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

Bo Xu\textsuperscript{1}, Hongtong Zhang\textsuperscript{1}, Jian Wang\textsuperscript{1}, Xiaokun Zhang\textsuperscript{1}, Dezhi Hao\textsuperscript{1}, Linlin Zong\textsuperscript{2}\textsuperscript{*}, Hongfei Lin\textsuperscript{1}, Fenglong Ma\textsuperscript{3}

\textsuperscript{1}School of Computer Science and Technology, Dalian University of Technology, China
\textsuperscript{2}School of Software, Dalian University of Technology, China
\textsuperscript{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,
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 Kuaishou¹, 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-

¹https://www.kuaishou.com
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.

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 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.

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 scientific 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 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.

Table 2: The statistics of the RealMedDial dataset.
Table 3: The hospital departments of the RealMedDial dataset.

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

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_KNpTMbyy7BO4c8bXMNX351RCPp/view
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 dialogue 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 a patient, related departments can be matched for the patient, which could be helpful to get more accurate medical services and fine-tune the pretrained 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 proportion of training set to test set is 8:2. We set the number of epochs to 10, and the batch size to 32.

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

<table>
<thead>
<tr>
<th></th>
<th>CDial-GPT</th>
<th>GPT2</th>
<th>BERT-GPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perplexity</td>
<td>21.25</td>
<td>16.40</td>
<td>29.95</td>
</tr>
<tr>
<td>NIST-4</td>
<td>10.30</td>
<td>9.30</td>
<td>0.55</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>1.196</td>
<td>1.123</td>
<td>0.068</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0.481</td>
<td>0.439</td>
<td>0.028</td>
</tr>
<tr>
<td>METEOR</td>
<td>1.403</td>
<td>1.385</td>
<td>0.009</td>
</tr>
<tr>
<td>Entropy-4</td>
<td>7.00</td>
<td>6.21</td>
<td>8.99</td>
</tr>
<tr>
<td>Dist-1</td>
<td>0.178</td>
<td>0.215</td>
<td>0.090</td>
</tr>
<tr>
<td>Dist-2</td>
<td>0.602</td>
<td>0.647</td>
<td>0.469</td>
</tr>
</tbody>
</table>

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

Figure 4: Another example of generated responses on the RealMedDial test set.
Table 5: Performance of department routing on the RealMedDial dataset.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>CPT</th>
<th>RoBerta</th>
<th>BERT</th>
<th>ALBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.749</td>
<td>0.705</td>
<td>0.686</td>
<td>0.611</td>
</tr>
<tr>
<td>m-Prec.</td>
<td>0.552</td>
<td>0.485</td>
<td>0.452</td>
<td>0.377</td>
</tr>
<tr>
<td>m-Recall</td>
<td>0.540</td>
<td>0.487</td>
<td>0.481</td>
<td>0.371</td>
</tr>
<tr>
<td>m-F1</td>
<td>0.536</td>
<td>0.480</td>
<td>0.462</td>
<td>0.359</td>
</tr>
<tr>
<td>w-Prec.</td>
<td>0.707</td>
<td>0.654</td>
<td>0.627</td>
<td>0.553</td>
</tr>
<tr>
<td>w-Recall</td>
<td>0.749</td>
<td>0.705</td>
<td>0.686</td>
<td>0.611</td>
</tr>
<tr>
<td>w-F1</td>
<td>0.723</td>
<td>0.674</td>
<td>0.652</td>
<td>0.566</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>CPT</th>
<th>RoBerta</th>
<th>BERT</th>
<th>ALBERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.621</td>
<td>0.523</td>
<td>0.552</td>
<td>0.348</td>
</tr>
<tr>
<td>m-Prec.</td>
<td>0.379</td>
<td>0.247</td>
<td>0.263</td>
<td>0.115</td>
</tr>
<tr>
<td>m-Recall</td>
<td>0.375</td>
<td>0.277</td>
<td>0.293</td>
<td>0.144</td>
</tr>
<tr>
<td>m-F1</td>
<td>0.353</td>
<td>0.240</td>
<td>0.256</td>
<td>0.110</td>
</tr>
<tr>
<td>w-Prec.</td>
<td>0.564</td>
<td>0.409</td>
<td>0.439</td>
<td>0.228</td>
</tr>
<tr>
<td>w-Recall</td>
<td>0.621</td>
<td>0.523</td>
<td>0.552</td>
<td>0.348</td>
</tr>
<tr>
<td>w-F1</td>
<td>0.565</td>
<td>0.438</td>
<td>0.465</td>
<td>0.247</td>
</tr>
</tbody>
</table>

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 $\epsilon$ 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, 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
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


Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, Soheil Keshmiri, Hidenobu Sumioka, Ryuji Yamazaki, Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018.


