Explainable Pulmonary Disease Diagnosis with Prompt-Based Knowledge Extraction

Haodi Zhang[1], Chenyu Xu[1,2], Jiahong Li[1], Peirou Liang[1], Xiangyu Zeng[1], Hao Ren[2], Weibin Cheng[2], Kaishun Wu[1]
[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2]Institute for Healthcare Artificial Intelligence Application, Guangdong Second Provincial General Hospital, Guangzhou, China

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995532

***Abstract*—Recent studies show that deep learning models perform well in many medical tasks such as medical imaging and automated diagnosis. With qualified t raining d atasets, some models can achieve or even surpass expert-level performance on some tasks. However, as a typical black-box-style approach, deep learning lacks theoretical interpretability, which is especially important for medical tasks. On the other hand, there are many sources of domain knowledge for medical diagnosis from human experts, such as clinical guidelines. How to sufficiently integrate human knowledge in the model is crucial for explainable diagnosis. In this paper, we propose a novel framework for explainable automated diagnosis that leverages explicit medical knowledge. We automate the knowledge extraction from textual clinical guidelines with prompt-based learning, train a set of weighted first-order l ogical r ules w ith c onstructed e vidence d atabase, and finally i nfer t he d iagnosis r esult w ith i ntegrated k nowledge and multi-sourced data. We instantiate the framework for pulmonary disease diagnosis, and our experiments on a real dataset show that our method outperforms the state-of-the-art baselines in accuracy and interpretability.**

***Index Terms*—knowledge extraction, explainable diagnosis**

I. Introduction

In the past few years, lots of studies show that deep learning models have the potential expertise in solving medical tasks such as text-based automated diagnosis. In particular, large-scale pre-trained language models (PLMs) have been widely used and achieve expressive performance for text-based diagnosis classification, f or i nstance. A fter fine-tuning on proper corpus in the medical domain, they can even achieve expert-level performance. However, these black-box models learn from the data and embed the knowledge in an implicit form of hyper-parameters. Consequently, most deep learning-based methods lack theoretical interpretability for the final results or decisions, which is especially important for both doctors and patients in disease diagnosis.

In the healthcare field, p lenty o f k nowledge b ases are available. These knowledge bases such as clinical guidelines play an important role during the human practice of medical diagnosis and treatment [1]. Supposedly, the effective integration of explicit medical knowledge into the machine-learning model helps to make a trustworthy decision.

This work is supported by Guangdong Basic and Applied Basic Research Foundation (2022A1515010675). Weibin Cheng and Kaishun Wu are the corresponding authors.

Therefore, we propose an explainable diagnosis framework that leverages medical knowledge from textual clinical guidelines. We automatically recognize useful predicates and formalize the knowledge into first-order logical (FOL) rules. We then build an evidence database from the multi-source data in EMRs. With the integrated knowledge and the multi-source evidence database, we construct and train a Markov Logic Network (MLN) to infer the final diagnosis result. We instantiate the framework for pulmonary disease diagnosis and conduct experiments on a real dataset from Guangdong Second Provincial General Hospital. The main contributions are listed as follows:

- We propose a framework that automatically extracts medical knowledge from authoritative clinical guidelines with prompt learning.
- We integrate the explicit knowledge into the data-driven decision-making model and infer the diagnosis.
- We instantiate the framework for Pulmonary Disease Diagnosis, and conduct experiments on a real-world dataset collected from GD2H. The experimental results show that our proposal outperforms the previous state-of-the-art baselines.

II. Related work

In this section, we briefly discuss some of the existing works related to ours.

A. Prompt-Based Learning

Recently, prompt learning [2] has become very popular in the natural language processing field. The original input *X* is modified into a textual string prompt *X'* that has some masks. The PLM performs prediction tasks by filling the masks with probability, from which the final output *Y* can be derived.

The basic version of prompting does not require fine-tuning, and instead predicts directly the masked tokens based only on the prompt template. Masked language models, such as BERT [3], and ERNIE [4], are naturally suitable for tuning-free prompting.

B. Markov Logic Network

MLN is a statistical relational learning model that combines Markov networks with FOL in uncertainty inference [5]. The basic idea is to combine FOL rules with weights while allowing for a relaxation of those hard rules. For example, if

one of the rules is violated, its existence is less likely, but not impossible. The larger weight of a rule, the more significant the rule is. If we increase the weights infinitely, the results derived by MLNs converge to those derived by first-order logic [6]. We use mature MLN learning and inference algorithms for disease diagnosis.

III. Method

In this section, we propose a novel framework of explainable diagnosis with prompt-based knowledge extraction. As shown in Figure 1, the framework consists of three main parts, namely, prompt-based knowledge extraction, evidence database construction and diagnosis inference. In the first part, a prompt learning module is utilized to automatically extract FOL rules from textual clinical guidelines after named entity recognition. The second part constructs the evidence database for the downstream training by concatenating a NER module and a well-designed schema matching procedure. The formalized knowledge rules and the evidence database are delivered to the third part for weight learning and diagnosis inference.

A. Prompt-Based Knowledge Extraction

The knowledge extraction module automatically formalizes the first-order logic rules set from diagnostic criteria in the clinical guidelines. The extraction process is similar to cause-effect relation extraction but simpler. The causes include clinical manifestations and lung imaging findings, and the effect is the establishment of the clinical diagnosis of pulmonary disease.

We first use the named entity recognition (NER) module and schema matching to extract logical symbols. The details of the processing are available in the supplementary materials. We then utilize PLM prompting and consistency parser to extract logical relations between logic symbols.

For the clinical diagnostic criteria in Chinese, we choose BERT-Base-Chinese [7] for zero-shot prompt learning. The following is a sample prompt.

> *If item 2 and any of item 1 are met, the clinical diagnosis can be established. If item [MASK] is met, a clinical diagnosis can be established.*

The first part is the last sentence of the diagnostic guideline, where item 1 refers to the clinical manifestations and item 2 refers to the lung imaging findings, and the second part makes a mask of the causes. Then we use PLM to prompt an answer by means of filling mask.

We employ an off-the-shelf constituency parser with a trained Chinese language model on this sentence. We design a generic logic identification approach that uses a parser tree to further extract nouns and gerundial phrases as a constituency. For example, we extract "item 2 and any of item 1" from the sentence as a constituency. In addition, further parsing this constituency, we can determine that item 2 and item 1 are coordinating conjunctions, and the number of conjuncts k=2.

Technically, the [MASK] position can only prompt one token at a time and "2" and "1" are the candidate tokens for prompting. We regard the output of top k as the final result. For example, the top 2 outputs by prompting are "2" and "1". Based on the logical symbols and logical relation, we can combine item 2 and item 1 to output the FOL rules.

B. Construction of Evidence Database

We use NER and Schema Matching to build the evidence database. NER is a fundamental task to extract medical entities, such as diseases, clinical symptoms and medical procedures, etc. In our framework, we fine-tune a Chinese pre-trained medical model ERNIE-Health [8] on downstream task CMeEE (Chinese Medical Named Entity Recognition Dataset) V2 [9] to extract entities from EMRs. The fine-tuned model reaches an advanced F1 score of 74% on the CMeEE V2 validation set.

After getting the NER results, we need to do some post-processing with entities (see supplementary materials in detail) and begin to construct an evidence database. First, we put the required predicates contained in FOL rules into a dictionary as the query object. However, some predicates may not be found in the NER results. Therefore, we use regular expression matching to retrieve them from semi-structured EMRs. Finally, we get the full predicates from each patient's case and construct the evidence database for each patient.

C. Inference with Integrated Knowledge and Multi-sourced Data

We split the evidence database into training set and test set in a ratio of 8:2 for each disease. We use the training set to learn the weight of each rule. Finally, we infer disease probability with these weighted rules.

1) Learning weights of diagnostic rules: To learn the weights, we first initialize the weights to zero, and then we use the MLN weight learning method [10] to train them. Specifically, we treat all rules as a linear equation array and use the numerical iterative algorithm conjugate gradient method [11] to solve for weights. If a rule is satisfied more times, it means that the more effective the rule is in real data, the greater the weight of the rule will be.

We directly call the open-source MLN tool alchemy [12] for training and inference. A training sample is a case, and the label is the diagnosed disease of the case. Since the disease predicates are query predicates, they need to be labeled as non-evidence predicates during training. For example, after training, we get the sample results as follows.

$$CXR\_PN \wedge Fever \Rightarrow Pneumonia : 2.478$$

2) Diagnosis inference: After training, we obtain a set of weighted rules, and we use the MLN infer method to make disease inferences for the test set. Specifically, to speed up the inference, we only keep rules with weight not 0. A weight of 0 means that the rule has never been satisfied in the training set.

As mentioned before, MLN can be seen as a template for Markov network, which demonstrates the joint distribution probabilities of a set of observed variables and unobserved

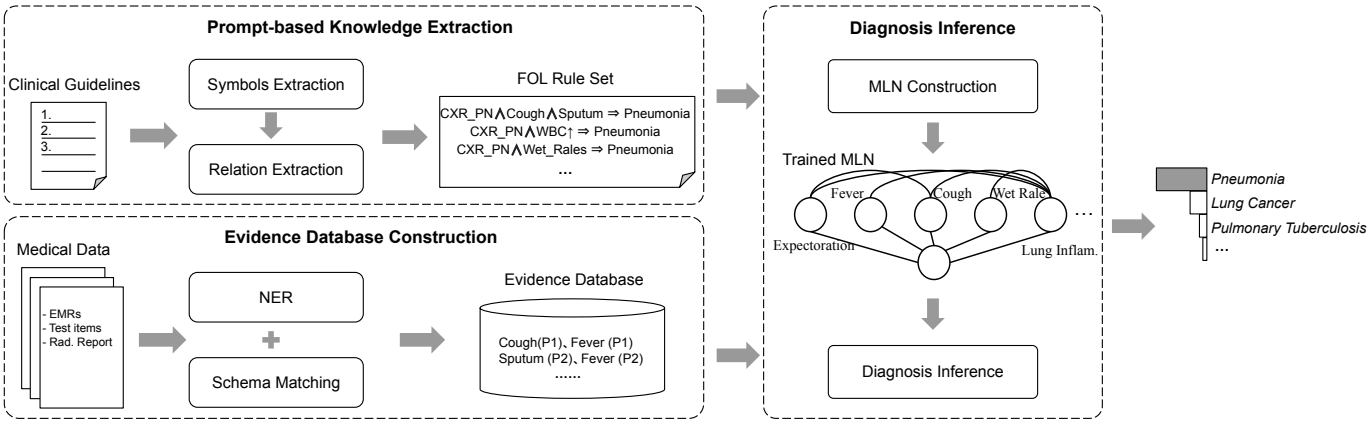

Fig. 1: Overview of the proposed explainable diagnosis with integrated knowledge

variables. The diagnosis inference is to calculate the marginal probability of query predicates.

In this experiment, disease diagnosis can be viewed as a disease multi-classification task. Each case has only one diagnosis result. Therefore, the final marginalization will only be left with an unobserved variable, i.e., the marginal probability of a certain disease to be inferred about. When performing probabilistic inference, we have two major classes of inference algorithms, exact and approximate inference. Since we have a large amount of data, we use the classical MCMC sampling method for inference.

For the final diagnosis inference, we need to mark disease predicates as query predicates when testing. For example, the final inference result is:

$$\begin{aligned} &\max_{D\in\mathcal{D}} P(D, c\_id) = P(Pneumonia, c\_id) = 0.975 \\ &\arg\max_{D\in\mathcal{D}} P(D, c\_id) = Pneumonia \end{aligned} \tag{1}$$

where the c_id is the case ID.

IV. Experiments

In this section, we introduce how we construct our experiment dataset and the evaluation results. By comparing with other models, we found that our model has good performance in disease diagnosis inference. From the ablation study, we can see the effect of dataset size and the long-tail data distribution.

A. Dataset

In the experiment, we collect four common Pulmonary Diseases EMRs from the hospital as shown in Table I. After screening and preprocessing, we obtain 10354 EMRs in total, of which 7016 are in training sets and 1754 in test sets.

In fact, an EMR usually does not have only one diagnosis, so we choose the primary diagnosis in the EMR as the ground truth label. Since the incidence of each disease is different in reality, it is common that the dataset has a long-tailed problem. Therefore, we use oversampling to deal with the imbalanced classes. Unfortunately, we cannot make the dataset public because of the data privacy protocol.

TABLE I: The statistics of the datasets. #EMRs is the number of the collected EMRs.

Pulmonary Disease	#EMRs
Pneumonia	2667
Tuberculosis	1609
Lung cancer	4547
Pulmonary embolism	1531
All	10354

TABLE II: Performance of our framework against baselines.

Model	Accuracy	AUC	F1	Sensitivity	Specificity
BERT	0.819	0.885	0.802	0.818	0.875
BertGCN	0.820	0.896	0.808	0.822	0.897
ERNIE -Health	0.824	0.898	0.833	0.851	**0.930**
Ours	**0.892**	**0.927**	**0.886**	**0.892**	0.851

B. Comparison with baselines

In this main experiment, we select 3 deep learning based text multi-classification baselines to compare with our model. The details of evaluation metrics refer to the supplementary materials.

1) Performance: As shown in Table II, our model achieved the highest performance in accuracy, AUC and F1 score. In particular, three deep learning based baselines achieved an accuracy of 0.81 to 0.83, while our model achieved an accuracy close to 0.90. Since there is a long-tail problem with the dataset, accuracy is not a fair evaluation metric on imbalanced data, we compute other evaluation metrics.

Among the 3 baseline models, ERNIE-Health has the best performance. ERNIE-Health learns more medical knowledge and medical entities from a large-scale medical corpus, like electronic medical records, medical journals, etc., helping it to perform better in the medical domain task. Our model improves 0.03, 0.05, 0.04 in AUC, F1 score and sensitivity over ERNIE-Health. However, it is not as good as ERNIE-

Health in terms of specificity.

In the medical field, sensitivity and specificity are two commonly used metrics, which represent the true positive rate and true negative rate, respectively, and can represent the rate of missed diagnosis and misdiagnosis. Our model has fewer missed diagnoses, while ERNIE-Health has fewer misdiagnoses. The results show that our model outperforms other competing baselines in most evaluation metrics and it is also interpretable.

2) Interpretability: Leveraging domain knowledge explicitly, our model infers the disease probability with the weighted rules set in the probabilistic inference method. We illustrate the interpretability with the visualization of MLN in Figure 1. Each node represents a predicate defined before and the arc lines represent the logical relation of the connected nodes.

For example, if only fever is extracted from a patient's EMR, then pneumonia will be predicted with a low probability because there are many situations causing fever. If fever and lung inflammation exist at the same time, then pneumonia will be predicted with a high probability. The proposed method can explain the prediction of pneumonia according to the diagnostic rules set.

C. Ablation study

In this section, we conduct ablation experiments on the effect of dataset size and long-tail data distribution on the model.

TABLE III: Ablation study with dataset size and long-tail data distribution on the model.

Model	Accuracy	F1	Sensitivity	Specificity
Ours	**0.892**	**0.886**	**0.892**	**0.851**
Ours (w. small size)	0.874	0.865	0.874	0.824
Ours (w. long-tailed distribution)	0.842	0.839	0.842	0.824

1) The effect of dataset size: In this section, we explore the effect of dataset size on our model. In this experiment, we only take 1/10 of each disease case in Table I and experiment on the data subset. Table III lists the results of original size and small size multimodal model performance. We find that although the dataset size is much smaller than the original, the performance does not drop too much. Even if the size is only 1/10 of the original, the accuracy reaches 0.874, and the F1 score reaches 0.865, which is still better than the ERNIE-Health model. The experimental results show that dataset size has little impact on our model. Even with a small dataset, we can build a disease diagnosis system in the startup phase and obtain decent performance. As we have more and more data over time, it can continuously help the system to improve performance.

2) The effect of long-tail data distribution: As we mentioned before, different diseases have different prior probabilities of prevalence in the population, which leads to a general problem of long-tail disease data distribution. In our dataset, the number of cases of pneumonia and lung cancer is much larger than that of tuberculosis and pulmonary embolism.

In this experiment, we compare the effect of the data long-tail problem on model performance. As shown in the Table III, the result shows a general decrease of 3%-5% in each evaluation metric. It indicates that our model is also affected by the unbalanced class data. Specifically, we find that the model learns the prior probability of a disease in data-driven learning. If a case does not satisfy any of the diagnostic rules, it will be likely to be classified into the major class.

Therefore, it is necessary for oversampling tuberculosis and pulmonary embolism cases to improve the model performance. In addition, we can manually set a prior probability for each disease in practice to avoid the model learning the wrong prior probability in the data.

V. Conclusion

In this paper, we propose a framework of explainable diagnosis with prompt-based knowledge extraction. The framework successfully leverages expert knowledge for automated diagnosis. Our experiments on a real pulmonary disease dataset show the effectiveness of our method.

References

[1] D. A. Grimes, D. Hubacher, K. Nanda, K. F. Schulz, D. Moher, and D. G. Altman, "The good clinical practice guideline: a bronze standard for clinical research," *The Lancet*, vol. 366, no. 9480, pp. 172–174, 2005.

[2] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.

[3] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[4] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "Ernie: Enhanced language representation with informative entities," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1441–1451.

[5] M. Richardson and P. Domingos, "Markov logic networks," *Machine learning*, vol. 62, no. 1, pp. 107–136, 2006.

[6] L. Getoor and B. Taskar, "Markov logic: A unifying framework for statistical relational learning," 2007.

[7] Z. Yang, Z. Zhixiong, L. Huan, and D. Liangping, "Classification of chinese medical literature with bert model," *Data Analysis and Knowledge Discovery*, vol. 4, no. 8, pp. 41–49, 2020.

[8] Q. Wang, S. Dai, B. Xu, Y. Lyu, Y. Zhu, H. Wu, and H. Wang, "Building chinese biomedical language models via multi-level text discrimination," *arXiv preprint arXiv:2110.07244*, 2021.

[9] N. Zhang, M. Chen, Z. Bi, X. Liang, L. Li, X. Shang, K. Yin, C. Tan, J. Xu, F. Huang *et al.*, "Cblue: A chinese biomedical language understanding evaluation benchmark," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7888–7915.

[10] P. Singla and P. Domingos, "Discriminative training of markov logic networks," in *AAAI*, vol. 5, 2005, pp. 868–873.

[11] B. T. Polyak, "The conjugate gradient method in extremal problems," *USSR Computational Mathematics and Mathematical Physics*, vol. 9, no. 4, pp. 94–112, 1969.

[12] S. Kok, M. Sumner, M. Richardson, P. Singla, H. Poon, D. Lowd, J. Wang, and P. Domingos, "The alchemy system for statistical relational ai," Technical report, Department of computer science and engineering, university ..., Tech. Rep., 2005.