Towards Building Automatic Medical Consultation System: Framework, Task and Dataset

Anonymous ACL submission

Abstract

In this paper, we propose two frameworks to support automatic medical consultation, namely doctor-patient dialogue understanding and diagnosis-oriented interaction. A new medical dialogue dataset with multi-level fine-grained annotations is introduced and five evaluation tasks are established, including medical named entity recognition, dialogue act classification, symptom recognition, medical report generation and diagnosis-oriented dialogue system. We report a set of benchmark results for each track, which shows the usability of the dataset and sets a baseline for future studies.

1 Introduction

Online medical consultation has shown great potential in improving the quality of healthcare services while reducing cost (Al-Mahdi et al., 2015; Singh et al., 2018), especially in the era of raging epidemics such as Coronavirus. This fact has accelerated the emergence of online medical communities such as SteadyMD and Haodafu. These platforms provide an environment for doctors and patients to communicate with each other via textual messages and images. Figure 1 demonstrates a doctor-patient dialogue record.

Recently, researchers have paid attention to developing automatic approaches to facilitate online consultation service. Research topics include medical entity recognition (Zhou et al., 2021), drug recommendation (Zheng et al., 2021), automatic diagnosis (Chen et al., 2020), question answering (He et al., 2020), medical report generation (Zhang et al., 2020) and dialogue system (Wei et al., 2018). Although progresses have been made to support online consultation from different perspectives, there is still a large gap between existing work and real application. There are three major limitations. (1) Lack of systematical frameworks for automatic medical consultation. (2) Lack of unified design of tasks. (3) Lack of benchmark datasets to support the development of research and application.

In this paper, we make the first step to build a framework for automatic medical consultation and propose several tasks to cover the entire procedure. Two modes of frameworks are proposed to support both static and dynamic scenarios, namely, doctor-patient dialogue understanding and diagnosis-oriented interaction. Understanding framework takes the entire doctor-patient dialogue record as input and aims to generate some labels to support medical diagnosis. Interaction framework follows the setting of task-oriented dialogue system (Wei et al., 2018) plays the role of agent to collect symptoms from the patient and provide professional suggestions and diagnosis. We build a corpus with multi-level annotations to support the research and application development of these five tasks. (1) Self-reported symptoms. (2) Doctor/patient consultation. (3) Doctor's physical examination for the baby. (4) Doctor's suggestion to give the child a stool routine examination. (5) Doctor's diagnosis of the baby.

Table 1: An example of the doctor-patient dialogue record. It consists of the self-report of patient, the dialogue plain text and disease diagnosis result.

---

4https://www.steadymd.com
5https://www.haodf.com/
tasks under the two modes. We conduct a comprehensive analysis of our corpus and tasks to show great future opportunities. Some baseline results are shown for references. Both the corpus and baseline implementation codes will be published.

2 Automatic Medical Consultation Tasks

We introduce our framework and tasks in this section. For dialogue understanding, we propose four tasks including medical entity recognition, dialogue act classification, symptom recognition and medical report generation. For interaction, we introduce diagnosis-oriented dialogue system.

2.1 Notation

Suppose $T = \{T_i\}_{i=1}^{|T|}$ is a piece of dialogue. It consists of three parts - self-report (SR), dialogue (DL) and disease diagnosis (DD). $n_i = |T_i|$ represents the number of utterance in $T_i$. $T_i^{(u)} = \{T^{(a)}_i\}_{j=1}^{m(u)}, u = 0, \ldots, n_i$ stands for the $u$-th utterance in the $i$-th dialogue which consists of $m(u)$ tokens and $D_i = d_i$ represent the result of disease diagnosis for the $i$-th dialogue. For simplicity, $T_i^{(0)}$ stands for the self-report. In addition, we define a unified symptom dictionary $S = \{s_i\}_{i=1}^{|S|}$.

Each token in the utterance might be specific entities. $y^{(u),j}_i$ is the label corresponding to the $j$-th token of the $u$-th utterance in the $i$-th dialogue. $Y^{(u)}_i$ is dialogue action of the $u$-th utterance in the $i$-th dialogue. And $E_i = \{e_i^1 : a_i^1, e_i^2 : a_i^2, \ldots\}$ is the entity attributes of $i$-th dialogue, where $e_i^j \in S$ is symptom name and $a_i^j$ is the corresponding status. Furthermore, $U_i = \{u_i^{(a)}\}_{j=1}^{|E_i|}$ stands for the medical report summarized from SR and DL.

2.2 Doctor-Patient Dialogue Understanding

Medical Named Entity Recognition MNE recognition requires dialogue plain text $\{T^{(a)}_i\}_{u=1}^{n_i}$ as input, and prediction is based on the token level, namely $y^{(u)}_i = \{\hat{y}^{(u),j}_i\}_{j=1}^{m(u)}, u = 1, \ldots, n_i$.

Dialogue Act Classification DA classification requires dialogue plain text $\{T^{(a)}_i\}_{u=1}^{n_i}$ as input with utterance-level action tag prediction $\hat{Y}_i^{(u)}, u = 1, \ldots, n_i$.

Symptom Recognition Symptom recognition is an entity linking with attributes classification task in our setting. It requires self-report along with dialogue plain text $\{T^{(a)}_i\}_{u=0}^{n_i}$ as input with a predicted list $\hat{E}_i = \{\hat{e}_i^1 : \hat{a}_i^1, \hat{e}_i^2 : \hat{a}_i^2, \ldots\}$.

Medical Report Generation MR generation is a text generation task, which takes both self-report and dialogue plain text $\{T^{(a)}_i\}_{u=0}^{n_i}$ as input and a series of medical summary $\hat{U}_i$ as output.

2.3 Diagnosis-oriented Interaction

The diagnosis-oriented dialogue system is designed to simulate the process of a doctor’s diagnosis during conversations. For the doctor, the purpose of the dialogue is to request the patient for enough symptoms to make disease diagnosis. The whole dialogue is abstracted as a sequence of entities (Wei et al., 2018). The diagnosis-oriented dialogue system takes the sequence of EA $E_i$ as input and the output is the predicted disease $\hat{D}_i$. In particular, $E_i = E_i^{ex} \cup E_i^{im}$ where $E_i^{ex}$ is the explicit symptoms extracted from SR and $E_i^{im}$ is the implicit symptoms extracted from the DL.

3 Medical Dialogue Corpus: DialoIMC

The raw doctor-patient conversations are collected from a Chinese online health community that provides professional medical consulting service to patients by doctors with certification. We collect fine-grained annotations on top of MCRs to form our corpus DialoIMC. Several experts with medical background help us design the annotation scheme with consideration of actual scene of online consultation. We include detailed annotated sample and explanation of different labels in the appendix.

3.1 Annotation Scheme

Medical Named Entity (MNE) We define 5 categories of medical named entities, i.e., symptom, drug name, drug category, examination and operation. Among them, drug name represents a specific drug name while drug category represents a class of drugs with a certain efficacy. Inside–outside–beginning (BIO) (Ramshaw and Marcus, 1999) tagging scheme is employed and results in 11 possible tags for tokens. We assign an initial label to each sentence using a rule-based algorithm (Aho and Corasick, 1975) to prompt the annotation process.

Dialogue Act (DA) Dialogue act can be broadly divided into two big categories: request (R) and inform (I), one means "ask the other for information", and another means "tell the other the information". We further categorize the content of information as follows:

1. patient needs medical help and asks for help
2. patient describes symptoms
3. doctor asks patient for symptoms
4. doctor asks patient for medical history
5. doctor asks patient for background information
6. doctor asks patient for treatment history
7. patient expresses gratitude
8. patient asks for a consultation
9. patient asks for an appointment
10. patient asks for a prescription
11. patient asks for a referral

http://muzhi.baidu.com
mation conveyed as: physical characteristic (PC), symptom (SX), etiology (ETIOL), existing examination and treatment (EET), medical advice (MA), drug recommendation (DR) and other. There are both request and inform versions for all categories except MD and other. Therefore, there are 16 types of fine-grained dialogue acts in our scheme. In the following, we always use abbreviations to indicate a certain dialog act.

**Entity Attribute (EA)** We focus on the symptom entity and its two attributes: the standardized name (SN) and whether the patient has the symptom (Has). Symptoms are expressed in a variety of ways in utterance, such as verbs, nouns, abbreviations, and aliases. We collect all symptom entities and ask annotators to manually cluster them, resulting 328 standardized names normalized from 1,910 unique symptoms extracted by BIO tag. Further, for each dialogue, we collect all standardized symptoms mentioned in the conversation, and ask annotators to annotate whether the patient has the symptom (Yes, No, or Uncertain) for each symptom.

**Medical Report (MR)** Based on patient’s SR and doctor-patient dialogue, annotators are required to write a report to summarize the consulting case. It contains six parts: 1) chief complaint: patient’s main symptoms or signs; 2) present disease: description of main symptoms; 3) auxiliary examination: the patient’s existing examinations, examination results, records, etc; 4) history of past disease: previous health conditions and illnesses; 5) diagnosis: diagnosis of disease; 6) suggestions: doctor’s suggestions of inspection recommendations, drug treatment and precautions. Annotators are required to construct the report following the format. If some part of information is not mentioned in the case, the annotator would leave it as blank.

## 3.2 Inter-Annotator Agreement

To annotate medical conversations more conveniently, we design a web-based tool which can be used for general-purpose multi-level dialogue annotation tasks. We recruited 10 annotators, all of whom have medical degrees. Two annotations per dialogue were gathered resulting in 168,847 unique turns, and to estimate the inter-annotator agreement, we use Cohen’s kappa coefficient (Banerjee et al., 1999). For medical named entities, dialogue acts and entity attributes (Has), the kappa coefficients are 83.11%, 76.41% are 80.92% respectively; For medical reports, both reports are remained for golden reference.

## 3.3 Corpus Statistics

Samples in the DialoIMC are related to 10 types of pediatric diseases, and contains 4,116 dialogues, with an average of about 42 utterances and 539 words per dialogue. Table 1 shows the comparison between DialoIMC and other medical datasets. Compared with existing datasets in medical scenarios, DialoIMC is highly competitive both in annotation granularity and scale.

The detailed statistics about the annotated content in DialoIMC are shown in Figure 2. The distribution of types of medical named entities and dialogue acts are shown in Figure 2(a) and 2(b). Briefly, symptom entities appear the most, about 58.3%, followed by examination, drug name, drug category, and operation. This indicates that doctor-patient conversations mainly talk about patient-related symptoms.

Similar to entity types, the highest proportion of dialogue acts are I-SX, R-SX, L-DR, L-EET, and so on. Most types of dialog acts come either entirely from doctors or only from patients due to the defined fine-grained classification schema.

Figure 2(c) present the positional characteristics of dialogue acts. We divide utterances in a dialogue into five parts according to their locations.

---

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th># Diseases</th>
<th># Dialogues</th>
<th># Utterances</th>
<th># Entities</th>
<th>Annotation Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>MZ (Wei et al., 2018)</td>
<td>Pediatrics</td>
<td>4</td>
<td>710</td>
<td>-</td>
<td>70</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
<tr>
<td>DX</td>
<td>Pediatrics</td>
<td>5</td>
<td>527</td>
<td>2,816</td>
<td>46</td>
<td>✓</td>
</tr>
<tr>
<td>CMDMD</td>
<td>Pediatrics</td>
<td>4</td>
<td>2,067</td>
<td>87,005</td>
<td>161</td>
<td>✓</td>
</tr>
<tr>
<td>MIE</td>
<td>Cardiology</td>
<td>6</td>
<td>1,120</td>
<td>18,129</td>
<td>71</td>
<td>✓</td>
</tr>
<tr>
<td>MedDG</td>
<td>Gastroenterology</td>
<td>12</td>
<td>17,864</td>
<td>385,951</td>
<td>160</td>
<td>✓</td>
</tr>
<tr>
<td>Ours</td>
<td>Pediatrics</td>
<td>10</td>
<td>4,116</td>
<td>164,731</td>
<td>328/4,092</td>
<td>✓ ✓ ✓ ✓</td>
</tr>
</tbody>
</table>

Table 1: Comparison between DialoIMC and other medical dialogue corpus, where MNE, DA, EA, MR are the abbreviations of Medical Named entity, Dialog Act, Entity Attribute, and Medical Report respectively.
example, 0-20% means the sentences appeared in the first fifth of the conversation. We conclude that with the in-depth of medical consultation, the focus gradually shifts from symptoms to drugs, treatments and precautions.

Figure 2(d) shows the distribution of symptom attribute (Has). Explicit symptoms account for only about 20%, which means that only a small part of relevant symptoms appears in the patient’s SR. For implicit symptoms, No and Not Sure accounted for more than 30%, this indicates that a large proportion of symptoms in the conversation are potentially unrelated to the patient.

A total of 8,232 medical reports are obtained with an average of about 68 words, where the Present disease and Suggestions part has about 30 and 20 words on average respectively.

4 DialoIMC as a New Benchmark

In order to further show the characteristics of DialoIMC, we demonstrate experiment results of some baselines for five tasks. Detailed experiment results are shown in Appendix.

4.1 Medical Named Entity Recognition

We treat it as a sequence labeling task and present some baselines including LSTM (Dyer et al., 2015), BERT-base (Devlin et al., 2018) and BERT-base with CRF. Experiment results show that BERT with CRF generate the best F1 score of 89%. Details are reported in Table 2.

4.2 Dialogue Act Classification

We treat this task as a sentence classification one and use accuracy for evaluation. In terms of models, we try non-pre-trained models represented by TextCNN (Kim, 2014) and DPCNN (Johnson and Zhang, 2017), and pre-trained models represented by BERT (Devlin et al., 2018). We adopt same settings in the training, where the batch size is 128, the epoch is 20, and the learning rate is 1e-5. The accuracy of sentence classification on the test set is reported in Table 3.

4.3 Symptom Recognition & Inference

We treat it as an entity alignment with attributes classification task and use F1 score for evaluation. Two frameworks are set as baselines - a multi-task learning (MTL) method on the basis of NER, and a multi-label classifier based on the whole dialogue. Results are reported in Table 4. The performance of the model based on multi-task learning is slightly better than that of the multi-label classification model, exceeding 72%.

4.4 Medical Report Generation

We treat this task as a text generation one and use ROUGE (Lin, 2004) as the evaluation metric. Three widely used text generators are used as baselines: Seq2Seq with attention mechanism (Nallapati et al., 2016), Pointer generator (See et al., 2017) and BERT-Transformer (Vaswani et al., 2017). The overall results are shown in Table 5. BERT outperform the others with obvious advantages.

4.5 Diagnostic-oriented Dialogue System

We treat this task as a sequence decision task based on reinforcement learning, and use symptom recall and disease classification accuracy as evaluation metrics. We use reinforcement learning systems such as DQN and HRL as the baseline models. Experimental results show that HRL can reach a better performance with disease accuracy at 71.5% and symptom recall at 46.7%. Details are in Table 6.

5 Conclusion

This paper proposes a framework for automatic medical consultation and present a dataset with multiple-level annotations as benchmark. We also demonstrate experiment results of some baselines on the dataset to give an insight about the difficulty of different tasks.
Ethical Statement

In this paper, different ethical restrictions deserve discussion. All the data in our self-constructed corpus are available online. When crawling data from the web platforms, we strictly abide by the platform’s policies and rules. We did not use any author-specific information in our research.

We recruited undergraduates and postgraduates in medical school to annotate our corpus and strictly evaluated each annotating work. The reward for annotating is counted by the number of dialogue that the annotator dealt with. We pay $0.5 for each dialogue. All annotators are people who are willing to participate and over the age of 18.

What we need to declare is that the framework of automatic medical consultation system proposed in this paper is only an assistant role, not a complete replacement for doctors’ face-to-face consultation. When our assistant consultation system presents information that is contrary to medical common sense, it is necessary to attach importance to the judgment of doctors.

References


An example of our corpus with annotations, including named entities, dialogue acts, symptom attributes, and medical record is shown in Figure 3.

B Details of Annotation Scheme

B.1 Token-level Annotations

Token-level annotations mainly served for medical named entity recognition task. There are totally 5 kinds of medical entity in our corpus, namely symptom, drug name, drug category, examination and operation. We followed the widely used BIO tagging scheme. “B” and “I” determine the boundary of an entity, in particular, “B” stands for the beginning of the entity and “I” means inside. So there are totally 11 candidate labels for each token - O, B-symptom, B-drugname, B-drugcategory, B-examination, B-operation and I-symptom, I-drugname, I-drugcategory, I-examination, I-operation. The predicted $\hat{y}_{i,j}$ should be selected in these 11 candidates.

B.2 Utterance-level Annotations

Utterance-level annotations works for dialogue action classification. There are 16 types of fine-grained dialogue acts in our scheme - both request (R) and inform (I) for physical characteristic (PC), symptom (SX), etiology (ETIOL), existing examination and treatment (EET), medical advice (MA), drug recommendation (DR), precautions (PRCTN) and two single dialogue action make diagnosis (MD) and other. The predicted $\hat{y}_{i}$ should be selected in these 16 candidates.

B.3 Dialogue-level Annotations

Report generation, symptom recognition and diagnosis-oriented dialogue system all need the dialogue-level annotations. First, the human annotated medical report summarizes the dialogue in 6 main parts - chief complaint, present disease, auxiliary, past disease history, diagnosis and suggestions. Secondly, human annotators extract the symptoms involved in SR and DL, each symptom has 4 different status, namely not mentioned, no, has, not clear. And lastly, dialogue system will use the sequence of annotated symptoms as request sequence to predict the disease.

C Experimental Results for Different Tasks

C.1 Results for MNE Recognition

Results of medical named entity recognition are shown in Figure 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi-LSTM (Dyer et al., 2015)</td>
<td>80.54</td>
</tr>
<tr>
<td>Bi-LSTM-CRF (Huang et al., 2015)</td>
<td>85.76</td>
</tr>
<tr>
<td>Bi-LSTM-CNN-CRF (Ma and Hovy, 2016)</td>
<td>85.31</td>
</tr>
<tr>
<td>BERT (Devlin et al., 2018)</td>
<td>86.18</td>
</tr>
<tr>
<td>BERT-CRF (Devlin et al., 2018)</td>
<td>89.44</td>
</tr>
</tbody>
</table>

Table 2: Results for medical named entity recognition.

C.2 Results for DA Classification

Results of dialogue action classification are shown in Figure 3.
The baby suffers from diarrhea and it does not improve after taking Medilac-Vita for five days.

**Dialogue**

---

**Doctor**

Is it the stool watery? or with undigested milk disc?

---

**Patient**

No milk disc.

---

**Doctor**

Taking Medilac-Vita does not improve.

---

**Patient**

I suggest to take a stool routine examination for the baby.

---

**Patient**

What medicine can the baby take to relieve?

---

**Doctor**

I suggest to take a stool routine examination for the baby.

---

**Normalisation (Symptom Entity)**

- Watery stool (watery stool)
- Undigested milk disc (loose stool)
- Indigestion (indigestion)

**Attributes (Symptom Entity)**

- Watery stool (not sure)
- Loose stool (no)
- Indigestion (yes)

**Medical Record**

- Chief complaint: diarrhea
- Present disease: the baby has diarrhea and is taking Medilac-Vita now.
- Auxiliary: N.A
- Past disease history: N.A
- Diagnosis: dyspepsia, reasons pending.
- Suggestions: stool routine, note the light diet.

---

**Table 3:** Results for dialogue act classification.

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TextCNN (Kim, 2014)</td>
<td>80.92</td>
</tr>
<tr>
<td>TextRNN (Liu et al., 2016)</td>
<td>80.61</td>
</tr>
<tr>
<td>TextRNN w/ Attn (Zhou et al., 2016)</td>
<td>81.23</td>
</tr>
<tr>
<td>TextRCNN (Lai et al., 2015)</td>
<td>81.76</td>
</tr>
<tr>
<td>DPCNN (Johnson and Zhang, 2017)</td>
<td>79.82</td>
</tr>
<tr>
<td>BERT (Devlin et al., 2018)</td>
<td>82.35</td>
</tr>
</tbody>
</table>

**Table 4:** Results for Symptom Attribute Inference.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAI-MLC</td>
<td>69.89</td>
</tr>
<tr>
<td>SAI-MTL</td>
<td>72.28</td>
</tr>
</tbody>
</table>

**C.3 Results for SRI**

Results of symptom attributes inference are shown in Figure 4.

**Table 5:** Results of Medical Report Generation.

<table>
<thead>
<tr>
<th>Model</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2seq+attention w/o other</td>
<td>58.91</td>
<td>40.88</td>
<td>56.79</td>
</tr>
<tr>
<td>Pointer-generator w/o other</td>
<td>62.67</td>
<td>44.30</td>
<td>57.60</td>
</tr>
<tr>
<td>BERT-Transformer w/o other</td>
<td>63.31</td>
<td>43.82</td>
<td>57.28</td>
</tr>
</tbody>
</table>

**C.4 Results for Report Generation**

Results of medical report generation are shown in Figure 5.

**Table 6:** Results for Diagnostic-oriented Dialogue System.

<table>
<thead>
<tr>
<th>Model</th>
<th>Disease Accuracy (%)</th>
<th>Symptom Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>DQN-Flat</td>
<td>43.333</td>
<td>28.683</td>
</tr>
<tr>
<td>HRL</td>
<td>71.489</td>
<td>46.689</td>
</tr>
</tbody>
</table>