of EMRs. Specifically, the bidirectional maximum matching method and similarity-based approach are used to obtain the entities set contained in EMRs and linked to the COKG. The final knowledge representation is obtained by a weight-based disease prediction algorithm, and it is fused with the text representation through a linear weighting method. Experiment results show that our approach can bring about +3.53 F1 score improvements upon the strong BERT baseline in the diagnosis assistant task.

## 1 Introduction

Health service relations on the health of millions of people, and it is a livelihood issue in our country. Specifically in China, which has a huge population, the total amount of medical resources is still insufficient. The imbalance between the supply and demand for medical services is still the focus of China's healthcare industry. Although the implementation of China's Universal Two-child Policy in 2016 achieved many benefits, it also leads to an increase in the proportion of older pregnant women and the incidence of various complications (Yang and Yang, 2016). Compared to the overall supply of the medical industry, the lack of obstetric medical resources is prominent.

Since the issue of the Basic Norms of Electronic Medical Records (Trial) (China's Ministry of Health, 2010) by the National Health and Family Planning Medical Affairs Commission in 2010, medical institutions have accumulated many obstetric Electronic Medical Records (EMRs). EMRs are detailed records of medical activities, dominated by the semi-structured or unstructured texts. There is a lot of medical knowledge and health information in EMRs, which is the core medical big data. The first course record in EMRs can be divided into the chief complaint, physical examination, auxiliary examination, admitting diagnosis, diagnostic basis, and treatment plan. In general, there is not a single diagnosis in the admitting diagnosis, it usually includes normal obstetric diagnosis, medical diagnosis, and complications. As a consequence, the diagnosis assistant task based on the Chinese obstetric EMRs can be treated as a multi-label text classification problem, in which the different diagnoses can be regarded as the variable labels. However, the doctor's diagnosis and treatment process are based on comprehensive clinical experience and knowledge in the medical field to make a diagnosis and formulate a corresponding treatment plan. At the same time, they can also explain the corresponding diagnosis basis to the patient in detail. Therefore, rich clinical experience and solid medical knowledge play a vital role in the diagnosis procedure. In order to simulate the diagnosis and treatment process of doctors, we need to introduce external knowledge that



Proceedings of the 19th China National Conference on Computational Linguistics, pages 1156-1166, Hainan, China, October 31 - Novermber 1, 2020.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

is not available in EMRs. The introduction of medical domain knowledge requires formal expression so that it can be easily used in the diagnosis assistant model. To solve this problem, we adopt the Chinese Obstetric Knowledge Graph (COKG)[0] to introduce external medical domain knowledge.

In this paper, we use the BERT (Bidirectional Encoder Representation from Transformers) (Devlin et al., 2019) to generate the text representation of EMRs. The numerical information in EMRs is also important for the diagnosis results, it is being used to enhance the text representation with the multi-head self-attention (Vaswani et al., 2017). For entity acquisition, we compare the bidirectional maximum matching method and the Bi-LSTM-CRF method respectively, and choose the former method to obtain the entity sets from EMRs. Then the entities are linked to the COKG by a similarity-based method. Due to the fact that the negative words in EMRs will have an impact on the semantics, we employ a negative factor to deal with the negative words in EMRs and propose a weight-based disease prediction algorithm to obtain the final knowledge representation. Finally, a linear weighting method is employed to fuse the text representation and knowledge representation. The experiments on the Obstetric First Course Record Dataset support the effectiveness of our approach.

The main contributions of this paper are summarized as follows:

- In this paper, we propose the KEDA (**K**nowledge-**E**nabled **D**iagnosis **A**ssistant) model to integrate external knowledge from COKG into diagnosis assistant task.
- A weight-based disease prediction algorithm named WBDP is used to limit the influence of negative words in EMRs and generate the final knowledge representation.

## 2 Related Work

In this paper, we treat the obstetric diagnosis assistant task as a multi-label classification problem. The multi-label classification in traditional machine learning is usually regarded as a binary classification problem or adjust the existing algorithm to adapt to the multi-label classification task (Zhang and Zhou, 2007; Zhang and Zhou, 2006; Read et al., 2011; Tsoumakas et al., 2010).

With the development and application of deep learning, CNN and RNN are widely used in multi-label text classification tasks. For example, Kurata G et al. (2016) use CNN-based word embedding to obtain the direct relationship of the labels. Chen et al. (2017) propose a model that combined CNNs and RNNs to represent the semantic information of the text, and modeling the high-order label association. Baker S and Korhonen A (2017) use row mapping to hide the layers that map to the label co-occurrence based on a CNN architecture to improve the model performance. Ma et al. (2018b) propose a multi-label classification algorithm based on cyclic neural networks for machine translation. Yang et al. (2018) propose a Sequence Generation Model (SGM) to solve the multi-label classification problem. In recent years, the pre-training technology has grown rapidly, ELMo (Peters et al., 2018), OpenAI GPT (Radford et al., 2018), and BERT (Devlin et al., 2019) model have achieved significant improvements in multiple natural language processing tasks. They can be applied to various tasks after fine-tuning. However, due to the little knowledge connection between specific and open domain, these models do not perform well on domain-specific tasks. One way to solve this problem is to pre-train the model on a specific domain, but it is time-consuming and computationally expensive for most users. The models in this way are like ERNIE (Sun et al., 2019), BERT-WWM (Cui et al., 2019), Span-BERT (Joshi et al., 2020), RoBERTa (Liu et al., 2019), XLNET (Yang et al., 2019b), and so on. Moreover, if we can integrate knowledge at the fine-tuning process, it may bring better results. Several studies integrate external knowledge into the model. Chen J et al. (2019) use BiLSTM to model the text and introduce external knowledge through C-ST attention and C-CS attention. Li M et al. (2020) use BiGRU to extract word features, and use a similar matrix based on convolutional neural network and self-entity and parent-entity attention to introduce knowledge graph information. Yang A et al. (2019a) use knowledge base embedding to enhance the output of BERT for machine reading comprehension.

In terms of the diagnosis assistant based on Chinese obstetric EMRs, Zhang et al. (2018) utilize four multi-label classification methods, backpropagation multi-label learning (BP-MLL), random k-labelsets

[0] http:/47.106.35.172:8088/

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1156-1166, Hainan, China, October 31 - Novermber 1, 2020.

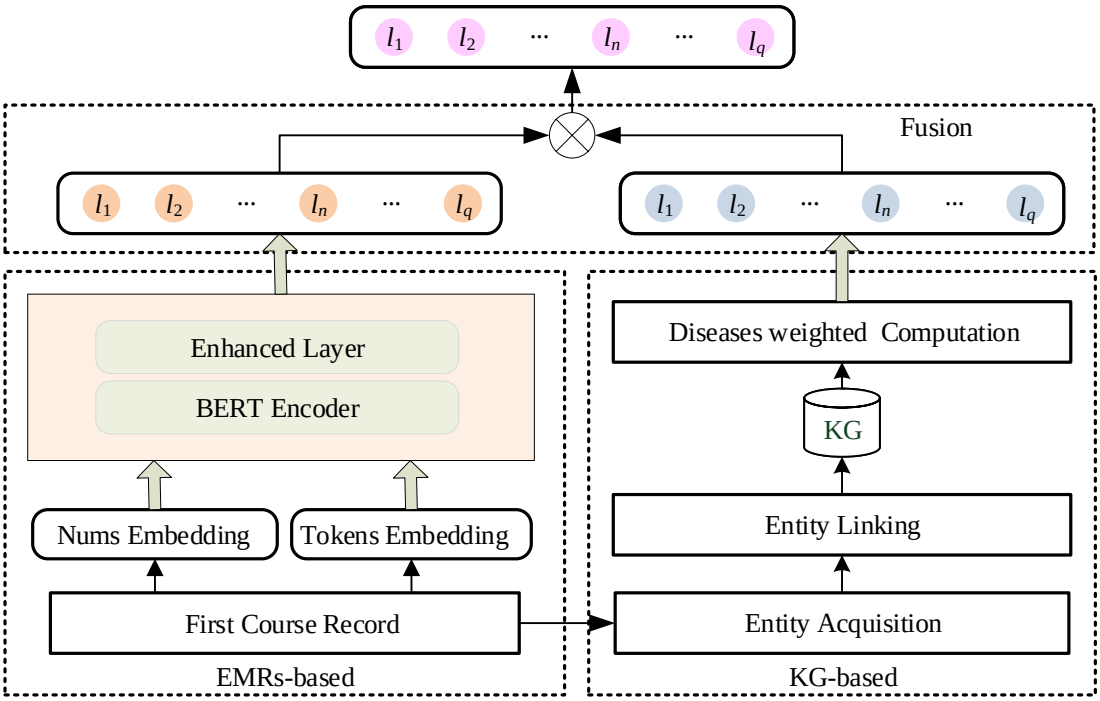


Figure 1: The architecture of the KEDA model

(RAkEL), multi-label k-nearest neighbor (MLKNN), and Classifier Chain (CC) to build the diagnosis assistant models. Ma et al. (2018a) fuse numerical features by employing the concatenated vector to improve the performance of the diagnosis assistant. Zhang et al. (2019) encode EMRs with BERT, and propose an enhanced layer to enhance the text representation for diagnosis assistant.

## 3 Methodology

### 3.1 Model Architecture

As shown in Figure 1, the KEDA model can be divided into three parts: EMRs-based module, KG-based module, and Fusion module. For any given EMR, the EMRs-based module generates the text representation by the BERT encoder firstly, then the numerical information contained in EMR is employed to enhance the text representation. Meanwhile, the KG-based module obtains the entities set and links to COKG through the entity acquisition and entity linking methods. Finally, the final knowledge representation is computed by a weight-based disease prediction algorithm and fused with the text representation through a linear weighting method. The following will introduce the implementation details of this model.

### 3.2 EMRs-based Module

The function of this module is to generate the text representation of EMRs. Similar to the BERT model, the input of KEDA model is composed of four parts: Token embedding, Position embedding, Segment embedding, and Nums embedding which contains the numerical information in EMRs.

**BERT encoder**

In this paper, we utilize the BERT as an encoder to obtain the text representation of EMRs.The input text sequence is as follows.

$$[CLS]ElectronicMedicalRecordText[SEP]$$

Where $[CLS]$ is a specific classifier token and $[SEP]$ is a sentence separator which is defined in BERT. For the diagnosis assistant task, the input of the model is a single sentence.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1156-1166, Hainan, China, October 31 - Novermber 1, 2020.

**Enhanced Layer**

The enhanced layer aims to enhance the text representation obtained by the BERT encoder through the numerical information in EMRs. Since the maximum length of the input sequence of BERT is 512, and the average length of EMRs is about 790 characters, we need to reduce the length of the input sequence. The information contained in the EMRs text can be divided into textual information and numerical information. Numerical information usually includes certain examinations or indications characterized by numerical values(For example, it contains the age, body temperature, pulse, respiration, respiration, and so on), which is also important information for diagnosis. So we separately extract the numerical information in EMRs to enhance the textual information, which not only can meet the limit of the input length, but also can better use the numerical information in the EMRs for diagnosis.

Then we adopt a multi-head self-attention proposed in Transformer (Vaswani et al., 2017) to integrate the numerical information into text representation of EMRs, as shown in Equation (1)-(4).

$$Q = K = V = W^S Concat([C]; Num_{1...M}) \tag{1}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

$$[C^{'}] = Concat(head_i, ..., head_h)W^O \tag{4}$$

Where $[C]$ is the hidden layer state representation of [CLS], $[C^{'}]$ is the text representation after fusing numerical information. $Num_{1...M}$ is the Nums embedding containing $M$ values, which is obtained by standardizing and normalizing the numerical information in EMRs. $W^S$, $W^Q$, $W^K$, $W^V$, and $W^O$ are trainable parameters, where $Q \in d^{model}$.

### 3.3 KG-based Module

**Entity Acquisition**

Through the analysis of obstetric EMRs, we found that the entities such as symptoms, signs, and diseases in EMRs are high-value information for the intelligent diagnosis, so we mainly identify these entities contained in EMRs.

To achieve better performance, we compared two ways for entity acquisition. One way is a dictionary-based method, the Chinese Symptom Knowledge Base(CSKB)[1], diseases set in ICD-10, and the entity sets of diseases and symptoms in COKG are used as dictionaries. We utilize the bidirectional maximum matching algorithm used in Chinese word segmentation (Gai et al., 2014) for entity acquisition, the obtained set includes a total of 9,836 entities. Another way is to use the Bi-LSTM-CRF model for entity acquisition, the texts labeled when constructing COKG is used as the training corpus. The Detailed analysis of experimental comparison results can be found in section 4.

**Entity Linking**

For the entity sets obtained above, it is necessary to establish a link relationship with the nodes in the knowledge graph. In this paper, the similarity-based approach is used to link the entities in the knowledge graph.

For a given identified entity $E_R$, we need to find the $n$ entities that are most similar to the knowledge graph COKG, the set of candidate entities is denote as $S = \{E_{K_1}, E_{K_2}, ..., E_{K_i}, ..., E_{K_n}\}$. Then we calculate the similarity between entities r and k, and select the entity with the highest similarity as the entity linked to COKG. The Levenshtein distance, Jaccard coefficient and the longest common substring are used to calculate the similarity respectively, as shown in Equation (5)-(7).

$$Sim_{ld} = \frac{levE_R, E_{K_i}(|E_R|, |E_{K_i}|)}{max(|E_R|, |E_{K_i}|)} \tag{5}$$

[1] http://www5.zzu.edu.cn/nlp/info/1015/1865.htm

$$Sim_{jacc} = jaccard(bigram(|E_R|), bigram(|E_{K_i}|)) \tag{6}$$

$$Sim_{lcs} = \frac{|lcs(E_R, E_{K_i})|}{max(|E_R|, |E_{K_i}|)} \tag{7}$$

These three similarity algorithms measure the similarity of two entities from different angles, and the average value is used as the final score of the similarity of two entities, as shown in Equation (8).

$$Sim(E_R, E_{Ki}) = (Sim_{ld} + Sim_{jacc} + Sim_{lcs})/3 \tag{8}$$

However, the negative words in EMRs will have an impact on the semantics of components in their jurisdiction. For example, for the descriptions of *There is no discomfort such as vaginal bleeding*(无阴道流血等不适) and *There is involuntary vaginal fluid*(不自主阴道流液) contain the negative words 无 and 不. The first word will change the actual semantics, but the latter word is only a description of *vaginal fluid*.

Therefore, we utilize the negative factor $f_{neg}$ to limit the influence of negative words on semantics. If the negative words that do not change or partially change semantics, the entities described by those words will be linked to COKG, and the negative factor is 1 or 0.5, respectively. For those negative words that will change semantics, their negative factor is -1.

**Diseases Weighted Computation**

Through entity linking above, we can obtain the symptoms set $S_R = \{s_{R_1}, s_{R_2}, ..., s_{R_i}, ..., s_{R_m}\}$ and the diseases set $D_R = \{(d_{R_1} : f_{R_1}), (d_{R_2} : f_{R_2}), ..., (d_{R_i} : f_{R_i}), ..., (d_{R_q} : f_{R_q})\}$, where $f_{R_i}$ is the frequency of disease entity and $f_{R_1} \leq f_{R_2} \leq ... \leq f_{R_i} \leq ... \leq f_{R_q}$.

Then we propose a weight-based disease prediction algorithm named WBDP. The disease and symptom sets in COKG are denoted as $D_K$ and $S_K$. Through the matching of tail entities, we can get a set $D_i = \{d_{i_1}, d_{i_2}, ..., d_{i_j}, ..., d_{i_n}\}$ of n candidate disease entities in COKG for symptom $s_{R_i}$, the disease candidate set corresponding to all symptoms is denoted as $D$. For each disease $d_{ij}$ in candidate set D, there is a symptom set $S_{d_{ij}} = \{s_{d_{ij}1}, s_{d_{ij}2}, ..., s_{d_{ij}l}, ..., s_{d_{ij}M}\}$ containing m symptoms in COKG associated with it, and $Q_{ij} = S_R \cap S_{d_{ij}}$. The purpose of WBDP is to compute the weight of disease $d_{ij}$, as shown in Equation (9).

$$W_{d_{ij}} = \sum_{s_{R_i} \in S_R} \frac{f_{neg} \times p(s_{R_i}, d_{ij})}{\sum_{q_r \in Q_{ij}} p(q_r, d_{ij})} \log_2 \frac{|D|}{|D_i| + 1} \tag{9}$$

Where $|D_i|$ and $|D|$ are the number of diseases in set $D_i$ and $D$, $f_{neg}$ is the negative factor of $s_{R_i}$, $p(s_{R_i}, d_{ij})$ is the co-occurrence probability of symptom $s_{R_i}$ and disease $d_{ij}$ in COKG.

We adopt two methods to deal with the disease set $D_R$ contained in EMRs. If the disease negative factor $f_{neg}$ is -1, it will be removed from the candidate set. Otherwise, if the candidate set associated with symptoms already contains $d_{R_i}$, the weight $W'_{d_{R_i}}$ will be computed according to the $W_{d_{R_i}}$ and the frequency $f_{R_i}$, as shown in Equation (10).

$$W'_{d_{R_i}} = W_{d_{R_i}}(1 + \frac{f_{R_i}}{\sum_{f_{R_i} \in D_R} f_{R_i}}) \tag{10}$$

If the candidate set associated with symptoms does not contain $d_{R_i}$, it will be add to the candidate set. Its weight is $\beta$ times of the average weight, where $\beta$ is a hyper-parameter and $\beta \geq 1$, the Equation is shown in (11). It is means that the diseases in EMRs have more influence on the diagnosis results than the symptoms.

$$W_{d_{R_i}} = f_{neg} \times \frac{\beta}{|D|} \sum_{d_i \in Dise} W_{d_i} \tag{11}$$

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1156-1166, Hainan, China, October 31 - Novermber 1, 2020.

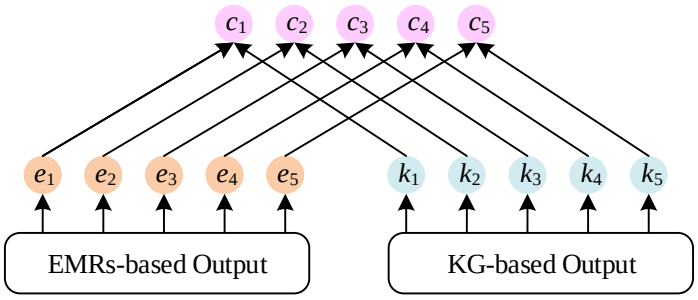


Figure 2: The fusion module of KEDA model

### 3.4 Fusion Module

The fusion module is aimed to integrate the output of the KG-based module into the output of the EMRs-based module. Inspired by the method proposed by (Chen et al., 2019), we employ a linear weighting method to fuse those representations, as shown in Figure 2.

The output of KG-based module and EMRs-based module is denoted as $K = [k_1, k_2, ..., k_i, ..., k_q]$ and $E = [e_1, e_2, ..., e_i, ..., e_q]$, where $k_i$ is the normalized representation of the weights mentioned above. The fusion process is shown in Equation (12).

$$c_i = \sigma(\gamma_i e_i + (1 - \gamma_i)k_i) = \frac{1}{1 - \exp(-(\gamma_i e_i + (1 - \gamma_i)k_i))} \tag{12}$$

Where $\sigma$ is the sigmoid function, $\gamma$ can be seen as a soft switch to adjust the importance of two representations. There are various ways to set the $\gamma$. The simplest one is to treat $\gamma$ as a hyper-parameter and manually adjust. Alternatively, it can also be learned by a neural network automatically, as shown in Equation (13).

$$\gamma = \sigma(W^T[K; E] + b) \tag{13}$$

Where $W$ and $b$ are trainable parameters.

### 3.5 Training

To train the KEDA model, the objective function is to minimize the cross-entropy in Equation (14).

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log P_i + (1 - y_i)\log(1 - P_i))] \tag{14}$$

Where $y_i \in \{0, 1\}$, $N$ is the number of labels, and $P$ is the model's prediction.

## 4 Experiments

### 4.1 The Procedure of Diagnosis Assistant

As shown in Figure 3, the procedure of diagnosis assistant can be divided into four parts: entity acquisition, entity linking, disease weighted computation, and weights fusion. For any given EMR, we obtain the entity sets through entity acquisition firstly, then the entities in those sets are linked to the COKG by a similarity-based method. As a result, we can get the disease nodes set and symptom nodes set from COKG. The WBDP algorithm is employed to compute the disease weights, and the negative factor $f_neg$ is used to limit the influence of negative words in EMRs for disease or symptom entities. Ultimately, the disease weights are regarded as the final knowledge representation to fuse the text representation so that we can get the diagnosis results.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1156-1166, Hainan, China, October 31 - Novermber 1, 2020.

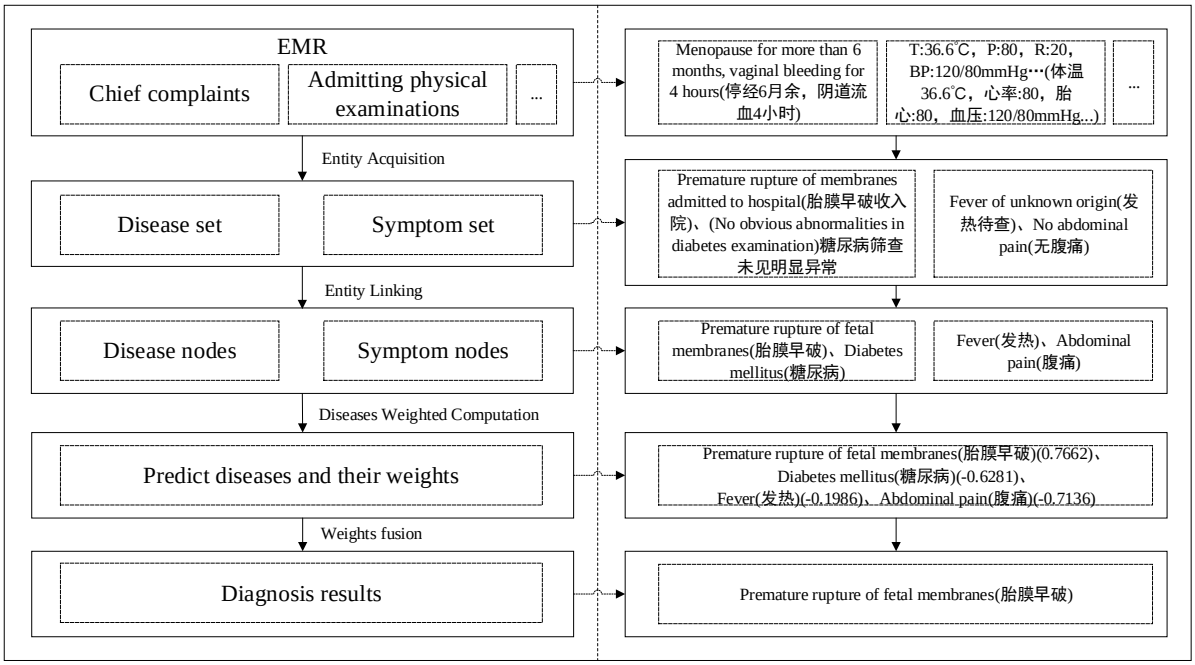


Figure 3: The procedure of diagnosis assistant

### 4.2 Dataset

We conducted experiments on the obstetric first course record dataset and COKG.

**Obstetric First Course Record Dataset.** The first course records include 24,339 EMRs from multiple hospitals in China. They were pre-processed through the steps of anonymization, data cleaning, structuring, and diagnostic label standardization. 21,905 of them were used for training and 2,434 were used for testing.

**COKG.** COKG uses the MeSH-like framework as the knowledge ontology to define the entity and relationship description system with obstetric diseases as the core. It contains knowledge from various sources such as the professional thesaurus, obstetrics textbooks, clinical guidelines, network resources, and other multi-source knowledge. COKG includes a total of 15,249 kinds of relations. Among them, 5,790 kinds of relations are semi-automatically extracted, and 9,459 kinds of relations are automatically extracted. The number and source of relations are shown in Table 1.

### 4.3 Experimental Setup

In this paper, the EMRs are preprocessed by de-identifying, data cleaning, structuring, data filtering, and standardization of diagnostic labels. During the data filtering process, the information that is duplicated and has little effect on the diagnosis is removed. On the one hand, it can meet the limitation of the input length of the BERT model, and on the other hand, it can also retain the useful information. The version of BERT model we used is BERT-base-Chinese, the main parameters are hidden size 768, max position embedding 512, num attention heads 12, num hidden layers 12, maximum input length 512, learning rate 5e-5, batch size 6, training epoch 20. All our experiments are run on an RTX2080ti GPU(12G).

### 4.4 Results

Experimental results on the obstetric first course record dataset are shown in Table 2. F1 (F1-micro), Hamming Loss, One Error, and AP (Average Precision) were used as evaluation metrics. BERT indicates the results of the baseline Google BERT, SGM is the results of SGM(Sequence Generation Model)(Yang et al., 2018), BERT+A, and BERT+A-AP are from (Zhang et al., 2019), which experiments are carried out on the same dataset as this paper. The KG-based means only use knowledge graph information, and KEDA is our proposed model.

Table 1: The relations statistics in COKG.

| Relation | Semi-automatic extraction | Automatic extraction | Total |
|---|---|---|---|
| disease-disease | 1,053 | 942 | 1,995 |
| disease-symptom | 1,680 | 3,199 | 4,879 |
| disease-anatomic site | 78 | 63 | 141 |
| disease-check | 529 | 815 | 1,344 |
| disease-medicine | 447 | 612 | 1,059 |
| disease-operation | 225 | 2 | 227 |
| disease-other treatments | 323 | 0 | 323 |
| disease-prognosis | 17 | 0 | 17 |
| disease-epidemiology | 160 | 84 | 244 |
| disease-sociology | 878 | 367 | 1,245 |
| disease-others | 170 | 2,889 | 3,059 |
| disease-synonym | 262 | 486 | 748 |

Table 2: The results on obstetric first course record dataset.

| Model | F1(%) | Hamming Loss | One Error | AP(%) |
|---|---|---|---|---|
| SGM | 60.00 | 0.0200 | 0.0630 | 39.00 |
| BERT | 79.58 | 0.0132 | 0.0961 | 84.97 |
| BERT+A | 80.26 | **0.0129** | 0.0863 | 85.42 |
| BERT+A-AP | 80.28 | **0.0129** | 0.0891 | 85.74 |
| KG-based | 53.57 | 0.0220 | 0.2417 | 52.13 |
| KEDA | **83.11** | 0.0143 | **0.00152** | **88.90** |

From Table 2, it can be seen that the improvements in our model over the BERT baseline and other results from (Zhang et al., 2019) are significant and consistent overall evaluation metrics. The AP of KG-based is only 52.13%, which is far lower than the result of KEDA. There may be two reasons for this situation, one of them may be some diagnoses are not obstetric diseases. Another possibility is that COKG is constructed from multi-source texts, which have different levels of detail for different diseases, it may make the number of triples of some diseases insufficient for accurate prediction.

Although the KG-based method does not have an advantage in various indicators, the results of the KEDA are better than BERT and others, indicating that the fusion of knowledge graph can improve the performance of diagnosis assistant. By further analyzing the diagnostic labels in the results, we find that the integration of knowledge graph is more obvious for the improvement of low-frequency labels. For example, the label *Placental abruption*(胎盘早剥) only appeared 5 times in the dataset, due to the scarcity of samples, it is difficult to make accurate predictions using only the BERT-based method. But there are 47 triplets in COKG that describe its symptoms, signs, and related diseases. After introducing the corresponding knowledge graph information, the accuracy of this type of disease has been significantly improved.

### 4.5 The Results of Entity Acquisition

As mentioned above, in order to choose a better entity acquisition method, we compared the bidirectional maximum matching and Bi-LSTM-CRF on the manually labeled 100 EMRs, the results are shown in Table 3. It is can be seen that the effect of the bidirectional maximum matching method is better than Bi-LSTM-CRF in testing. Bi-LSTM-CRF is trained on texts such as obstetric teaching materials, national norms, clinical practice, etc.

The differences in training data and test data may have an impact on the effectiveness of the model. The dictionaries of the bidirectional maximum matching method come from CSKB and ICD-10, which are more suitable for the description and content in obstetric EMRs. This may be one of the reasons for

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1156-1166, Hainan, China, October 31 - Novermber 1, 2020.

Table 3: The results of entity acquisition.

| **Method** | **F1(%)** | **P(%)** | **R(%)** |
|---|---|---|---|
| Bidirectional Maximum Matching | 89.42 | 85.20 | 94.10 |
| Bi-LSTM-CRF | 86.53 | 88.10 | 85.03 |

Table 4: The setting of hyper-parameter $\gamma$ on KEDA.

| $\gamma$ | **F1(%)** | **P(%)** | **R(%)** | **AP(%)** |
|---|---|---|---|---|
| 0.1 | 62.46 | 63.25 | 60.23 | 64.70 |
| 0.3 | 64.24 | 65.32 | 63.68 | 66.57 |
| 0.5 | 75.30 | 77.38 | 74.19 | 78.95 |
| 0.7 | 77.23 | 79.86 | 74.52 | 80.90 |
| 0.9 | 71.25 | 73.19 | 68.26 | 74.28 |
| Trained | **83.11** | **87.21** | **79.36** | **88.90** |

its better effect on entity acquisition.

### 4.6 The Setting of Hyper-Parameter $\gamma$

The goal of this part is to verify the effectiveness of the fusion module. Firstly, We manually tune the hyper-parameter $\gamma$ to explore the relative importance of EMRs-based and KG-based. We adjust $\gamma$ from 0 to 1 with an interval of 0.2, and the results are shown in Table 4. When $\gamma$ is equal to 0 or 1, the model will become the KG-based or EMRs-based, its results can be found in Table 2. From these results, the model with $\gamma = 0.7$ performs best. When $\gamma$ gradually increases, the model performs better, but after 0.7, the performance of the KEDA will decline. This shows that too much introduction of knowledge will also affect the overall performance of the model.

Moreover, the hyper-parameter $\gamma$ is treated as a trainable parameter to train with the model, the results are shown in the last row of Table 4. Compared with manual adjustment, the way to use $\gamma$ as a trainable parameter is a better choice.

### 4.7 Error Analysis

In this section, we analyze the bad cases induced by our KEDA model. Most of bad cases can be divided into two categories.

First, some entities in EMRs are not obstetric disease or symptom, which can not find their corresponding nodes in COKG. For example, those entities like *otitis media*(中耳炎), *glaucoma*(青光眼) and so on, there are not enough descriptions in COKG. Thus, the model can not make the correct diagnosis.

Second, COKG is constructed on multi-source obstetric disease texts, which have different levels of detailed description of different diseases. Among them, the proportion of diseases with less than 10 triplets accounts for more than 60%. If some diseases have fewer triplets in COKG, the model also cannot achieve good performance.

## 5 Conclusion

In this paper, the obstetric diagnosis assistant task is treated as a multi-label classification problem. We propose a KEDA model for this task, which integrates the numerical information from EMRs and external knowledge from COKG to improve the performance of diagnosis. We utilize the bidirectional maximum matching method to get the entities in EMRs, and the similarity-based approach is used to link the entities in knowledge graph COKG. Then we propose a WBDP algorithm to compute the weights of the entities in the candidate set. Finally, a linear weighting method is employed to fuse the text representation and knowledge representation. The results on the obstetric EMRs support the effectiveness of our approach compared to the BERT model. It turns out that even though the pre-training of BERT

Proceedings of the 19th China National Conference on Computational Linguistics, pages 1156-1166, Hainan, China, October 31 - Novermber 1, 2020.

involves a large number of corpora, the knowledge graph of the specific domain can still provide useful information.

In the future, we will incorporate more valuable information into deep neural networks to further improve the performance of the diagnosis assistant. We find that some disease entities in EMRs are not included in COKG(For example, the disease entity 'patella fracture' is a diagnosis label in EMRs, but it is not an obstetric disease), to introduce other knowledge graphs that contain more disease entities is an effective feature for diagnosis.

## Acknowledgements

This work has been supported by the National Key Research and Development Project (Grant No. 2017YFB1002101), Major Program of National Social Science Foundation of China (Grant No. 17ZDA138), China Postdoctoral Science Foundation (Grant No. 2019TQ0286), Science and Technique Program of Henan Province (Grant No. 192102210260), Medical Science and Technique Program Co-sponsored by Henan Province and Ministry (Grant No. SB201901021), Key Scientific Research Program of Higher Education of Henan Province (Grant No. 19A520003, 20A520038), the MOE Layout Foundation of Humanities and Social Sciences (Grant No. 20YJA740033), and the Henan Social Science Planning Project (Grant No. 2019BYY016).

## References

Simon Baker and Anna Korhonen. 2017. Initializing neural networks for hierarchical multi-label text classification. In *BioNLP 2017*, pages 307–315, Vancouver, Canada,, August. Association for Computational Linguistics.

Guibin Chen, Deheng Ye, Zhenchang Xing, Jieshan Chen, and Erik Cambria. 2017. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2377–2383. IEEE.

Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. 2019. Deep short text classification with knowledge powered attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6252–6259.

China's Ministry of Health. 2010. Basic specification of electronic medical records (trial). Technical Report 3.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Rong Li Gai, Fei Gao, Li Ming Duan, Xiao Hui Sun, and Hong Zheng Li. 2014. Bidirectional maximal matching word segmentation algorithm with rules. In *Advanced Materials Research*, volume 926, pages 3368–3372. Trans Tech Publ.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Gakuto Kurata, Bing Xiang, and Bowen Zhou. 2016. Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–526.

Mingchen Li, Gabtone Clinton, Yijia Miao, and Feng Gao. 2020. Short text classification via knowledge powered attention with similarity matrix based cnn. *arXiv preprint arXiv:2002.03350*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hongchao Ma, Kunli Zhang, and Yueshu Zhao. 2018a. Study on obstetric multi-label assisted diagnosis based on feature fusion. *Journal of Chinese Information Processing*, 32(5):128–136.

Shuming Ma, Xu Sun, Yizhong Wang, and Junyang Lin. 2018b. Bag-of-words as target for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 332–338, Melbourne, Australia, July. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning*, 85(3):333.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2010. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079–1089.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Hui-li Yang and Zi Yang. 2016. Effect of older pregnancy on maternal and fetal outcomes. *Chinese Journal of Obstetric Emergency(Electronic Editon)*, 5(3):129–135.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: sequence generation model for multi-label classification. pages 3915–3926.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019a. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, 18(10):1338–1351.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048.

Kunli Zhang, Hongchao Ma, Yueshu Zhao, Hongying Zan, and Lei Zhuang. 2018. The comparative experimental study of multilabel classification for diagnosis assistant based on chinese obstetric emrs. *Journal of healthcare engineering*, 2018.

Kunli Zhang, Chuang Liu, Xuemin Duan, Lijuan Zhou, Yueshu Zhao, and Hongying Zan. 2019. Bert with enhanced layer for assistant diagnosis based on chinese obstetric emrs. In *2019 International Conference on Asian Language Processing (IALP)*, pages 384–389. IEEE.