RealMedDial: A Real Telemedical Dialogue Dataset Collected from Online Chinese Short-Video Clips

Bo Xu[1], **Hongtong Zhang**[1], **Jian Wang**[1], **Xiaokun Zhang**[1], **Dezhi Hao**[1], **Linlin Zong**[2]*, **Hongfei Lin**[1], **Fenglong Ma**[3]
[1]School of Computer Science and Technology, Dalian University of Technology, China
[2]School of Software, Dalian University of Technology, China
[3]College of Information Sciences and Technology, Pennsylvania State University, USA
{xubo, wangjian, llzong, hflin}@dlut.edu.cn
{dutzht, kp5861213, kun}@mail.dlut.edu.cn, fenglong@psu.edu

Abstract

Intelligent medical services have attracted great research interests for providing automated medical consultation. However, the lack of corpora becomes a main obstacle to related research, particularly data from real scenarios. In this paper, we construct RealMedDial, a Chinese medical dialogue dataset based on real medical consultation. RealMedDial contains 2,637 medical dialogues and 24,255 utterances obtained from Chinese short-video clips of real medical consultations. We collected and annotated a wide range of metadata with respect to medical dialogue including doctor profiles, hospital departments, diseases and symptoms for fine-grained analysis on language usage pattern and clinical diagnosis. We evaluate the performance of medical response generation, department routing and doctor recommendation on RealMedDial. Results show that RealMedDial are applicable to a wide range of NLP tasks with respect to medical dialogue.

1 Introduction

The COVID-19 pandemic has dramatically changed how outpatient care is delivered in healthcare practices. To decrease the risk of transmitting the virus to either patients or healthcare workers within their health practice, providers are deferring or selectively prohibiting in-person visits, but fortunately, they are usually converting in-person visits to telemedicine visits (Mann et al., 2020). During the telemedicine, patients describe their symptoms of suffered diseases and/or adverse reactions of the taking drugs to doctors, while doctors provide medical consultations through online video conferences. Although telemedicine is convenient and timely for disease diagnoses, the continuous growth of telemedicine visits significantly increases the burden and workload of doctors, and meanwhile, the health conditions of remote patients become increasingly difficult to be tracked. Thus, how to relieve the burden of doctors and effectively track the patient's health conditions remains an open research question.

Researchers from related fields are trying to solve this issue by developing medical dialogue systems to serve as virtual doctors, which greatly facilitates users to obtain medical and healthcare information. Recent advances in medical dialogue systems have benefited medical applications such as psychological consultation (Das et al., 2022), elderly care (Keshmiri et al., 2019), and disease pre-diagnosis (Nasreen et al., 2021). To build effective medical dialogue systems, related studies are focusing on optimizing medical dialogue from various aspects, including automatic diagnosis (Wei et al., 2018; Xu et al., 2019), medical information extraction (Zhang et al., 2020), medical slot filling (Shi et al., 2020), and medical conversational summarization (Joshi et al., 2020). Although these researches have improved the performance of medical dialogue, this challenging task is still facing great difficulty in generating effective responses due to the particularity and professionalism of the medical field.

In general, several key challenges have not been thoroughly considered in the current medical dialogue datasets. First, as shown in Table 1, most existing medical dialogue datasets extract the data from online medical or healthcare community, which is **non-real time communication** records between doctors and patients. In fact, such data are more similar to question and answering (Q&A) data, instead of medical dialogue. Besides, the static Q&A data are largely different from real-time medical consultations in language expressions and interaction patterns. In real-time medical consultations, doctors make accurate diagnosis predictions not only based on symptom descriptions from patients, but also according to observations of patient health status and medical examination results,

*Corresponding author

Proceedings of the 29th International Conference on Computational Linguistics, pages 3342–3352
October 12–17, 2022.

Dataset Name	Source	#Dialogues	#Utterances	#Diseases	Department
MZ(Wei et al., 2018)	Online health community	710	-	4	Pediatrics
DX(Xu et al., 2019)	Online health community	527	2,186	5	Pediatrics
CMDD(Lin et al., 2019)	Online health community	2,067	87,005	4	Pediatrics
MIE(Zhang et al., 2020)	Online health community	1,120	18,129	6	Cardiology
MedDG(Liu et al., 2020)	Online health community	17,864	385,951	12	Gastroenterology
COVID-EN(Zhou et al., 2021)	-	603	-	1	COVID
COVID-CN(Zhou et al., 2021)	-	1,088	-	1	COVID
MedDialog-EN(Zeng et al., 2020)	Online health community	257,332	514,664	96	29 Departments
MedDialog-CN(Zeng et al., 2020)	Online health community	3,407,494	11,260,564	172	51 Departments
RealMedDial (Ours)	Online short-video clips	2,637	24,255	55	17 Departments

Table 1: Comparison between our dataset and other existing medical dialogue datasets.

which are usually missing in the current medical dialogue datasets. Thus, it is essential to build a real-time medical dialogue dataset, which can be used for developing workable dialogue systems.

Second, when patients use online health community to ask for help from doctors, they usually input their symptoms as detailed as possible. Then doctors predict possible diagnoses based on patients' inputs. Such a working procedure leads to a common shortage of existing medical dialogue data extracted from online health community, that is, they only have **a few communication rounds or utterances**. For instance, the average number of utterances in a dialogue is only 3.3 in the largest Chinese medical dialogue dataset MedDialog-CN (Zeng et al., 2020). In real-world medical consultations, a doctor seldom makes any decisions just based on limited number of interactions with patients. Therefore, such datasets may be not suitable for training a real medical dialogue model.

Third, doctors, especially domain experts, are usually very busy and do not have enough time to answer online questions frequently. In order to attract more patients to use the health community, the companies have to hire graduate students studying in medical schools as online doctors. They will be paid when replying patients' questions. Compared with experts' replies, the quality of the answers from graduate students is usually not very high in some dialogues. The **low quality** issue of existing datasets also impedes the development and learning of medical dialogue models.

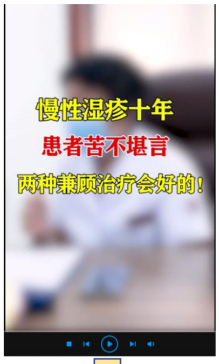
To tackle all the aforementioned limitations, in this paper, we construct a real-time, high-quality, and large-scale medical dialogue dataset named **RealMedDial**, which is extracted from online Chinese short-video clips. In particular, the videos are downloaded from a popular Chinese video-based social media named Kuaishou[1], where many medical physicians record the short-videos when they communicate with their online or offline patients and post them to Kuaishou. Those short-videos are all real-time medical consultations, which are high quality and representative for diagnosing diseases. Moreover, the contents between doctors and patients not only include disease diagnoses but also treatment plans as well as prognoses. We transcribe the real-scenario medical conversations into text, which is used to simulate real doctor-patient consultations for training effective medical dialogue models. Besides, we also extract video titles, doctor profiles, disease, symptoms, and hospital departments. An example is shown in Figure 1.

Compared with existing medical dialogue datasets based on online health community, our dataset also has **two extra advantages**. The first advantage is that it enables us to conduct the *modeling of language usage patterns*. On online health community, doctor-patient conversations are often completed offline. Offline language usage tends to adopt written expressions, which is quite different from the oral expressions in real medical consultation. Since our dataset is realistic starting from the dialogue scenario, dialogue models based on our dataset are more conducive to better modeling the language usage patterns for training a robust response generation model.

The second advantage is *comprehensiveness of information*. The constructed dataset not only contains medical conversations between doctors and patients during clinical consultations. We also collected and annotated a wide range of meta-data with respect to medical conversations including doctor profiles, hospital departments, diseases and symptoms. The meta-data can be used for fine-grained modeling and analysis on medical dialogue. For example, doctor profiles could be incorporated into an expertise-specific dialogue model for personalized response generation. Diseases and symptoms can be used to mine patients' dialogue intents for precise clinical treatments.

The main **contributions** of this work are sum-

[1] https://www.kuaishou.com



Video duration: 58 seconds
Doctor profile: Associate Chief Physician of Dermatology, Doctor of Medicine, with nearly 30 years of experience, specializing in the treatment of common and difficult diseases of skin diseases, such as psoriasis, dermatitis and eczema, vitiligo, alopecia, neuralgia after herpes zoster, and prurigo nodularis , Scleroderma, etc.

Title of the video clip
慢性湿疹十年，患者苦不堪言，相信我两种方法兼顾治疗 (Chronic eczema has been suffering for ten years, and the patient is miserable. Believe me that the two methods are both treatments)
Dialogue
医生：我看你那个手上和脚上都有湿疹是吧？(Doctor: I observe you have eczema on your hands and feet, right?)
病人：是的，慢性的，十年了。(Patient: Yes, chronic, for ten years.)
医生：什么时间能加重？你能找出原因吗？(Doctor: When can it get worse? Can you find out the reason?)
病人：好像就是跟心情有关系，容易焦虑。(Patient: It seems that it has something to do with mood, and I'm easy to be anxious)
医生：那我先给您开中药调着看，可能比较慢但是有效果，这个病很可能是神经性皮炎，但是从肉眼上有的时候分不出来，我用中药给您治一下。(Doctor: I will prescribe Chinese medicine for you. It may be slower but effective. This disease is probably neurodermatitis, but sometimes it cannot be distinguished from the naked eye. I will treat you with Chinese medicine.)
.....
Disease and Symptom
疾病：神经性皮炎 (Disease: Neurodermatitis)
症状：湿疹 (Symptoms: eczema)
Hospital Department
皮肤病科 (Dermatology)

Figure 1: An exemplar offline medical consultation, which includes (1) a video clip with doctor profile, (2) video title, (3) dialogue between doctor and patient, and (4) disease, symptom and department.

marized as follows.

- We construct a large-scale medical dialogue dataset - RealMedDial, which contains (1) 2,637 real-scenario medical conversations from pre-recorded video clips by 59 doctors in their daily clinical consultations, and (2) comprehensive metadata of medical dialogues, such as the expertise of doctors, hospital departments and diseases that each doctor is good at treating. To the best of our knowledge, RealMedDial is the first medical dialogue dataset based on real consultations.
- We annotate all the medical dialogues with dialogue-specific diseases and symptoms, which can be used for medical information extraction and intent mining. Combined with doctor profiles, personalized medical dialogue model could be developed to meet diversified medical intents and improve automatic healthcare services.
- We validate the usability of RealMedDial on three tasks, including medical response generation, department routing and doctor recommendation. Experimental results demonstrated the usefulness of our dataset, meanwhile indicating a large space for future research.

The rest of this paper is organized as follows. Section 2 reviews related work to our paper. Section 3 introduces our data collection including data source, cleaning, annotation strategy and statistics. Section 4 provides our experiments of dialogue generation, department routing and doctor recommendation on the constructed dataset. Section 5 concludes this work and provides future directions for the constructed dataset.

2 Related Work

Our work primarily concerns two lines of related work: medical dialogue systems and medical dialogue datasets.

2.1 Medical Dialogue Systems

Recent research on medical dialogue systems has mostly focused on natural language understanding and dialogue management. Various natural language understanding tasks have been investigated in medical dialogue, such as medical information extraction (Lin et al., 2019; Du et al., 2019a,b; Zhang et al., 2020), medical slot filling (Shi et al., 2020), and medical conversational summarization (Joshi et al., 2020). For dialogue management, inspired by the successful application of reinforcement learning in dialogue management strategy (Dhingra et al., 2017; Li et al., 2017; Peng et al., 2018), Wei et al. (2018) first addressed automatic diagnosis in medical dialogue using reinforcement learning framework. Xu et al. (2019) further proposed an end-to-end relational dialogue system to enhance medical diagnosis using knowledge-routed deep Q-network. Xia et al. (2020) proposed a GAN-based policy gradient framework for automatic diagnosis. However, most previous work merely fo-

cused on a single module of the pipeline-based medical dialogue system, and built task-specific datasets for model evaluation. Our work aims to build a real-time based medical dialogue dataset that contains as much information as possible to facilitate various tasks of medical dialogue.

2.2 Medical Dialogue Datasets

For medical dialogue datasets, the MZ dataset (Wei et al., 2018) and the DX dataset (Xu et al., 2019) were first launched for symptom extraction using self-reports of patients and conversations in online healthcare community. Similarly, Shi et al. (2020) collected a dialogue dataset with the purpose of medical slot filling. Since these datasets are collected for symptom extraction tasks, they are hardly applied to other medical dialogue tasks. Lin et al. (2019) released the CMDD dataset with 2,067 pediatric-related dialogues. Zhang et al. (2020) collected the MIE dataset with 1,120 cardiovascular-related dialogues. Zhou et al. (2021) collected two dialogue datasets, CovidDialog in English and in Chinese, containing doctor-patient conversations about COVID-19. CMDD, MIE and CovidDialog are built for understanding disease-specific natural language, but not for dialogue generation. Zeng et al. (2020) built two large-scale medical dialogue datasets, MedDialog-CN and MedDialog-EN from different healthcare communities. Liu et al. (2020) released a large-scale high-quality medical dialogue dataset related to 12 types of common gastrointestinal diseases. Existing medical dialogue datasets are mostly built from the offline question answering in online healthcare communities to simulate real-time dialogue, which partly hinders the performance of medical dialogue systems. Unlike previous work, we build our dataset based on real medical consultations to enhance medical dialogue performance. We compare our dataset and other existing medical dialogue datasets in Table 1.

3 Data Collection

3.1 Dataset Overview

Our raw data are crawled from the short-video clips of Kuaishou, which is one of the largest Chinese short-video clip platform with over three hundred professional doctors out of about 300 million users. Doctors regularly release their daily medical consultation video clips for healthcare services. The video clips of medical consultation record the whole process of medical diagnosis with conversations between doctors and patients.

# dialogues	2,637
# utterances	24,255
Avg. # of utterances in a dialogue	9.20
Median # of utterances in a dialogue	8
Max # of utterances in a dialogue	48
Min # of utterances in a dialogue	2
# doctors	59
Max # of dialogues of a doctor	184
Min # of dialogues of a doctor	4
# departments	17
Avg. # of dialogues of a department	155.12
Avg. # of doctors of a department	3.47
Avg. # of diseases of a department	4.26
Avg. # of symptoms of a department	8.48

Table 2: The statistics of the RealMedDial dataset.

We manually searched and selected 125 professional doctor accounts with totally 4.7K video clips as our primary data source. To avoid potential ethical risks and ensure the quality of the data, we manually filtered the video clips that have been edited or only included introduction to popular science medical knowledge, retaining those containing complete multi-turn patient-doctor dialogues. Finally, we obtained 2,637 video clips released by 59 doctor accounts with multi-turn doctor-patient dialogue. Table 2 shows the statistics of our dataset containing medical dialogues transcribed from real-scenario medical consultations.

Besides, since the collected dialogues are from real doctors, we crawled the profiles of doctors from Kuaishou user homepages and Baidu Encyclopedia[2] as supplement metadata for fine-grained medical dialogue research. We categorized the doctors according to hospital departments, which could be used to build fine-grained dialogue model for different hospital departments. We show the statistics of hospital departments of our datasets in Table 3. In the following sections, we describe our data cleaning , annotation strategy and quality control in detail.

3.2 Data Cleaning

We transcribed the contents of the selected video clips as text. Fifteen graduate students participated in the transcription process with five-fold cross validation to ensure the quality of the transcribed text.

[2]https://baike.baidu.com/

ID	Department	# doctors	# dialogues
1	Cardiovascular	7	240
2	Andrology	3	35
3	Dermatology	6	272
4	Internal Medicine	6	326
5	Gastroenterology	9	150
6	Orthopedics	4	182
7	Anorectal	3	116
8	Obstetrics	3	130
9	Gynecology	2	48
10	Rheumatology	6	424
11	Chinese Medicine	4	202
12	Urology	1	26
13	Endocrinology	2	86
14	Nephrology	3	260
15	Brain	1	26
16	Respiratory	1	4
17	Pediatrics	2	110

Table 3: The hospital departments of the RealMedDial dataset.

Figure 2: The interface for data annotation.

Each transcribed medical dialogue contains four fields: video title, multi-turn dialogue, diseases and symptoms. The video titles often appear in the form of question sentences, indicating the medical problems that the video contents aim to solve. The dialogue is the entire process of real-time medical consultations with multi-turn patient-doctor question answering. The diseases and the symptoms are annotated based on Chinese medical subject headings (Li et al., 2001). We removed personal information, duplicate video clips and single turn dialogues by rule-based filtering.

3.3 Annotation Strategy and Quality Control

We annotated the transcribed medical dialogue with user intents, including diseases and symptoms. Other medical intents can be extended in future studies. The annotation process is achieved by fifteen native Chinese graduate students under the guidance of a professional expert. The annotators followed detailed annotation instructions with standard principles and potentially occurred difficulties. In the annotation process, formal training lessons and regular seminars are carried out to exchange ideas and discuss problems on annotation once a week during the six-week annotation process. The annotation guidelines changed three times as we added information on newly found annotation difficulties during the entire annotation period.

Specifically, we divided fifteen students into five groups, and each group consisted of three student annotators. Using cross-validated annotation, the three-member groups annotated the user intents, and the expert participated in the final decision when there was divergence. If an agreement could not be reached on certain data annotation, everyone discussed and determined the annotation to ensure its accuracy and consistency.

We used a standardized method to achieve high-quality annotation. An interface, shown in Figure 2, was provided to allow the annotators to precisely enter intent information. To promote the correction of the entered terms, we employed Chinese medical subject headings as a support tool to obtain more specialized expressions of user intents. Since the annotation process is based on the annotators' intuition, the results may be subjective. To verify the reliability of annotations, we adopt Kappa score (Sidney and John, 1988) to measure inter-annotator agreement, which is widely used in annotation scheme of computational linguistics. To measure inter-annotator agreement, the annotators were given the same 1,000 medical dialogues to annotate the intents. The agreements on the intents of diseases and symptoms were 0.78 and 0.74, which means the annotation is substantially reliable.

4 Experiments

The RealMedDial dataset is built from real-scenario medical consultations, and thus, it can be used to simulate medical dialogue in a real environment for developing effective automatic medical chatbot. Except for generating useful responses of medical dialogue, RealMedDial can also be used for the department routing task and the doctor recommendation task. Next, we provide detailed evaluation on these three tasks, respectively.

4.1 Medical Response Generation

Medical response generation is one of the most important tasks for medical dialogue, aiming to

generate informative and instructive responses in consideration of the dialogue context and health conditions of patients. Since we collect the doctor-patient conversations from real medical consultations, the data can largely cover language usage patterns of real human-to-human oral conversations. Therefore, our dataset is more conducive to capturing and modeling semantic information in the dialogue by the machine, thereby simulating artificial language patterns to generate useful responses and fully grasp the contextual information of the current dialogue.

4.1.1 Model Pretraining

We trained several response generation models on the RealMedDial dataset as benchmark results for future comparison. Response generation can be generally formulated as a language modeling process in recent proposed models. Given the dialogue context with multi-turn conversations, the probability on the sequence of tokens in the response is modeled as follows:

$$p(r|c) = p(r_1|c) \prod_{i=2}^{n} p(r_i|c, r_1, ..., r_{i-1}), \quad (1)$$

where c denotes the multi-turn dialogue context, and r denotes the next token in the generated response.

Based on this idea, the pretrained GPT2 model (Radford et al., 2019) is proposed to use Transformer decoder to model the generative conditional probability, which enhances the GPT model (Radford et al., 2018) with a few modifications. GPT2 achieves good performance on several text generation tasks reported from existing work (Mass and Roitman, 2020; Bai et al., 2021).

BERT-GPT (Wu et al., 2020; Lewis et al., 2020) is another pretrained language model that integrates the BERT-based encoder and the GPT-based decoder. In BERT-GPT, BERT is used to encode the input token sequence with masks, which is then fed into the GPT decoder for recovering the masked tokens and generating the dialogue responses.

CDial-GPT is a recently proposed pretrained model for Chinese dialogue generation, which is built on a large-scale cleaned Chinese conversation dataset LCCC (Wang et al., 2020). CDial-GPT fills up the gaps in the pre-trained Chinese GPT language models, and provides a reliable model for Chinese dialogue generation. We use the pretrained GPT2, BERT-GPT and CDial-GPT, and fine-tune these models on the RealMedDial dataset to examine their performance for dialogue generation.

We split the RealMedDial dataset into a training set, a validation set, and a test set with the ratio of 8:1:1. The split was carried out separately in different departments, which was based on dialogues instead of source-target pairs. For CDial-GPT and GPT2, we used the implementation by THU-COAI[3], and followed the default hyperparameter settings in the original CDial-GPT (Wang et al., 2020). For BERT-GPT, we used the implementation by UCSD-AI4H[4], and also followed the default hyperparameter setting of the original model. The maximum length of input sequences was truncated to 300, and that of output sequences was truncated to 100. Top-k random sampling (Fan et al., 2018) with k=50 was used for decoding in all the used models.

We evaluated the trained models using automatic metrics including Perplexity, NIST-n (Doddington, 2002) (where n is the size of n-gram and is set as 4), BLEU-n (Papineni et al., 2002) (where n is set as 2 and 4), METEOR (Lavie and Agarwal, 2007), Entropy-n (Zhang et al., 2018) (where n is set as 4), and Dist-n (Li et al., 2016) (where n is set as 1 and 2). Perplexity measures the language quality of the generated responses. NIST, BLEU, and METEOR measure the similarity between the generated responses and the ground truths via n-gram matching. Entropy and Dist measure the lexical diversity of the generated responses. CDial-GPT was pretrained on LCCC-base (a large-scale cleaned Chinese conversation dataset), which is filtered from 79 million conversations from one of the largest Chinese social media website Weibo. BERT-GPT was pretrained by UCSD-AI4H on Chinese corpus collected from a large scale Chinese corpus for NLP[5]. GPT2 was pretrained by UCSD-AI4H on Chinese Chatbot Corpus[6] containing 14 million dialogues and 500K Chinese dialogues[7].

4.1.2 Evaluation Results

Table 4 shows the response generation performance on the RealMedDial dataset. From the table, we

[3] https://github.com/thu-coai/CDial-GPT

[4] https://github.com/UCSD-AI4H/Medical-Dialogue-System

[5] https://github.com/brightmart/nlp_chinese_corpus

[6] https://github.com/codemayq/chinese_chatbot_corpus

[7] https://drive.google.com/file/d/1nEuew_KNpTMbyy7BO4c8bXMXN351RCPp/view

	CDial-GPT	GPT2	BERT-GPT
Perplexity	21.25	**16.40**	29.95
NIST-4	**10.30**	9.30	0.55
BLEU-2	**1.196**	1.123	0.068
BLEU-4	**0.481**	0.439	0.028
METEOR	**1.403**	1.385	0.009
Entropy-4	7.00	6.21	**8.99**
Dist-1	0.178	**0.215**	0.090
Dist-2	0.602	**0.647**	0.469

Table 4: Performance of response generation on the RealMedDial dataset.

Conversation history: 我的痘为什么严重了呢？就是在暑假吃凉的，也吃的比较油腻，在家里。天天吃肉，我觉得一天三顿都是肉。(Why is my acne getting worse? I eat cold and greasy food during the summer, and I eat meat at home every day, three meals a day.)
Groundtruth response: 吃肉没关系，就是少油。瘦的肉少油。然后呢，注意夏天不吃茄子，茄子是很油的，茄子特别吸油，尤其是烧茄子，什么茄丁面呀，烧茄子，这些都特别油。(It's okay to eat meat, but less oil is needed. The lean meat has less oil. Then, be careful, do not eat eggplant in the summer, eggplant is very oily, eggplant especially absorb oil, especially roasted eggplant, eggplant diced noodles, etc., these are particularly oily.)
CDial-GPT: 你这种情况你不能吃肉。(In your case you can't eat meat.)
GPT2: 你就是有点儿严重了，所以你要注意饮食，保持好的睡眠，少吃肉，少吃甜的，油炸食品。(You're just a little bit more serious, so you need to pay attention to your diet, get a good night's sleep, eat less meat, less sweet, fried foods.)
BERT-GPT: 应该是内火大，主要是脏腑功能失调所致，所以在饮食方面呢，一定要少吃油腻的和油炸的，这样呢对你皮肤是有影响的，所以平时就早睡。我给你开点中药，你试试，先吃一段时间。(It should be a big internal fire, mainly due to the dysfunction of the internal organs, so in terms of diet, it must eat less greasy and fried, so that it has an impact on your skin, so usually go to bed early. I'll give you some Chinese medicine, you try, first eat a period of time.)

Figure 3: An example of generated responses on the RealMedDial test set.

observe that GPT2 achieved the better performance than the other two models in terms of Perplexity, Dist-1 and Dist-2. Since these three metrics are used to measure the informativeness and diversity of the generated text, it indicates that GPT2 can generate more diverse and informative responses. Although CDial-GPT yielded better performance on four machine translation metrics, NIST-4, BLEU-2, BLEU-4 and METEOR, these metrics are all auxiliary metrics that evaluate the performance of dialogue generation in terms of n-gram matching (Liu et al., 2016). The CDial-GPT model is pretrained using the social media data which may not be well applied in medical dialogue task, thus leading to much lower values of these metrics compared with other tasks, such as machine translation. To further illustrate the comparison of different models, we provide two examples of the generated responses in Figure 3 and Figure 4.

4.2 Department Routing

The task of department routing is to attribute current dialogue with patient descriptions to the corresponding hospital departments for optimizing di-

Conversation history: 就是这个手指头，就是肿着疼着呢，就想这个握拳都不能握，不知道是什么原因，但是我的手指好像变形都好几年了，今年好像 10 号有点疼。(My fingers are swollen and painful, and I can't even make a fist if I want to, I don't know what the reason is, but my fingers seem to be deformed for several years, and this year it seems to be a little painful on the 10th.)
Groundtruth response: 你听我说啊，你的这个远端这个手指头这都是长的骨刺，它不是类风湿，这个呢就是骨关节炎，这骨关节炎呢一个是跟遗传有关系，另外一个就是年轻的时候可能摸凉水，干活儿干的太过了，它现在呢就会退化的，长骨刺更加早一些，平时呢尽量的这种少摸一些凉水，少用手干一干活，省着点儿用。另外呢我开点儿中药连喝再洗，好不好？(You have a bone spur in your finger, it is not rheumatoid, but osteoarthritis. This osteoarthritis is related to genetics, and the other is that when you are young, you may touch the cold water, work too much, and it will degenerate now, the bone spur is even earlier, usually try to touch less cold water, less hand work, save it. In addition, I will give you some Chinese medicine, you drink and wash, okay?)
CDial-GPT: 你这个可能是关节有问题，需要去医院检查一下。(You may have a problem with your joints and need to go to the hospital for a checkup.)
GPT2: 建议你去拍个片子，平时注意保护手指。(It is recommended that you get an x-ray and pay attention to protecting your fingers in general.)
BERT-GPT: 变形了好几年，首先你要去医院检查一下，拍 x 片，看看有没有坏死，然后呢你伸指的时候疼不疼？有没有麻木？如果一点儿感觉都没有，我们就给你开点中药试试。(Deformation for several years, first of all you have to go to the hospital to check, take x-rays, see if there is necrosis, and then you stretch your fingers when the pain is not painful? Is there any numbness? If you don't feel anything at all, we'll give you some Chinese medicine to try.)

Figure 4: Another example of generated responses on the RealMedDial test set.

alogue models according to the characteristics of the departments. Since the RealMedDial dataset contains a wide range of hospital departments with respect to different medical domains, we design this task for providing more targeted dialogue services for patients. The department routing task can be tackled by multi-class classification. Namely, given a brief description of the health issues of an patient, related departments can be matched for the patient, which could be helpful to get more accurate medical services and fine-tune the pretained dialogue model for generating more personalized and useful responses.

4.2.1 Model Pretraining

We use three BERT-based models, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2020), for department routing. These models have been proved to be effective in various NLP tasks. Besides, we also adopt the CPT (Shao et al., 2021) model, which is designed as a Chinese pre-trained unbalanced Transformer to utilize the shared knowledge between natural language understanding and natural language generation through a shared encoder, an understanding decoder, and a generation decoder. We input health descriptions into these models and predict the department that the corresponding disease or symptom belongs to.

4.2.2 Implementation Details

We labeled the dialogue texts with index of departments and doctors respectively, and split the dataset into a training set and a test set with the

Metrics	CPT	RoBerta	BERT	ALBERT
Accuracy	**0.749**	0.705	0.686	0.611
m-Prec.	**0.552**	0.485	0.452	0.377
m-Recall	**0.540**	0.487	0.481	0.371
m-F1	**0.536**	0.480	0.462	0.359
w-Prec.	**0.707**	0.654	0.627	0.553
w-Recall	**0.749**	0.705	0.686	0.611
w-F1	**0.723**	0.674	0.652	0.566

Table 5: Performance of department routing on the RealMedDial dataset.

Metrics	CPT	RoBerta	BERT	ALBERT
Accuracy	**0.621**	0.523	0.552	0.348
m-Prec.	**0.379**	0.247	0.263	0.115
m-Recall	**0.375**	0.277	0.293	0.144
m-F1	**0.353**	0.240	0.256	0.110
w-Prec.	**0.564**	0.409	0.439	0.228
w-Recall	**0.621**	0.523	0.552	0.348
w-F1	**0.565**	0.438	0.465	0.247

Table 6: Performance of doctor Recommendation on the RealMedDial dataset.

ratio of 4:1. Pretrained language models are the primary ingredients of the state-of-the-art text classifiers including BERT, RoBERTa, ALBERT and CPT. These models are trained on the training set, and the weighting parameters were learned with AdamW (Loshchilov and Hutter, 2017), whose ϵ was set as 1e-8. The initial learning rate was set as 4e-5. The learning rate scheduler was set as Linear. We evaluate the models with metrics including Accuracy, macro/weighted Precision (m/w-Prec.), macro/weighted Recall (m/w-Recall) and macro/weighted F1 score (m/w-F1).

4.2.3 Evaluation Results

Table 5 shows the experimental results for department routing. From the table, we can observe that CPT outperforms other models in all the evaluation metrics. This is because CPT can capture specific knowledge of this task using a shared encoder, an understanding decoder, and a generation decoder. Compared with other models, CPT takes full advantage of previous pre-trained models and achieves better performance in department routing.

4.3 Doctor Recommendation

Doctor recommendation aims to recommend suitable doctors based on patients' health status. Since RealMedDial contains real medical consultation records of multiple doctors, we can model different doctors according to the characteristics of their language usage and unique forms of question-answering by building doctor profiles. The doctor profiles can be used to build chatbots to assist in the completion of various medical health services.

4.3.1 Evaluation Results

Similar to department routing, doctor recommendation task is also a multi-class classification problem, and we still use BERT, RoBERTa, ALBERT and CPT as baselines. Table 6 shows the experimental results for doctor recommendation. From the table, we can observe that CPT still outperforms other models in terms of different evaluation metrics. Like the performance trends of department routing, CPT contributes to effective modeling of language usage pattern and profiles of doctors by incorporating more comprehensive domain-specific information into the learned model, and yields better doctor recommendation results.

4.4 Further Discussion

We validate the usability of RealMedDial on medical response generation, department routing and doctor recommendation. Experimental results have shown the usefulness of RealMedDial in these tasks, which also provides benchmark results for future studies. Advanced models trained using RealMedDial could consider more special nature of medical consultation for generating accurate and low risk medical responses. To this end, more effective models trained on RealMedDial can be devised by comprehensively using the wide range of metadata of our corpus, such as doctor profiles and disease descriptions. Although experiments in this work are our preliminary attempts on demonstrating the usefulness of our corpus, we will extend our future work to consider more domain-specific information to develop more effective generation models. Although our dataset is in Chinese, the application scenario is not limited to Chinese applications. To adapt RealMedDial to research in other languages, dialogue contents can be tokenized to token IDs, which can thus be used to train dialogue models for other language-based research. We will also use automated methods, such as the automatic transcription software [8], to reduce the time cost and manual labor in constructing and expanding our dataset in future.

5 Conclusion and Future Work

To facilitate automatical medical consultation, we construct RealMedDial, a high-quality Chinese

[8] https://sonix.ai/

dataset of medical dialogue based on real scenario medical consultation from online short-video clips. Real medical consultation contributes to learning more powerful and human-like dialogue models by considering communications in reality between doctors and patients instead of question answering-based communications on online health community. We collected and annotated a wide range of meta-data with respect to the medical dialogue including titles of short-video clips, doctor profiles, hospital departments, diseases and symptoms for fine-grained analysis on language usage pattern and clinical diagnosis. We evaluated the performance of medical response generation, department routing and doctor recommendation on RealMedDial. Results show that RealMedDial is applicable to various medical dialogue tasks. As for future work, we will build personalized dialogue models by incorporating more professional knowledge into medical response generation.

6 Ethical Consideration

The original short-video clips of our study are collected from Kuaishou, one of the largest Chinese short-video clip platform with over three hundred professional doctors out of about 300 million users. The doctor profiles are collected from Kuaishou user homepages and Baidu Encyclopedia. All the collected data are public available, which do not contain any personal privacy information of patients and doctors. We have ensured that the doctors obtain patient consent in the first place to post videos of consultations, and the short videos are without any personal information of patients. The constructed corpus is completely anonymous, and the identity of the patient or doctor cannot be inferred from it. The transcribed texts in RealMedDial are randomly shuffled so that it is hard to find connections between the original videos and the transcribed text without the identity of the doctor. Therefore, there is no privacy issue for the data we use. When annotating the dataset, all annotators were paid based on their workload and submitted all required consent forms. Since this work only focuses on medical dialogue without additional identified and private information, the protection of privacy is preserved.

7 Acknowledgements

This work is partially supported by grant from the Natural Science Foundation of China (No. 62006034), Natural Science Foundation of Liaoning Province (No. 2021-BS-067) and the Fundamental Research Funds for the Central Universities (No.DUT21RC(3)015). We would like to thank our reviewers for their insightful comments, which help us greatly enhance our work.

References

He Bai, Peng Shi, Jimmy Lin, Luchen Tan, Kun Xiong, Wen Gao, Jie Liu, and Ming Li. 2021. Semantics of the unwritten: The effect of end of paragraph and sequence tokens on text generation with GPT2. In *Proceedings of the ACL-IJCNLP 2021 Student Research Workshop, ACL 2021, Online, July 5-10, 2021*, pages 148–162. Association for Computational Linguistics.

Avisha Das, Salih Selek, Alia R. Warner, Xu Zuo, Yan Hu, Vipina Kuttichi Keloth, Jianfu Li, W. Jim Zheng, and Hua Xu. 2022. Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues. In *Proceedings of the 21st Workshop on Biomedical Language Processing, BioNLP@ACL 2022*, pages 285–297. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 484–495. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurence statistics. In *Proceedings of the second international conference on Human Language Technology Research, 2002*.

Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019a. Extracting symptoms and their status from clinical conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 915–925. Association for Computational Linguistics.

Nan Du, Mingqiu Wang, Linh Tran, Gang Lee, and Izhak Shafran. 2019b. Learning to infer entities, properties and their relations from clinical conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4978–4989. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*.

Soheil Keshmiri, Hidenobu Sumioka, Ryuji Yamazaki, and Hiroshi Ishiguro. 2019. Decoding the perceived difficulty of communicated contents by older people: Toward conversational robot-assistive elderly care. *IEEE Robotics Autom. Lett.*, 4(4):3263–3269.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, pages 228–231. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Danya Li, Tiejun Hu, Zhu Wenyan, Qing Qian, Huiling Ren, Junlian Li, and Bin Yang. 2001. Retrieval system of chinese medical subject headings. *Chinese Journal of Medical Library*, 4:1–9.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 733–743. Asian Federation of Natural Language Processing.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5032–5041. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2122–2132. The Association for Computational Linguistics.

Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *CoRR*, abs/2010.07497.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Devin M Mann, Ji Chen, Rumi Chunara, Paul A Testa, and Oded Nov. 2020. Covid-19 transforms health care through telemedicine: evidence from the field. *Journal of the American Medical Informatics Association*, 27(7):1132–1135.

Yosi Mass and Haggai Roitman. 2020. Ad-hoc document retrieval using weak-supervision with BERT and GPT2. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4191–4197. Association for Computational Linguistics.

Shamila Nasreen, Julian Hough, and Matthew Purver. 2021. Rare-class dialogue act tagging for alzheimer's disease diagnosis. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2021*, pages 290–300. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2182–2192. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. CPT: A pre-trained unbalanced transformer for both chinese language understanding and generation. *CoRR*, abs/2109.05729.

Xiaoming Shi, Haifeng Hu, Wanxiang Che, Zhongqian Sun, Ting Liu, and Junzhou Huang. 2020. Understanding medical conversations with scattered keyword attention and weak supervision from responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8838–8845.

Siegel Sidney and Castellan N. John. 1988. Nonparametric statistics for the behavioral sciences. *McGraw-Hill*.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *The CCF International Conference on Natural Language Processing and Chinese Computing*, pages 91–103.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang 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.

Qingyang Wu, Lei Li, Hao Zhou, Ying Zeng, and Zhou Yu. 2020. Importance-aware learning for neural headline editing. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9282–9289. AAAI Press.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 1062–1069. AAAI Press.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7346–7353.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9241–9250. Association for Computational Linguistics.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1815–1825.

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. Mie: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469.

Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, Yichen Zhang, Qingyang Wu, Zhou Yu, Kun Xu, Eric P. Xing, and Pengtao Xie. 2021. On the generation of medical dialogs for COVID-19. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 886–896. Association for Computational Linguistics.