Task-oriented Dialogue System for Automatic Diagnosis

Qianlong Liu[1], Zhongyu Wei[1], Baolin Peng[2], Xiangying Dai[3], Huaixiao Tou[1], Ting Chen[1], Xuanjing Huang[4], Kam-fai Wong[2,5]

1 School of Data Science, Fudan University, Shanghai, China
2 The Chinese University of Hong Kong, Hong Kong
3 Baidu Inc., China
4 School of Computer Science, Fudan University, Shanghai, China
5 MoE Key Lab of High Confidence Software Technologies, China

Introduction

Our research aims to develop an task-oriented dialogue system that make the diagnosis for patients automatically, which can converse with patients to collect additional symptoms.

- Most works involving automatic diagnosis is based on electronic health records (EHRs), which is very expensive to collect.
- Task-oriented dialogue system (DS) has been well researched and reached a promising performance in some specific tasks.
- The cost of collecting data from patients will be reduced greatly by applying DS to medical domain.

Contributions:

- We annotate the first medical dataset for dialogue system.
- We proposed a reinforcement learning based framework for medical DS.

Dataset

- Collected from the pediatric department in a Chinese online healthcare community.
- Three annotators are invited to label all the symptom phrases in both self-reports and conversational data.

Self-report

宝宝嗓子有痰，腹泻并伴有拉水的症状。请问要吃什么药?
The little baby get sputum in throat and have watery diarrhea.
what kind of medicine needs to be taken?

Conversation

……
Doctor: 宝宝现在咳嗽拉肚子吗?
Does the baby have a cough or diarrhea now?
Patient: 不咳嗽，拉肚子。
No cough, but diarrhea.
Doctor: 平常呛奶吗?
Does the baby choking milk?
Patient: 偶尔会吐奶。
He vomits milk sometimes.
……

Symptom Extraction:

- Each Chinese character is assigned a label of “B”, “I” or “O”.
- Each extracted symptom expression is tagged with *True* or *False* indicating whether the patient suffers from this symptom or not.
- The Cohen’s kappa coefficient between annotators are 71% and 67% for self-reports and conversations respectively.

Symptom Normalization:

- Each symptom expression is linked to the most relevant concept on SNOMED CT for normalization.
- User goals are derived from user records.

Extracted symptom expression	Related concept in SNOMED CT
咳嗽(cough)	咳嗽(cough)
喷嚏(sneez)	打喷嚏(sneezing)
鼻涕(cnot)	鼻流涕(cnot)
拉肚子(have loose bowels)	腹泻(diarrhea)
温 度37.5-37.7之 间(body tempera-ture between 37.5-37.7)	低热(low-grade fever)


```
{
    "disease_tag": "小儿支气管炎(children's bronchitis)",
    "request_slots": {
        "disease": "unknow"
    },
    "explicit_symptoms": {
        "咳嗽(cough)": true,
        "鼻流涕(snot)": true
    },
    "implicit_symptoms": {
        "咽痛(sore throat)": true,
        "发烧(fever)": true,
        "粗糙呼吸音(harsh breath sounds)": false,
        "呕吐(emesis)": false
    }
}
```


- *disease_tag* is the disease that the user suffers.
- *explicit_symptoms* are symptoms extracted from the user self-report.
- *implicit_symptoms* are symptoms extracted from the conversational data.
- *request_slots* is the disease slot that the user would request.

Figure 1. An example of user goal

Disease	User goal #	Ave # of explicit symptoms	Ave # of implicit symptoms
Infantile diarrhea	200	2.13	2.71
Children functional dyspepsia	150	1.70	3.20
Upper respiratory infection	160	2.56	3.55
Children’s bronchitis	200	2.87	3.64

Table 1. Overview of the dataset

Proposed Framework

User Simulator

- Sampling a user goal from the experiment dataset to initiate a dialogue session.
- Taking one of the three actions including *True*, *False* and *not_sure*.
- The dialogue session will be terminated as successful by the user if the agent informs correct disease. Otherwise it will be terminated as failed.

Dialogue system

- Both natural language understanding and natural language generator are implemented with template-based models.
- Dialogue state consists of the symptoms requested by the agent and informed by the user till the current time *t*, the previous action of the user, the previous action of the agent and the turn information.
- An action is composed of a dialogue act and a slot.
- The dialogue policy is trained via DQN.
- ε-greedy and experience replay are applied.

Experiments and Results

Experimental setup:

- The maximum dialogue turn is 22.
- The reward for successful and failure dialogue session are +44 and -22 respectively.
- A step penalty of -1 for each turn is applied.
- 80% of the user goals for training and 20% for testing.

Metrics:

- success rate, average reward, average number of turns per dialogue session.

Baseline:

- SVM: takes the automatic diagnosis as a multi-class classification problem.
  - SVM-ex&im: takes both explicit and implicit symptoms as input.
  - SVM-ex: takes only explicit symptoms to predict the disease.
- Random agent: takes an action randomly at each turn.
- Rule-based agent: takes an action based on handcrafted rules.

Disease	SVM-ex&im	SVM-ex
Infantile diarrhea	0.91	0.89
Children functional dyspepsia	0.34	0.28
Upper respiratory infection	0.52	0.44
Children’s bronchitis	0.93	0.71
Overall	0.71	0.59

Table 2. Accuracy of classification models

Model	Success	Reward	Turn
Random Agent	0.06	-24.36	17.51
Rule Agent	0.23	-13.78	17.00
DQN Agent	0.65	20.51	5.11

Table 3. Performance of dialogue system

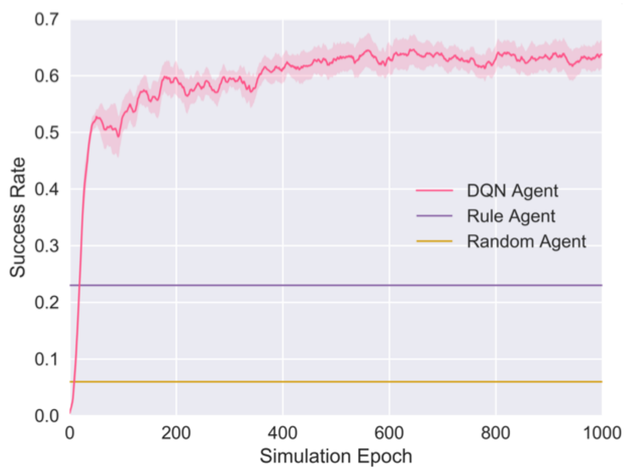

Figure 2. Learning curve of policy learning

- The implicit symptoms can greatly improve the accuracy of disease identification.
- Our rule-based agent is well designed and outperforms the random agent greatly.
- DQN agent outperforms SVM-ex by collecting additional implicit symptoms.
- The gap between DQN agent and SVM-ex&im indicates that there is still rooms for the improvement of the dialogue system.