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.

Self-Report

The baby suffers from diarrhea and it does not improve after taking Medilac - Vita for five days
孩子有点拉肚子，吃妈咪爱五天不见好

Dialogue

Doctor	Is it the stool watery? or with undigested milk disc? 孩子是浠水便吗？有未消化的奶瓣吗？
Patient	No milk disc 没有奶瓣
......	
Patient	Taking Medilac-Vita does not improve 吃妈咪爱也没有好转
Doctor	The baby looks like to have indigestion according to the current stool 孩子现在大便看着是得了消化不良
......	
Patient	What medicine can the baby take to relieve 需要吃些什么药能缓解
Doctor	I suggest to take a stool routine examination for the baby 我建议给孩子查个大便常规

Disease Diagnosis: 消化不良 (Indigestion)

Figure 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.

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*[1]. This fact has accelerated the emergence of online medical communities such as *SteadyMD*[2] and *Haodafu*[3]. 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 develop 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

[1] https://en.wikipedia.org/wiki/COVID-19
[2] https://www.steadymd.com
[3] https://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 $\mathcal{T} = \{T_i\}_{i=1}^{|\mathcal{T}|}$ 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. $T_i^{(u)} = \{T_i^{(u),j}\}_{j=1}^{m_{(u)}}, u = 0, \ldots, n_i$ stands for the u-th utterance in the 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 $\mathcal{S} = \{s_i\}_{i=1}^{|\mathcal{S}|}$.

Each token in the utterance might be specific entities. $y_i^{(u),j}$ is the label corresponding to the j-th token of the u-th utterance in the i-th dialogue. $Y_i^{(u)}$ 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 $e_i^j \in \mathcal{S}$ is symptom name and a_i^j is the corresponding status. Furthermore, $U_i = \{u_i^j\}_{j=1}^{|U_i|}$ 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 $\{T_i^{(u)}\}_{u=1}^{n_i}$ as input, and prediction is based on the token level, namely $\hat{\mathbf{y}}_i^{(u)} = \{\hat{y}_i^{(u),j}\}_{j=1}^{m_{(u)}}, u = 1, \ldots, n_i$.

Dialogue Act Classification DA classification requires dialogue plain text $\{T_i^{(u)}\}_{u=1}^{n_i}$ 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 $\{T_i^{(u)}\}_{u=0}^{n_i}$ 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 $\{T_i^{(u)}\}_{u=0}^{n_i}$ 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 are 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, $E_i = E_i^{\text{ex}} \cup E_i^{\text{im}}$ 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[4] 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 infor-

[4] http://muzhi.baidu.com

Dataset	Domain	Annotation Scale				Annotation Granularity			
		# Diseases	# Dialogues	# Utterances	# Entities	MNE	DA	EA	MR
MZ (Wei et al., 2018)	Pediatrics	4	710	-	70			✓	
DX	Pediatrics	5	527	2,816	46		✓		
CMDD	Pediatrics	4	2,067	87,005	161		✓		
MIE	Cardiology	6	1,120	18,129	71				✓
MedDG	Gastroenterology	12	17,864	385,951	160		✓		
Ours	Pediatrics	10	4,116	164,731	328 / 4,692	✓	✓	✓	✓

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.

mation conveyed as: *physical characteristic* (PC), *symptom (SX)*, *etiology (ETIOL)*, *existing examination and treatment (EET)*, *medical advice (MA)*, *drug recommendation (DR)*, *precautions (PRCTN)*, *make diagnose (MD)* 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.

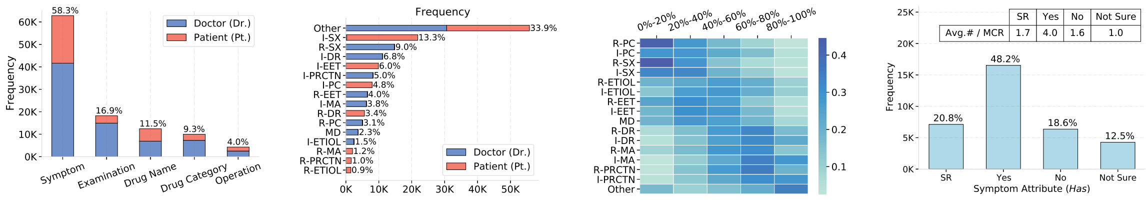
3.3 Corpus Statistics

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, I-DR, I-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. For



(a) # of named entities (b) # of dialogue acts (c) position of dialogue acts (d) # of symptom labels

Figure 2: Statistics of annotations for dialogue acts and medical named entities.

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

Alfred V Aho and Margaret J Corasick. 1975. Efficient string matching: an aid to bibliographic search. Communications of The ACM, 18(6):333–340.

Ibrahim Al-Mahdi, Kathleen Gray, and Reeva Lederman. 2015. Online medical consultation: A review of literature and practice. In Proceedings of the 8th Australasian Workshop on Health Informatics and Knowledge Management, pages 27–30.

Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. 1999. Beyond kappa: A review of interrater agreement measures. Canadian journal of statistics, 27(1):3–23.

Jun Chen, Xiaoya Dai, Quan Yuan, Chao Lu, and Haifeng Huang. 2020. Towards interpretable clinical diagnosis with Bayesian network ensembles stacked on entity-aware CNNs. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3143–3153, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. arXiv: Computation and Language.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4604–4614, Online. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXiv: Computation and Language.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 562–570.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In Twenty-ninth AAAI conference on artificial intelligence.

Chin-Yew Lin. 2004. Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough? In NTCIR.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. CoRR, abs/1603.01354.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:1602.06023.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In Natural language processing using very large corpora, pages 157–176. Springer.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

Ajeet Pal Singh, Hari Shanker Joshi, Arun Singh, Medhavi Agarwal, and Palveen Kaur. 2018. Online medical consultation: a review. International Journal of Community Medicine and Public Health, 5(4):1230.

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.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 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), pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5108–5120, Online. Association for Computational Linguistics.

Zhi Zheng, Chao Wang, Tong Xu, Dazhong Shen, Penggang Qin, Baoxing Huai, Tongzhu Liu, and Enhong Chen. 2021. Drug package recommendation via interaction-aware graph induction. In Proceedings of the Web Conference 2021, pages 1284–1295.

Baohang Zhou, Xiangrui Cai, Ying Zhang, and Xiaojie Yuan. 2021. An end-to-end progressive multi-task learning framework for medical named entity recognition and normalization. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6214–6224, Online. Association for Computational Linguistics.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), pages 207–212.

A A Sample of Annotated Data

An example of our corpus with annotations, including named entities, dialogue acts, symptom normalization, 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 $\hat{y}_i^{(u),j}$ 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 diagnose (MD)* and *other*. The predicted $\hat{Y}_i^{(u)}$ 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 complain, 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.

Model	F1 (%)
Bi-LSTM (Dyer et al., 2015)	80.54
Bi-LSTM-CRF (Huang et al., 2015)	85.76
Bi-LSTM-CNN-CRF (Ma and Hovy, 2016)	85.31
BERT (Devlin et al., 2018)	86.18
BERT-CRF (Devlin et al., 2018)	**89.44**

Table 2: Results for medical named entity recognition.

C.2 Results for DA Classification

Results of dialogue action classification are shown in Figure 3.

Self-Report

The baby suffers from diarrhea and it does not improve after taking Medilac-Vita for five days	diarrhea; Medilac-Vita
孩子有点拉肚子，吃妈咪爱五天不见好	腹泻；妈咪爱

Dialogue

......

Doctor	Is it the stool watery? or with undigested milk disc? 孩 子 是 浠 水 便 吗 ？ 有 未 消 化 的 奶 瓣 吗 ？ 0 0 0 B^s I^s I^s 0 0 0 0 0 0 0 B^s I^s 0 0	*R-SX*
Patient	No milk disc 没 有 奶 瓣 0 0 B^s I^s	*I-SX*

......

Patient	Taking Medilac-Vita does not improve 吃 妈 咪 爱 也 没 有 好 转 0 B^d I^d I^d 0 0 0 0 0	*I-EET*
Doctor	The baby looks like to have indigestion according to the current stool 孩 子 现 在 大 便 看 着 是 得 了 消 化 不 良 0 0 0 0 0 0 0 0 0 0 0 B^s I^s I^s I^s	*MD*

......

Patient	What medicine can the baby take to relieve 需 要 吃 些 什 么 药 能 缓 解 0 0 0 0 0 0 0 0 0 0	*R-DR*
Doctor	I suggest to take a stool routine examination for the baby 我 建 议 给 孩 子 查 个 大 便 常 规 0 0 0 0 0 0 0 0 B^e I^e I^e I^e	*I-MA*

......

Normalization (*Symptom Entity*)	**Attributes** (*Symptom Entity*)
stool watery (watery stool); milk disc (loose stool); indigestion (indigestion)	*watery stool (not sure); loose stool (no); indigestion (yes)*
浠水便（水样便）；奶瓣（稀便）；消化不良（消化不良）	水样便（不确定）；稀便（否）；消化不良（是）

Medical Record

Chief complaint: diarrhea	主诉：腹泻。
Present disease: the baby has diarrhea and is taking Medilac-Vita now.	现病史：患儿腹泻十天，喷射状。现服用妈咪爱。
Auxiliary: N.A	辅助检查：暂缺。
Past disease history: N.A	既往史：不详。
Diagnosis: dyspepsia, reasons pending.	诊断：少儿消化不良，原因待查。
Suggestions: stool routine, note the light diet.	建议：大便常规，注意清淡饮食。

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

Model	Acc. (%)
TextCNN (Kim, 2014)	80.92
TextRNN (Liu et al., 2016)	80.61
TextRNN w/ Att (Zhou et al., 2016)	81.23
TextRCNN (Lai et al., 2015)	81.76
DPCNN (Johnson and Zhang, 2017)	79.82
BERT (Devlin et al., 2018)	82.35

Table 3: Results for dialogue act classification.

C.3 Results for SRI

Results of symptom attributes inference are shown in Figure 4.

Model	F1 Score (%)
SAI-MLC	69.89
SAI-MTL	72.28

Table 4: Results for Symptom Attribute Inference

C.4 Results for Report Generation

Results of medical report generation are shown in Figure 5.

Model	R-1	R-2	R-L
Seq2seq+attention (Nallapati et al., 2016)	58.91	40.88	56.79
w/o other	60.18	42.17	57.23
Pointer-generator (See et al., 2017)	62.67	44.30	57.60
w/o other	62.91	44.41	57.88
BERT-Transformer (Vaswani et al., 2017)	63.31	43.82	57.28
w/o other	64.13	45.64	58.72

Table 5: Results of Medical Report Generation

C.5 Results for Diagnostic-oriented Dialogue System

Results of diagnostic-oriented dialogue system are shown in Figure 6.

Model	Disease Accuracy (%)	Symptom Recall
DQN-Flat	43.333	28.683
HRL	71.489	46.689

Table 6: Results for Disease accuracy & Symptom recall