Bioinformatics, 38(16), 2022, 3995–4001
https://doi.org/10.1093/bioinformatics/btac408
Advance Access Publication Date: 1 July 2022
Original Paper

OXFORD

Data and text mining

Hierarchical reinforcement learning for automatic disease diagnosis

Cheng Zhong [1,†], Kangenbei Liao[1,†], Wei Chen [1], Qianlong Liu[2], Baolin Peng[3], Xuanjing Huang[4], Jiajie Peng[5,*] and Zhongyu Wei[1,5,*]

[1]School of Data Science, Fudan University, 200433 Shanghai, China, [2]Alibaba Group, 310052 Hangzhou, China, [3]Microsoft Research, Redmond, WA 98052, USA, [4]School of Computer Science, Fudan University, 200433 Shanghai, China and [5]Research Institute of Intelligent Complex Symtems, Fudan University, 200433 Shanghai, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on March 21, 2022; revised on May 16, 2022; editorial decision on June 3, 2022; accepted on June 29, 2022

Abstract

Motivation: Disease diagnosis-oriented dialog system models the interactive consultation procedure as the Markov decision process, and reinforcement learning algorithms are used to solve the problem. Existing approaches usually employ a flat policy structure that treat all symptoms and diseases equally for action making. This strategy works well in a simple scenario when the action space is small; however, its efficiency will be challenged in the real environment. Inspired by the offline consultation process, we propose to integrate a hierarchical policy structure of two levels into the dialog system for policy learning. The high-level policy consists of a master model that is responsible for triggering a low-level model, the low-level policy consists of several symptom checkers and a disease classifier. The proposed policy structure is capable to deal with diagnosis problem including large number of diseases and symptoms.
Results: Experimental results on three real-world datasets and a synthetic dataset demonstrate that our hierarchical framework achieves higher accuracy and symptom recall in disease diagnosis compared with existing systems. We construct a benchmark including datasets and implementation of existing algorithms to encourage follow-up researches.
Availability and implementation: The code and data are available from https://github.com/FudanDISC/DISCOpen-MedBox-DialoDiagnosis
Contact: jiajiepeng@nwpu.edu.cn or zywei@fudan.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the development of electronic medical records (EMRs), researchers have explored different machine learning approaches for automatic diagnosis (Richens and Buchard, 2022; Shivade *et al.*, 2014). Although impressive results have been reported for the identification of various diseases, e.g. heart failure with preserved ejection fraction (HFpEF) (Jonnalagadda *et al.*, 2017) and autism spectrum disorders (ASDs) (Doshi-Velez *et al.*, 2013), they rely on well-established EMRs which are labor-intensive to build. Moreover, the automatic diagnosis of a certain disease requires EMRs of that type for model training and is difficult to be extended to other types of diseases.

To relieve the pressure of constructing EMRs, researchers (Wei *et al.*, 2018; Xu *et al.*, 2019) introduce the task-oriented dialog system to request symptoms automatically from patients for disease diagnosis. Since the disease consultation is an interactive procedure with multiple time steps, they formulate the task as Markov decision processes (MDPs) and employ reinforcement learning (RL) based methods for the policy learning. In each round of interaction, the agent either chooses a symptom to request or makes the diagnosis via selecting an action from the joint action space of symptoms and diseases. Correct symptom query and disease diagnosis will be rewarded, and the policy is learned by maximizing the expected cumulative rewards.

After that, reinforcement learning becomes the first choice for researchers in this field (Coronato *et al.*, 2020; Kao *et al.*, 2018; Peng *et al.*, 2018; Tang *et al.*, 2016; Yu *et al.*, 2021). Kao *et al.* (2018) presented a context-aware hierarchical reinforcement method, using policy gradients to make decisions based on the patient's personal information and explicit symptoms. Peng *et al.* (2018) proposed reward shaping and feature rebuilding techniques to help agents effectively learn a better strategy and Chen *et al.* (2019) introduced a new multiple action policy representation to help agents suggest medical tests to facilitate disease diagnosis. Also, many researchers tried to combine the RL-based and non-RL approaches. Xu *et al.* (2019) introduced the knowledge graph into their dialog system, Xia *et al.* (2020) applied GAN as a policy network to capture the relations between different symptoms, and Lin *et al.* (2020) proposed DSMAD method which was inspired by the introspective decision-making process of human to make the diagnosis process more reliable. Recently, Hou *et al.* (2021) proposed a multi-level reward modeling approach, and Teixeira *et al.* (2021) proposed an approach to automating the generation of a dialog manager to achieve the predictability and reliability. However, existing policies are designed with flat and monolithic structures (such as MLPs), which are not salable to deal with scenarios including a large number of diseases and symptoms.

In the actual diagnosis scenario, we find that the relationship between diseases and symptoms can help us classify the disease. In Figure 1, we present the proportion of symptoms related to four different diseases, i.e. children bronchitis, upper respiratory infection, children functional dyspepsia and infantile diarrhea. It shows that a particular disease is usually related to a certain group of symptoms. In offline consultation, doctors also do the pre-examination and triage according to the different symptoms that the patient suffered, then doctors in different departments will make a more detailed diagnosis. This method significantly reduces the workload of individual doctors and enables them to be more specialized in a certain field.

Hierarchical reinforcement learning (HRL) (Parr and Russell, 1998; Sutton *et al.*, 1999), in which multiple layers of policies are trained to perform decision-making, conforms to the problem-solving logic for disease diagnosis in the real environment. HRL has been successfully applied to different scenarios, inter alia, course recommendation (Zhang *et al.*, 2019), visual dialog (Zhang *et al.*, 2018), relation extraction (Feng *et al.*, 2018) and composite tasks with slot constraints (Lipton *et al.*, 2018). In each step, the agent chooses either a one-step 'primitive' action or a 'multi-step' action (option). Schatzmann *et al.* (2007) presented a POMDP dialog system for simulating user behavior, which can train and test a prototype system. Then, researchers showed that HRL dialog agents are feasible and promising in large-scale systems (Cuayáhuitl *et al.*, 2010). In order to improve the generalization ability of the HRL model, Florensa *et al.* (2017) proposed a general framework that first learns useful skills (high-level policies) in an environment and then leverages the acquired skills for learning faster in downstream tasks. Then, Budzianowski *et al.* (2017) applied HRL to multi-domain dialog management, which showed the potential of HRL to facilitate policy optimization for more sophisticated multi-domain dialog systems. Lipton *et al.* (2018) used HRL to efficiently learn the dialog manager that operates at different temporal scales. Up to now, HRL has been successfully applied to different tasks and reached promising results (Duan *et al.*, 2020; Feng *et al.*, 2018; Guo *et al.*, 2018; Takanobu *et al.*, 2018; Wan *et al.*, 2020; Wang *et al.*, 2018; Zhang *et al.*, 2018, 2019). Most existing works decompose the corresponding task into two steps manually, where the first step is treated by the high-level policy while the second step is treated by the low-level policy. This motivates us to divide the online diagnosis tasks into different levels following the setting of departments in the hospital and design a hierarchical structure for symptom acquisition and disease diagnosis.

In this article, we propose to build a dialog system with a hierarchy of two levels for automatic disease diagnosis using HRL methods. The high-level policy consists of a model named master and the low-level policy consists of several workers and a disease classifier. The master is responsible for triggering a model at a low level. Each worker is responsible for inquiring about symptoms related to a certain group of diseases while the disease classifier is responsible for making the final diagnosis based on information collected by workers. The proposed framework imitates a group of doctors from different departments diagnosing a patient together. Among them, each worker acts like a doctor from a specific department, while the master acts like a committee that appoints doctors to interact with this patient. When the information collected from workers is sufficient, the master would activate a separate disease classifier to make the diagnosis. Models in the two levels are trained jointly for better disease diagnosis. We test our model in three large real-world datasets and a synthetic dataset. Experimental results demonstrate that the performance of our hierarchical framework outperforms other state-of-the-art approaches.

We summarize our contribution as follows: (i) We propose a new RL-based method for automatic diagnosis. It simulates the real scene of clinical practice, and assigns patients to different workers through high-level policy, thereby reducing the action space and improving training efficiency. Also, the method can be compatible with different network models and training strategies. (ii) We perform a systematical evaluation to test the performance of our model on three public datasets from the real environment and synthetic datasets. Overall experiment results show the advantage of our model compared to state-of-the-art baselines and further analysis confirms the effectiveness of each component in our framework. (iii) We release a toolkit with the implementation of all existing baseline models and datasets. The toolkit can be used as a benchmark for dialog-based disease diagnosis.

2 Materials and methods

In this section, we introduce our HRL framework for disease diagnosis. We start with the formulation of the MDP for automatic disease diagnosis and then introduce the hierarchical setting.

2.1 MDP setup for disease diagnosis

As for RL-based models for automatic diagnosis, the action space of agent is $\mathcal{A} = D \cup S$, where D is the set of all diseases and S is the set of all symptoms that associated with these diseases. Given the state $s_t \in \mathcal{S}$ at turn t, the agent takes an action according to its policy $a_t \sim \pi(a|s_t)$ and receives an immediate reward $r_t = R(s_t, a_t)$ from the environment (patient/user). If $a_t \in S$, the agent chooses a symptom to inquire the patient/user. Then the user responds to the agent with *true/false/unknown*. If $a_t \in D$, the agent informs the user of the corresponding disease as the diagnosis result and the dialog session will be terminated as the success/failure in terms of the correctness of the diagnosis.

2.2 Hierarchical policy structure for disease diagnosis

To reduce the large action space, we extend the flat-RL structure to a hierarchical structure with two-layer policies for automatic diagnosis. Following the *options* framework (Sutton *et al.*, 1999), our

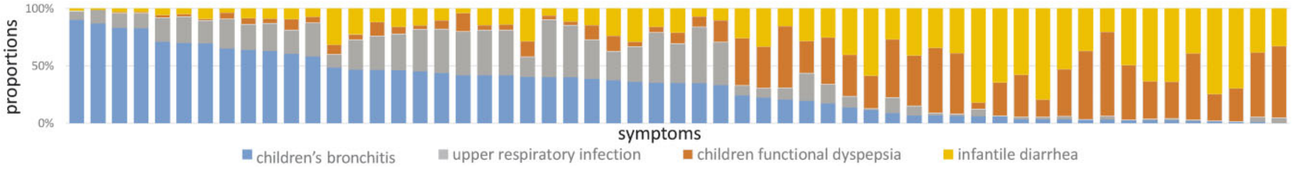
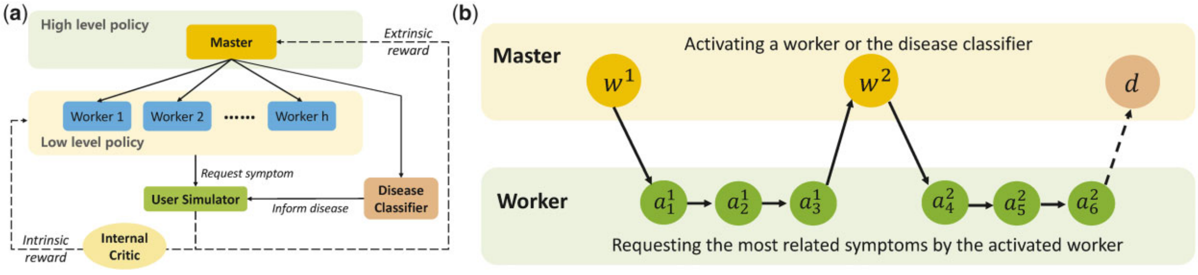


Fig. 1. Disease distribution over symptoms in MZ-4 (see Section 3.1). X-axis stands for symptoms and y-axis is the proportion. Each bar describes the disease distribution given a symptom. (a) Master–worker framework (training). (b) Master–worker framework (inference)



Master-worker framework (training)

Master-worker framework (inference)

Fig. 2. The training and inference process of our model. (a) The framework of our hierarchical reinforcement learning model with two-layer policies. (b) Illustration of the diagnosis process of our model with interactions between components of two levels. w^i is the action invoking worker w^i and d is the action invoking disease classifier

framework is illustrated in Figure 2a. There are four components in our framework: master, worker, disease classifier and user simulator. At turn t, the state s_t will be encoded as one-hot vectors that reflect the status of each symptom and number of turns for the master and worker network, and only symptom information will be encoded for the disease classifier. An illustration of the diagnosis process with interactions between models in two levels is presented in Figure 2b.

2.2.1 Master for classification

The master controls the higher policy of the agent. At each turn, the master can choose whether to activate the worker to collect symptom information or a disease classifier to make a decision. Once the master activates a worker, this worker will interact with the user for N turns until the subtask is terminated.

The learning problem of the master can be formulated as a semi-Markov decision process (SMDP), where the extrinsic rewards returned can be accumulated as the immediate rewards for the master (Ghavamzadeh, 2005). That is to say, after taking an action a_t^m, the reward r_t^m for the master can be defined as:

$$r_t^m = \begin{cases} \sum_{t'=1}^{N} \gamma^{t'} r_{t+t'}^e, & if \quad a_t^m = w^i \\ r_t^e, & if \quad a_t^m = d, \end{cases} \tag{1}$$

where γ is the discounted factor and w^i, d is the action to activate the worker w^i and disease classifier. The objective of the master is to maximize the extrinsic reward, thus we can write the master's loss function as follows:

$$\mathcal{L}(\theta_m) = \mathbb{E}_{s,a^m,r^m,s'\sim\mathcal{B}^m}\left[(y - Q_m(s,a^m;\theta_m))^2\right], \tag{2}$$

where $y = r^m + \gamma^N \max_{a^{m'}} Q_m(s', a^{m'}; \theta_m^-)$, θ_m, θ_m^- is the network parameter at current and previous iteration, $\mathcal{B}^m$ is the fixed-length buffer of samples for master.

2.2.2 Worker and disease classifier for interaction

The worker controls the lower policy of the agent and interacts with the patient to collect information for a specific group of symptoms. Once the worker w^i is invoked, it will take the corresponding state representation s^i from the master and generate an action $a^i \in \mathcal{A}_i^w$.

After taking action $a^i \in \mathcal{A}_i^w$, the state representation will be updated and worker w^i will receive an intrinsic reward r_t^i from the user simulator. So the objective of workers is to maximize the expected cumulative discounted intrinsic rewards. The loss function of worker w^i can be written in the following way:

$$\mathcal{L}(\theta_w^i) = \mathbb{E}_{s^i,a^i,r^i,s^{i'}\sim\mathcal{B}_i^w}[(y_i - Q_w^i(s^i,a^i;\theta_w^i))^2]. \tag{3}$$

Once the disease classifier is activated by the master, it will take the symptom information as input and output a vector $\mathbf{p} \in \mathbb{R}^{|D|}$, which represents the probability distribution over all diseases. The disease with the highest probability will be returned to the user as the diagnosis result. Also, the disease classifier will be jointly trained with the master and worker through supervised learning.

2.2.3 User simulator and internal critic for reward generation

Following (Wei *et al.*, 2018) and (Xu *et al.*, 2019), we set up a user simulator to interact with the agent. At the beginning of each dialog session, the user simulator samples a user goal from the training set randomly. Each piece of user goal contains two kinds of symptoms, namely explicit ones (obtained from the self-report) and implicit ones (obtained from the dialogs). Explicit symptoms will be directly provided to the agent as initial information at the beginning, and the agent needs to discover the implicit symptoms during the interaction with the patient. The simulator will initialize the dialog session based on the explicit symptoms and interacts with the agent based on the implicit symptoms.

In addition, the internal critic will generate a reward to the master and worker due to the action and the dialog status. In the higher level, the dialog session will be terminated as successful and get a positive reward if the agent makes the correct diagnosis, or failed if the informed disease is incorrect or the dialog reaches the maximal turn T. In the lower level, a worker is terminated as successful when a correct symptom is requested by the agent and failed when the number of turns reaches the upper limit of subtask turn T^{sub}.

2.3 Implementation details

To group diseases with similar symptoms, we formulate the dataset as a disease-symptoms vector, which represents the number of times that a symptom occurs in each disease. Then, we compare the similarities of the vectors on the training set and divide the diseases with higher similarity (above 0.5) into one group.

Also, to solve the problem of sparse action space and force the master to choose an efficient worker, we follow (Peng *et al.*, 2018) and use the reward shaping method to add auxiliary reward to the original extrinsic reward while keeping the optimal reinforcement learning policy unchanged.

In practice, the ϵ for the master and all the workers are all set to 0.1. For the master, the maximal dialog turns T is set to 10, it will receive an extrinsic reward of $+20$ if the master informs the right disease. Otherwise, it will receive an extrinsic reward of -100 if the dialog turn reaches the maximal turn. In other states, the extrinsic reward is 0. Moreover, the sum of the extrinsic rewards (after reward shaping) over one subtask will be the reward for the master. The maximal dialog turns T^{sub} is set to 2 for each worker.

For the master and all the workers, the neural network of DQN is a three-layer network with two dropout layers and the size of the hidden layer is 512, learning rate for the DQN network is set to 0.0005. All parameters are set empirically and settings for the datasets are the same. For the disease classifier, the neural network is a two-layer network with a dropout layer. Moreover, its trained every epoch during the training process of the master.

During the training process, it will take about 5000 epochs for the model to reach convergence, which takes about 18 h given an NVIDIA RTX 2080 Ti. For the best-performing model, γ is set to

0.95, discounted factor γ_w is set to 0.99, λ in reward shaping is set to +1.

3 Experiment implementation

We evaluate all methods on three dialog datasets [*MZ-4* (Wei *et al.*, 2018), *MZ-10* and *Dxy* (Xu *et al.*, 2019)] collected in the real environment and one synthetic dataset *SymCat-SD-90*. We construct MZ-10 as an expansion on the basis of MZ-4 by including more diseases and samples of patients. Table 1 shows the details of all datasets used in this article.

Also, we evaluate the performance of different methods in terms of disease and symptoms. In terms of disease, we use the accuracy of disease judgment as an indicator (*Acc.*). In terms of symptoms, we use match rate of symptoms (*M.R.*) which is the ratio of the number of corrected recalled implicit symptoms to the total number of implicit symptoms. At the same time, we report the average number of turns (*Avg. T*) conducted for dialog sessions for reference.

3.1 Real-world dataset

MZ-4. This is the first dataset collected from a real environment for the evaluation of a task-oriented dialog system (Wei *et al.*, 2018). It includes 4 diseases, 230 symptoms and 1733 user goals. Each user record consists of the self-report provided by the user and the conversation text between the patient and a doctor. Symptoms extracted from self-report are treated as explicit symptoms and the ones extracted from the conversation are implicit symptoms. The raw data are collected from the pediatric department on a Chinese online healthcare community (http://muzhi.baidu.com), and annotators will follow the begin-in-out (BIO) schema for symptom identification. After that, experts manually link each symptom expression to a concept on SNOMED CT (https://www.snomed.org/snomed-ct).

Dxy. A dialog medical dataset (Xu *et al.*, 2019) contains data from a Chinese online healthcare website (https://dxy.com/). They annotate five types of diseases, including allergic rhinitis, upper respiratory infection, pneumonia, children hand–foot–mouth disease and pediatric diarrhea. Also, they extract the symptoms and normalize them into 41 symptoms. This dataset contains 527 user goals, including 423 for training and 104 for testing.

MZ-10. It is expanded from MZ-4 to include 10 diseases, consisting of typical diseases of the digestive system, respiratory system and endocrine system. Following (Wei *et al.*, 2018), we collect medical consultation records for 10 pediatric diseases. Then, we annotate the samples to form the dataset. Based on the BIO schema, we tagged each symptom with an extra label: Positive, Negative, or Not Sure. Besides, we link all the symptoms to the most relevant concept on SNOMED-CT for normalization. For labeling, we developed a web-based tool and recruited undergraduates and postgraduates in medical school to annotate the corpus. All the annotators are people who are willing to participate and are over the age of 18. Each dialog is annotated twice and inconsistent parts are further finalized by a third annotator. The kappa coefficient of symptom labels is 92.71%, which represents a high consistency between the two annotations.

Table 1. Overview of datasets

Name	♯ of user goal	♯ of diseases	Avg. ♯ of im. sym.	♯ of sym.
MZ-4	1733	4	5.46	230
MZ-10	4116	10	6.60	331
Dxy	527	5	1.67	41
SymCat-SD-90	30 000	90	2.60	266

Note: We report the number of user goal, the number of diseases, the average number of implicit symptoms in each sample and the total number of symptoms in the dataset.

3.2 Synthetic dataset

The number of diseases and user goals in the real-world dataset is still limited. To show the effectiveness of the HRL method, we build a synthetic dataset (SD) following (Kao *et al.*, 2018) for further analysis of the method, named *SymCat-SD-90*. It is constructed based on a symptom-disease database called SymCat (www.symcat.com). There are 801 diseases in the database, and we classify them into 21 departments (groups) according to the International Classification of Diseases (ICD-10-CM) (https://www.cdc.gov/nchs/icd/). We choose 9 representative departments from the database, each department contains the top 10 diseases according to the occurrence rate in the Centers for Disease Control and Prevention (CDC) database.

In the SymCat database, each disease is linked with a set of symptoms, where each symptom has a probability indicating how likely the symptom is identified for the disease. Given a disease and its related symptoms, the generation of a user goal follows two steps. First, for each related symptom, we sample the symptom based on the probability. Second, a symptom is chosen randomly to be the explicit one (same as symptoms extracted from self-report in RD) and the rest of the true symptoms are treated as implicit ones.

3.3 Models for comparison

We compare our model with some state-of-the-art baselines for disease diagnosis.

Flat-DQN (Wei *et al.*, 2018)**.** This is the first work that treats the dialog-based disease diagnosis as an MDP problem and employs a one-layer policy structure based on DQN to choose actions in each dialog turn.

REFUEL (Peng *et al.*, 2018)**.** This work proposes two tricks to improve the performance of flat-DQN, namely reward shaping and feature rebuilding. Reward shaping aims to encourage the agent to discover positive symptoms more quickly by increasing the reward assigned to a correct symptom inquiry and penalizing incorrect ones. Feature rebuilding is introduced as an auxiliary component in the training process that aims to re-construct ground-truth symptom given the current information.

KR-DS (Xu *et al.*, 2019)**.** This model proposes to use external knowledge to further improve the diagnosis performance of DQN. The overall framework contains three components, namely a DQN-based policy network, a refinement module based on the co-occurrence relationship between diseases and symptoms and a conditional probability matrix based on a knowledge graph. The final decision is made by integrating output from the three components to choose actions in each dialog turn.

GAMP (Xia *et al.*, 2020)**.** This model integrates the generative adversarial network (GAN) with the reinforcement learning model. The DQN-based policy network is employed the generator to choose the action and a discriminator is trained to determine how good the action is with a discriminative reward. Moreover, an independent disease classifier is used to evaluate the contribution of the generated symptom (i.e. action) with a mutual information reward. Both rewards are combined as the final reward to update the policy network.

HRL-pretrained (Kao *et al.*, 2018)**.** This model utilizes a hierarchical policy structure of two levels. There are two major differences between their model and ours. Firstly, the master in the higher level and workers in the lower level are trained separately in their setup while we jointly train the master and workers to enforce the interaction between the two components. Secondly, the diagnosis of the disease is handed over to workers in their setup while we introduce an additional disease discriminator that helps to allow workers to focus on symptoms. To some extent, this model can be treated as a pipeline training version of our model.

In Supplementary material, we use a table to show the main differences between models. All these above-mentioned methods have not disclosed their codes. For a fair comparison and to encourage the follow-up researchers, we reproduce their algorithms and release a toolkit (https://github.com/FudanDISC/DISCOpen-MedBox-DialoDiagnosis). It can be used as the benchmark for dialog-based disease diagnosis.

Table 2. Overall performance on real-world datasets

	Dxy			MZ-4			MZ-10		
Model	Acc.	M.R.	Avg. T	Acc.	M.R.	Avg. T	Acc.	M.R.	Avg. T
Flat-DQN	0.731	0.110	1.96	0.681	0.062	1.27	0.408	0.047	9.75
REFUEL	0.721	0.186	3.11	0.716	0.215	5.01	0.505	0.262	5.50
KR-DS	0.740	0.399	5.65	0.678	0.177	4.61	0.485	0.279	5.95
GAMP	0.731	0.268	2.84	0.644	0.107	2.93	0.500	0.067	1.78
HRL (w/o grouped)	0.731	0.297	6.61	0.689	0.004	2.25	0.540	0.114	4.59
HRL (w/o discriminator)	–	**0.512**	8.42	–	**0.233**	5.71	–	**0.330**	8.75
HRL (ours)	**0.779**	0.424	8.61	**0.735**	0.229	5.08	**0.556**	0.295	6.99
Classifier lower bound	0.682	–	–	0.671	–	–	0.532	–	–
Classifier upper bound	0.846	–	–	0.755	–	–	0.612	–	–

Note: We conduct each experiment five times, and the reported number is the average. To make the results comparable, we keep all settings except the agent policy the same. Numbers in **Bold** are the best in each column.

–, missing numbers; Acc., *Accuracy*; M.R, *Match Rate*; Avg. T, *Average Turns*.

Table 3. Overall performance on SymCat-SD-90 dataset

Model	Acc.	M.R.	Avg. T
Flat-DQN	0.343	0.023	1.23
KR-DS	0.357	0.388	6.24
REFUEL	0.347	0.161	4.56
GAMP	0.267	0.077	1.36
HRL-pretrained	0.452	–	3.42
Ours	**0.504**	**0.495**	6.48
Classifier lower bound	0.308	–	–
Classifier upper bound	0.781	–	–

Note: We conduct all experiments five times, and the reported number is the average. Numbers in **Bold** are the best in each column.

–, missing numbers; Acc., *Accuracy*; M.R, *Match Rate*; Avg. T, *Average Turns*.

4 Results and discussions

We report the performance of different approaches on real-world datasets and the performance on the synthetic dataset. Then, we perform the ablation study to evaluate the effectiveness of different components in the model. After that, we present two analysis results to reveal the stability of the disease classifier and the efficiency of workers.

4.1 Performance on real-world datasets

In this section, we test all methods on three real-world datasets and compare the effectiveness of different models in terms of accuracy, average turns and match rate. Table 2 shows the overall results. We have the following findings.

- *Flat-DQN* has limited ability to extract symptom information. Model performance drops significantly as the number of symptoms rises.
- *REFUEL* performs better than *Flat-DQN* in symptom extraction and also has a slight improvement in disease diagnosis. This proves the effectiveness of reward shaping and feature rebuilding.
- Compared with other methods, *KR-DS* maintains a higher match rate and accuracy on the three datasets. This indicates the effectiveness of introducing external knowledge to the RL-based agent.
- The symptom extraction ability of *GAMP* declines significantly with the increase in the number of symptoms, which also limits its performance.
- Our proposed model *HRL* generates the best performance in symptom extraction and disease classification among all the models. This confirms the effectiveness of our proposed hierarchical model.

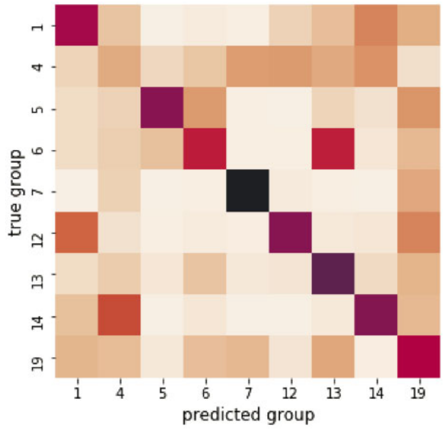


Fig. 3. The error analysis for the disease classifier. The square with true group *i* and predicted group j means a disease in group *i* is misclassified into group *j*, the darker the color, the higher the value

4.2 Performance on synthetic dataset

In order to further prove the effectiveness of our model, we conduct the experiment on a synthetic dataset. Experiment results of different approaches on synthetic datasets are shown in Table 3. Through the results, we can get the following findings.

- It is difficult for models with a flat policy structure to generate a good results on this dataset. *Flat-DQN, KR-DS, REFUEL* and *GAMP* perform similarly to each other. Besides, both *Flat-DQN* and *GAMP* tend to make the diagnosis with a very short length of dialog process. - The performance of *HRL-pretrained* is much better than that of other methods designed with flat and monolithic structure. This indicates the effectiveness of the hierarchical framework.
- Our proposed model *HRL* generates the best performance among all the models. This confirms the stability of our model.

4.3 Ablation studies

The result of the ablation study is shown in Table 2. We compare several versions of HRL models. We remove the master and merge diseases into a single group to form a model with flat policy structure [denoted as HRL (w/o grouped)]. Besides, we remove the separate disease discriminator and hand over the diagnosis action to workers [denoted as HRL (w/o discriminator)]. Experiment results show that HRL (ours) performs better than HRL (w/o

Table 4. The performance of different workers in SymCat-SD-90 dataset

Group id	Success rate (%)	Ave intrinsic reward	Match rate (%)	Activation times
1	48.6	0.031	16.74	0.615
4	54.6	−0.150	5.02	0.375
5	38.8	−0.013	7.96	3.252
6	48.0	−0.036	9.58	0.942
7	48.3	0.057	18.57	1.280
12	43.0	0.021	11.26	0.666
13	52.4	−0.138	7.18	0.823
14	72.2	−0.111	3.77	0.614
19	47.4	0.031	22.72	1.124
Average	50.3	−0.041	10.49	1.077

discriminator) and HRL (w/o grouped). This indicates the effectiveness of extra disease classifier and the hierarchical policy structure. It is noteworthy that after we remove the disease classifier, the match rate of HRL (w/o discriminator) performs better than HRL (ours) on all three real-world datasets. This may be because the reward for correctly predicting disease is higher than the reward for correctly predicting symptoms. This makes the model tend to make a final diagnosis rather than a symptom inquiry when it is confident enough, causing the model to ignore some symptoms that have less impact on the result. Therefore, the match rate of HRL will be smaller than the version without the disease discriminator.

4.4 Stability of disease classifier

To have a deeper analysis of the user goals which have been informed of the wrong disease by the agent, we collect all the wrong informed user goals and present the error matrix in Figure 3. It shows the disease prediction result for all the nine groups. We can see the color of the diagonal square is darker than the others, which means the wrongly predicted disease and the correct disease are in the same group. This is reasonable because diseases in the same groups usually share similar symptoms and are therefore difficult to be distinguished. On the other hand, it also proves that even if the model cannot make correct predictions, it can still assist in real consultations. From ICD-10-CM, Group 7 (Diseases of the eye and adnexa) is prone to misjudgment within the group. For misjudgments between groups, the most likely ones are Group 4 (Endocrine, nutritional and metabolic diseases) and Group 14 (Diseases of the genitourinary system), Group 6 (Diseases of the nervous system) and Group 13 (Diseases of the musculoskeletal system and the connective tissue).

It should be pointed out that although we say ‘the disease classifier can assist in real consultations’, this does not mean that the treatment for these diseases is the same. Diseases with similar symptoms may have completely different treatment options. In practice, we still need professional physicians to make the final decision. However, we believe that when the model can include more information (such as medical examinations, past medical history, etc.), more accurate judgments can be made.

4.5 Efficiency of different workers

While proving the effect of the disease classifier, we also hope that the workers in each group can also play a positive role during the judgment. We evaluate the performance of workers in terms of success rate, average intrinsic rewards and match rate. The results can be seen in Table 4. We found most of the workers can successfully exit by querying the symptoms that which patient suffered. It proves that workers in different groups can learn the symptom characteristics of the group and use the knowledge to guide the consultation process.

5 Conclusion

In this work, we formulate the problem of disease diagnosis as a hierarchical policy learning problem, where symptom acquisition and disease diagnosis are assigned to different kinds of workers at the lower level of the hierarchy. The experimental results on all datasets demonstrate that our hierarchical model outperforms other RL-based models in both disease accuracy and symptom recall. Since the input only contains symptom information, and the reinforcement learning model is sometimes not stable enough to have a statistically unbiased estimate of future expectations, the model cannot make a completely accurate diagnosis. But for now, we still believe that hierarchical architecture is the most reasonable structure in this field. At the same time, this method can also be integrated with other methods, such as the knowledge graph method, or adopting more advanced reinforcement learning techniques for master and worker. In the future, we would like to explore the dense representation of symptoms and diseases to improve the ability of generalization for automatic diagnosis.

Funding

This work was partially supported by Natural Science Foundation of China [71991471 and 6217020551], Science and Technology Commission of Shanghai Municipality Grant [20dz1200600 and 21QA1400600] and Zhejiang Lab [2019KD0AD01].

Conflict of interest: none declared.

References

Budzianowski,P. *et al.* (2017) Sub-domain modelling for dialogue management with hierarchical reinforcement learning. In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue.* ACL, pp. 86–92.

Chen,Y.E. *et al.* (2019) Effective medical test suggestions using deep reinforcement learning. *arXiv preprint arXiv:1905.12916.*

Coronato,A. *et al.* (2020) Reinforcement learning for intelligent healthcare applications: a survey. *Artif. Intell. Med.*, **109**, 101964.

Cuayáhuitl,H. *et al.* (2010) Evaluation of a hierarchical reinforcement learning spoken dialogue system. *Comput. Speech Lang.*, **24**, 395–429.

Doshi-Velez,F. *et al.* (2013) Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics,* **33**, e54–e63.

Duan,J. *et al.* (2020) Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data. *IET Intell. Transport Syst.*, **14**, 297–305.

Feng,J. *et al.* (2018) Relation mention extraction from noisy data with hierarchical reinforcement learning. *arXiv preprint arXiv:1811.01237.*

Florensa,C. *et al.* (2017) Stochastic neural networks for hierarchical reinforcement learning. *arXiv preprint: arXiv:1704.03012.*

Ghavamzadeh,M. (2005) Hierarchical reinforcement in continuous state and multi-agent environments. *Technical report*, MASSACHUSETTSUNIV AMHERST GRADUATE SCHOOL.

Guo,J. *et al.* (2018) Long text generation via adversarial training with leaked information. In: Thirty-Second AAAI Conference on Artificial Intelligence, *Vol. 32*. AAAI.

Hou,Z. *et al.* (2021) Imperfect also deserves reward: Multi-level and sequential reward modeling for better dialog management. *arXiv preprint arXiv: 2104.04748.*

Jonnalagadda,S.R. *et al.* (2017) Text mining of the electronic health record: an information extraction approach for automated identification and subphenotyping of HFpPEF patients for clinical trials. *J. Cardiovasc. Transl. Res.*, **10**, 313–321.

Kao,H.-C. *et al.* (2018) Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. AAAI.

Lin,J. *et al.* (2020) Towards a reliable and robust dialogue system for medical automatic diagnosis.

Lipton,Z. *et al.* (2018) BBQ-networks: efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. **32**. AAAI.

Parr,R. and Russell,S.J. (1998) Reinforcement learning with hierarchies of machines. In: *Advances in Neural Information Processing Systems.* MIT Press.

Peng,Y.-S. *et al.* (2018) Refuel: exploring sparse features in deep reinforcement learning for fast disease diagnosis. In: Advances in Neural Information Processing Systems, Vol. **31**. MIT Press, pp. 7322–7331.

Richens,J.G. and Buchard,A. (2022) Artificial intelligence for medical diagnosis. In: *Artificial Intelligence in Medicine*. Springer, pp. 181–201.

Schatzmann,J. *et al.* (2007) Agenda-based user simulation for bootstrapping a POMDP dialogue system. In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. ACL, pp. 149–152.

Shivade,C. *et al.* (2014) A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Inform. Assoc.*, **21**, 221–230.

Sutton,R.S. *et al.* (1999) Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. In: *Artificial Intelligence*. Elsevier.

Takanobu,R. *et al.* (2018) A hierarchical framework for relation extraction with reinforcement learning. *arXiv preprint arXiv:1811.03925*.

Tang,K.-F. *et al.* (2016) Inquire and diagnose: neural symptom checking ensemble using deep reinforcement learning. In: *Proceedings of NIPS Workshop on Deep Reinforcement Learning*. MIT Press.

Teixeira,M.S. *et al.* (2021) The interplay of a conversational ontology and ai planning for health dialogue management. In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. pp. 611–619.

Wan,G. *et al.* (2020) Reasoning like human: hierarchical reinforcement learning for knowledge graph reasoning. In: *IJCAI*. Morgan Kaufmann, pp. 1926–1932.

Wang,X. *et al.* (2018) Video captioning via hierarchical reinforcement learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 4213–4222.

Wei,Z. *et al.* (2018) Task-oriented dialogue system for automatic diagnosis. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. **2**. ACL, pp. 201–207.

Xia,Y. *et al.* (2020) Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In: *AAAI*, Vol. **34**. AAAI, pp. 1062–1069.

Xu,L. *et al.* (2019) End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In: *AAAI*, Vol. **33**. AAAI, pp. 7346–7353.

Yu,C. *et al.* (2021) Reinforcement learning in healthcare: a survey. *ACM Comput. Surv.*, **55**, 1–36.

Zhang,J. *et al.* (2018) Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. ACL, pp. 140–150.

Zhang,J. *et al.* (2019) Hierarchical reinforcement learning for course recommendation in MOOCs. *Psychology*, **5**, 5–65.