MedDialog: Large-scale Medical Dialogue Datasets

**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**
UC San Diego
p1xie@eng.ucsd.edu

Abstract

Medical dialogue systems are promising in assisting in telemedicine to increase access to healthcare services, improve the quality of patient care, and reduce medical costs. To facilitate the research and development of medical dialogue systems, we build large-scale medical dialogue datasets – MedDialog, which contain 1) a Chinese dataset with 3.4 million conversations between patients and doctors, 11.3 million utterances, 660.2 million tokens, covering 172 specialties of diseases, and 2) an English dataset with 0.26 million conversations, 0.51 million utterances, 44.53 million tokens, covering 96 specialties of diseases. To our best knowledge, MedDialog is the largest medical dialogue dataset to date. We pretrain several dialogue generation models on the Chinese MedDialog dataset, including Transformer, GPT, BERT-GPT, and compare their performance. It is shown that models trained on MedDialog are able to generate clinically correct and human-like medical dialogues. We also study the transferability of models trained on MedDialog to low-resource medical dialogue generation tasks. It is shown that via transfer learning which finetunes the models pretrained on MedDialog, the performance on medical dialogue generation tasks with small datasets can be greatly improved, as shown in human evaluation and automatic evaluation. The datasets and code are available at https://github.com/UCSD-AI4H/Medical-Dialogue-System

1 Introduction

Telemedicine refers to the practice of delivering patient care remotely, where doctors provide medical consultations to patients using HIPAA compliant video-conferencing tools. As an important complement to traditional face-to-face medicine practiced physically in hospitals and clinics, telemedicine has a number of advantages. First, it increases access to care. For people living in medically under-served communities (e.g., rural areas) that are in shortage of clinicians, telemedicine enables them to receive faster and cheaper care compared with traveling over a long distance to visit a clinician. Second, it reduces healthcare costs. In a study[1] by Jefferson Health, it is shown that diverting patients from emergency departments with telemedicine can save more than $1,500 per visit. Third, telemedicine can improve the quality of care. The study in (Pande and Morris, 2015) shows that telemedicine patients score lower for depression, anxiety, and stress, and have 38% fewer hospital admissions. Other advantages include improving patient engagement and satisfaction, improving provider satisfaction, etc. Please refer to (Wootton et al., 2017) for a more comprehensive review.

While telemedicine is promising, it has several limitations. First, it puts additional burden on physicians. In addition to practicing face-to-face medicine which already makes physicians very busy, physicians need to provide remote telemedicine consultations, which further increases the risk of physician burnout. Second, different from in-hospital patients, the progression of whose medical conditions can be easily tracked by clinicians, remote patients are difficult to track and monitor. To address such problems, there has been increasing research interest in developing artificial intelligence (AI) methods to assist in telemedicine. In particular, medical dialogue systems are being developed to serve as "virtual doctors". These "virtual doctors" are aimed to interact with patients via natural dialogues, asking about the medical conditions and history of patients and providing clinical advice. They can also proactively reach out to patients to ask about the progression of patients' conditions and provide timely interventions.

---

[1]https://www.healthleadersmedia.com/clinical-care/cost-savings-telemedicine-estimated-19-120-patient-visit

To build medical dialogue systems, a large collection of conversations between patients and doctors is needed as training data. Due to data privacy concerns, such data is difficult to obtain. The existing medical dialogue datasets (Xu et al., 2019; Yang et al., 2020) are limited in size or biased to certain diseases, which cannot adequately serve the purpose of training medical dialogue systems that can achieve doctor-level intelligence and cover many specialities in medicine.

To address the limitations of existing datasets, we build large-scale medical dialogue datasets – MedDialog – that contain 1) a Chinese dataset with 3.4 million conversations between patients and doctors, 11.3 million utterances, 660.2 million tokens, covering 172 specialties of diseases, and 2) an English dataset with 0.26 million conversations, 0.51 million utterances, 44.53 million tokens, covering 96 specialties of diseases. Both datasets cover almost all specialities in medicine, ranging from internal medicine to family medicine and covers a wide spectrum of diseases, including cancer, pneumonia, etc. To our best knowledge, they are the largest Chinese and English medical dialogue datasets to date. The data is open to the public. Each consultation starts with a description of medical conditions and history, followed by the conversation between doctor and patient. In certain consultations, doctors make diagnosis conclusions and give suggestions on treatment. The conversations have multiple turns.

On the Chinese MedDialog (MedDialog-CN) dataset, we train several dialogue generation models for the interested community to benchmark with. Generating a response given the conversation history can be formulated as a sequence-to-sequence (seq2seq) learning problem, where we use the Transformer (Vaswani et al., 2017) architecture to perform this task. Transformer consists of an encoder which embeds the conversation history and a decoder which generates the response. Both the encoder and decoder use self-attention to capture long-range dependency between tokens. In addition to training the Transformer on MedDialog-CN from scratch, we can pretrain the encoder and decoder on a corpora much larger than MedDialog-CN, then finetune them on MedDialog-CN. BERT-GPT (Wu et al., 2019; Lewis et al., 2019) is a pretrained model where the encoder is pretrained using BERT (Devlin et al., 2018) and the decoder is pretrained using GPT (Radford et al.). Besides

the seq2seq formulation, dialogue generation can be formulated as a language modeling problem which generates the next token in the response conditioned on the concatenation of the already generated tokens in the response and the conversation history. GPT (Radford et al.; Zhang et al., 2019) is a pretrained language model based on Transformer decoder. BERT-GPT and GPT are finetuned on MedDialog-CN. We perform evaluation of these models using automatic metrics including perplexity, BLEU (Papineni et al., 2002a), Dist (Li et al., 2015), etc. The generated responses are clinically informative, accurate, and human-like.

We utilize the models trained on the large-scale MedDialog-CN dataset to improve performance in low-resource dialogue generation tasks where the dataset size is small. The study is performed on COVID-19 dialogue generation on the CovidDialog (Yang et al., 2020) dataset, which contains 1,088 dialogues and 9,494 utterances. The small size of this dataset incurs high risk of overfitting, if directly training the large-sized neural models on it. To alleviate this risk, we take the weights of dialogue generation models pretrained on MedDialog-CN and finetune the weights on CovidDialog. Human evaluation and automatic evaluation show that pretraining on MedDialog-CN can greatly improve the performance on CovidDialog and generate clinically meaningful consultations about COVID-19.

The major contributions of this paper are:

- We build large-scale medical dialog datasets – MedDialog, which contain 1) a Chinese dataset with 3.4 million conversations between patients and doctors, 11.3 million utterances, 660.2 million tokens, covering 172 specialties of diseases, and 2) an English dataset with 0.26 million conversations, 0.51 million utterances, 44.53 million tokens, covering 96 specialties of diseases. To our best knowledge, they are the largest of their kinds to date.

- We pretrain several dialogue generation models on the Chinese MedDialog dataset, including Transformer, BERT-GPT, and GPT, and compare their performance using automatic metrics.

- Through human evaluation and automatic evaluation, we show that the pretrained models on MedDialog-CN can significantly improve performance on medical dialogue generation tasks where the dataset size is small, via transfer learning.

The rest of this paper is organized as follows. Section 2 and 3 present the datasets and dialogue generation models (DGMs). Section 4 gives experimental results of developing DGMs on Chinese MedDialog and studies the transferability of DGMs trained on MedDialog-CN to other low-resource medical dialogue generation tasks. Section 5 reviews related works and Section 6 concludes the paper.

2 Related Works

There have been several works investigating medical dialogue generation. Wei et al. (Wei et al., 2018) built a task-oriented dialogue system for automatic diagnosis. The system detects the user intent and slots with values from utterances, tracks dialogue states, and generates responses. Xu et al. (Xu et al., 2019) developed a medical dialogue system for automatic medical diagnosis that converses with patients to collect additional symptoms beyond their self-reports and automatically makes a diagnosis. This system incorporates a medical knowledge graph into the topic transition in dialogue management. Xia et al. (Xia et al.) developed a reinforcement learning (RL) based dialogue system for automatic diagnosis. They proposed a policy gradient framework based on the generative adversarial network to optimize the RL model.

3 Datasets

Our MedDialog consists of a Chinese dataset and an English dataset, collected from different sources.

3.1 The Chinese MedDialog dataset

The Chinese MedDialog (MedDialog-CN) dataset contains 3.4 million Chinese dialogues (consultations) between patients and doctors. The total number of utterances is 11.3 million. Each consultation starts with the narration of patient' medical condition and history, including present disease, duration of the disease, allergies, medications, past diseases, etc. Then it follows with the multi-turn conversation between patient and doctor. In the conversation, there are cases that multiple consecutive utterances are from the same person (either doctor or patient) and these utterances were posted at different time points. For such cases, we combine the consecutive utterances from the same person into a single utterance. Optionally, at the end of the consultation, the doctor makes diagnosis and

# dialogues	3,407,494
# utterances	11,260,564
# tokens	660,171,367
Avg. # of utterances in a dialogue	3.3
Max # of utterances in a dialogue	198
Min # of utterances in a dialogue	2
Avg. # of tokens in an utterance	55.6
Max # of tokens in an utterance	6,935
Min # tokens in an utterance	1

Table 1: Statistics of the Chinese MedDialog dataset

treatment suggestions to the patient. Table 1 shows statistics of the Chinese dataset. Figure 1 shows an exemplar consultation. The data is crawled from an online consultation website – haodf.com[2], which provides consultation service to patients. The dialogues cover 29 broad categories of specialties including internal medicine, pediatrics, dentistry, etc. and 172 fine-grained specialties including cardiology, neurology, gastroenterology, urology, etc. The consultations are conducted from 2010 to 2020.

3.2 The English MedDialog dataset

The English MedDialog (MedDialog-EN) dataset contains 0.26 million English consultations between patients and doctors. The total number of utterances is 0.51 million. Each consultation consists of two parts: (1) description of patient's medical conditions; (2) conversation between patient and doctor. The data is crawled from iclinic.com[3] and healthcaremagic.com[4], which are two online platforms of healthcare services, including symptom self-checker, video consultation, online chat with doctors, etc. The consultations cover 51 categories of communities including diabetes, elderly problems, pain management, etc. and 96 specialties including andrology, cardiology, nephrology, pharmacology, etc. The consultations were conducted from 2008 to 2020. Table 2 shows statistics of the English dataset.

3.3 Advantages of our datasets

To our best knowledge, MedDialog-CN and MedDialog-EN are the largest Chinese and English medical dialog dataset respectively. They have the following advantages.

- **Large number of conversations and utterances.** MedDialog-CN has about 3.4 mil-

[2]https://www.haodf.com/
[3]https://www.icliniq.com/
[4]https://www.healthcaremagic.com/

Description of medical conditions and history

疾病: 宝宝眼角红红的, 严重时轻微溃烂. (Disease: The baby's eyes are red and slightly ulcerated when becoming severe.)

病情描述: 宝宝眼角红红的氧, 用小手挠, 严重时轻微溃烂, 怎么回事. 用了紫草膏很快消失过两天又出来了. (Medical condition: The baby's eyes are red and itchy, scratched with hand, and slightly ulcerated when becoming severe. After using Burt's bee Res-Q ointment, it disappeared quickly but came out after two days.)

希望获得的帮助: 宝宝眼角红红怎么回事. (Help needed: What's wrong with baby's red eyes?)

患病多久: 一月内. (Hong long the condition has been: Less than one month)

过敏史: 无 (Allergies: No)

既往病史: 无 (Past medical history: No)

Dialogue

医生: 感谢您的信任, 病情资料我已详细阅读. 根据您现有的资料, 建议: 睑缘炎. 图片不是很清楚. 经常揉是吧? (Doctor: Thank you for your trust. I have read the medical information in detail. Based on the existing information, the diagnosis is blepharitis. The picture is not very clear. Scratch it often, right?)

病人: 出生到现在奶量一直很少, 嘴唇老是干干的, 也不像别的宝宝流口水. (Patient: Drinks little amount of milk since birth, and the baby's lips are always dry, and not drooling like other babies.)

医生: 眼部是局部炎症 (Doctor: Eyes have local arthritis.)

病人: 是的 (Patient: Yes)

医生: 给予典必殊眼膏一天两次 (Doctor: Use Tobramycin and Dexamethasone eye ointment twice a day)

病人: 这个怎么回事 (Patient: What's going on?)

医生: 考虑睑腺炎或者睑缘炎 (Doctor: Consider blepharitis or blepharitis)

……

Diagnosis and suggestions

病情摘要及初步印象: 睑腺炎 (Summary of the condition and initial impressions: Blepharitis)

总结建议: 局部炎症, 给予典必殊眼膏一天两次, 观察恢复情况, 必要时医院就诊. (Summary of recommendations: For local inflammation, use Tobramycin and Dexamethasone eye ointment eye ointment twice a day, monitor the recovery, and go to the hospital if necessary.)
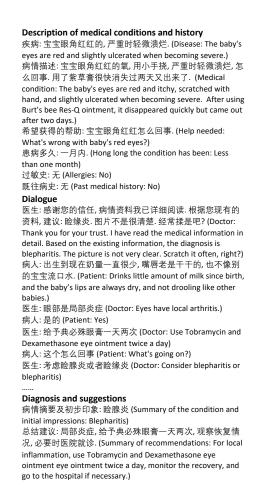
Figure 1: An exemplar consultation, which includes (1) description of medical conditions and history of patient, (2) dialogue between doctor and patient, and (3) diagnosis and treatment suggestions given by doctor.

lion conversations and 11.3 million utterances. MedDialog-EN has about 0.3 million conversations and 0.5 million utterances.

- **Broad coverage of medical specialities.** Consultations in MedDialog-CN are about 29 broad categories of specialties and 172 fine-grained specialties. Consultations in MedDialog-EN are about 96 categories of specialties.

- **Diversity of the patients.** The patients in MedDialog-EN are from all over the world, with different nationalities, ethics, age, gender, occupation, education, income, etc. The patients in MedDialog-CN are from 31 provincial-level administrative divisions in China, with different ethics, age, gender, occupation, education, income, etc. Such diversity

# dialogues	257,332
# utterances	514,664
# tokens	44,527,872
Avg. # of utterances in a dialogue	2
Max # of utterances in a dialogue	2
Min # of utterances in a dialogue	2
Avg. # of tokens in an utterance	86.5
Max # of tokens in an utterance	3,672
Min # tokens in an utterance	1

Table 2: Statistics of the English dataset

Dataset	#dialogs	#diseases
Muzhi (Wei et al., 2018)	710	4
Dxy (Xu et al., 2019)	527	5
COVID-EN (Yang et al., 2020)	603	1
COVID-CN (Yang et al., 2020)	1,088	1
MedDialog-CN	3,407,494	172
MedDialog-EN	257,332	96

Table 3: Comparison with other datasets.

greatly minimizes population biases in these two datasets.

Table 3 shows a comparison of our datasets with several other medical dialogue datasets. The number of dialogs and diseases in our datasets are both much larger than those in other datasets.

4 Methods

We train several dialogue generation models on the Chinese MedDialog dataset for the interested research community to benchmark with. During training, given a dialogue containing a sequence of alternating utterances between patient and doctor, we process it into a set of pairs $\{(s_i, t_i)\}$ where the target t_i is a response from the doctor and the source s_i is the concatenation of all utterances (from both patient and doctor) before t_i. A dialogue generation model takes s as input and generates t. This problem can be formulated either as a sequence-to-sequence learning problem where the goal is to generate t conditioned on s via an encoder-decoder model, or as a language modeling problem which generates the i-th token t_i in t conditioned on the concatenation of the conversation history s and the already generated sequence $t_1, \cdots, t_{i-1}$ in the response before t_i via a language model.

4.1 Dialogue Generation as Sequence-to-Sequence Modeling

The problem of response generation can be formulated as a sequence-to-sequence (seq2seq) learn-

ing (Sutskever et al., 2014) problem: given the conversation history s, generate the response t. We use the Transformer (Vaswani et al., 2017) architecture for seq2seq modeling. Transformer consists of an encoder which embeds the input sequence into a latent space and a decoder which takes the embedding of the input sequence as input and generates the output sequence. Different from LSTM-based seq2seq models (Sutskever et al., 2014) which learn representations of a sequence of tokens in a recurrent manner and therefore suffer computational inefficiency due to their sequential nature, Transformer uses self-attention to capture the long-range dependency among tokens by calculating the similarity between each pair of tokens in the sequence. Self-attention avoids sequential computation and greatly facilitates parallel computation. A building block in Transformer contains the following modules: a self-attention sub-layer, a token-wise feed-forward sub-layer, residual connections (He et al., 2016) between sub-layers, and layer normalization (Ba et al., 2016). Both the encoder and decoder are composed of a stack of such building blocks. The encoder generates an encoding for each token in the input sequence. These encodings are fed into the decoder to generate the output sequence. To generate the token at position i, the decoder encodes the generated tokens from 1 to $i - 1$ (like an encoder), calculates an attentional representation by performing attention between the encodings of input tokens and the encodings of output tokens $1, \cdots, i - 1$, then feeds the attentional representation into a softmax layer to generate token i. Transformer learns the weights in the encoder and decoder by maximizing the conditional likelihood of responses conditioned on conversation histories.

4.2 Dialogue Generation as Language Modeling

Besides the sequence-to-sequence formulation, response generation can be formulated as a language modeling problem as well. Given the conversation history s, a language model defines the following probability on the sequence of tokens $t = t_1, \cdots, t_n$ in the response:

$$p(t|s) = p(t_1|s) \prod_{i=2}^{n} p(t_i|s, t_1, \cdots, t_{i-1}), \quad (1)$$

where $s, t_1, \cdots, t_{i-1}$ denotes the concatenation of s and $t_1, \cdots, t_{i-1}$. GPT (Radford et al.) is a

Split	# Dialogs	# Utterances	# Pairs
Train	2,725,990	9,006,966	4,503,483
Validation	340,749	1,127,150	563,575
Test	340,755	1,126,448	563,224

Table 4: The split statistics of the Chinese MedDialog dataset.

pretrained language model which uses the Transformer decoder to model the conditional probability $p(t_i|s, t_1, \cdots, t_{i-1})$ in Eq.(1), which first encodes the tokens in $s, t_1, \cdots, t_{i-1}$, then predicts t_i based on the encodings. GPT learns the weights of the decoder by maximizing the likelihood (defined based on Eq.1) on the responses in the training data.

4.3 Pretraining

Before training Transformer and GPT on the MedDialog-CN dataset, we can first pretrain them on general-domain text datasets which are much larger than MedDialog-CN, to get a good initialization of the weight parameters. BERT-GPT (Wu et al., 2019; Lewis et al., 2019) is a pretraining approach of Transformer, which uses BERT (Devlin et al., 2018) to pretrain the Transformer encoder and uses GPT to pretrain the Transformer decoder. Given a sequence of tokens, BERT randomly marks out some of them. The masked sequence is fed into the transformer encoder, which aims to recover the masked tokens. The weights in the encoder are learned by maximizing the accuracy of recovery. In BERT-GPT, the BERT encoder generates representation of the input sequence, which is then fed into the GPT decoder to generate the response.

5 Experiments

5.1 Experiments on the Chinese MedDialog dataset

5.1.1 Experimental Settings

We split the Chinese MedDialog dataset into a training set, a validation set, and a test set with a ratio of 0.8:0.1:0.1. The split was based on dialogues, not based on source-target pairs. The split statistics are summarized in Table 4. The models were built at the Chinese character level. The validation set was used for hyperparameter tuning. The training procedure was stopped when the validation loss stopped to decrease. For Transformer, the implementation by HuggingFace[5] was used, where the hyperparameters followed the default settings in the original Transformer (Vaswani et al., 2017). In

[5]https://github.com/huggingface/transformers

9245

BERT-GPT, the BERT encoder and GPT decoder are Transformers with 12 layers. The hidden state size is 768. The optimization of weight parameters was performed using stochastic gradient descent, with a learning rate of 1e-4. The maximum length of input sequences was truncated to 400 and that of output sequences was truncated to 100. For GPT, the DialoGPT-small (Zhang et al., 2019) architecture was used, with 10 layers. We set the embedding size to 768 and the context size to 300. In layer normalization, the epsilon hyperparameter was set to 1e-5. In multi-head self-attention, we set the number of heads to 12. The weight parameters were learned with Adam (Kingma and Ba, 2014). The initial learning rate was set to 1.5e-4 and the batch size was set to 32. The learning rate scheduler was set to Noam, with 2000 warm-up steps. Top-k random sampling (Fan et al., 2018) with $k = 50$ was used for decoding in all methods. 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 to 4), BLEU-n (Papineni et al., 2002b) (where n is set to 2 and 4), METEOR (Lavie and Agarwal, 2007), Entropy-n (Zhang et al., 2018) (where n is set to 4), and Dist-n (Li et al., 2015) (where n is set to 1 and 2). Perplexity measures the language quality of the generated responses. The lower, the better. NIST, BLEU, and METEOR measure the similarity between the generated responses and groundtruth via n-gram matching. The higher, the better. Entropy and Dist measure the lexical diversity of generated responses. The higher, the better.

BERT-GPT is pretrained on Chinese corpus collected from the Large Scale Chinese Corpus for NLP[6]. The corpus includes Chinese Wikipedia containing 104 million documents, News containing 2.5 million news articles from 63,000 sources, Community QA containing 4.1 million documents belonging to 28 thousand topics, and Baike QA containing 1.5 million question-answering pairs from 493 domains. The total size of these datasets is 15.4 GB. GPT is pretrained on Chinese Chatbot Corpus[7] containing 14 million dialogues and 500k-Chinese-Dialog[8] containing 500K Chinese dialogues.

---

[6] https://github.com/brightmart/nlp_chinese_corpus
[7] https://github.com/codemayq/chinese_chatbot_corpus
[8] https://drive.google.com/file/d/1nEuew_KNpTMbyy7BO4c8bXMXN351RCPp/view

	Transformer	BERT-GPT	GPT
Perplexity	9.5	**8.2**	9.7
NIST-4	**0.39**	0.31	0.36
BLEU-2	4.9%	3.7%	**5.0**%
BLEU-4	0.9%	0.5%	**1.8**%
METEOR	**13.1**%	10.4%	12.1%
Entropy-4	13.5	**13.6**	**13.6**
Dist-1	**0.03**%	0.02%	0.02%
Dist-2	2.0%	**2.1**%	2.0%
Avg Len	27.9	27.3	28.3

Table 5: Performance on the MedDialog-CN test set.

5.1.2 Results

Table 5 shows the performance on the MedDialog-CN test set. From this table, we make the following observations. First, BERT-GPT achieves lower perplexity than Transformer. This is because BERT-GPT is pretrained on a large collection of corpora before being finetuned on MedDialog-CN. Pretraining enables the model to better capture the linguistic structure among words, which yields lower perplexity. Second, on machine translation metrics including NIST-4, BLEU-2, BLEU-4, and METEOR, BERT-GPT performs worse than Transformer. This indicates that Transformer is able to generate responses that have more overlap with the groundtruth. However, it is worth noting that the studies in (Liu et al., 2016) show that machine translation metrics are not reliable evaluation metrics for dialogue generation. Given the same conversation history, many responses are valid. A response should not be deemed as bad simply because it has little overlap with the response given by a doctor. Third, on diversity metrics, BERT-GPT and Transformer are on par, which indicates that they have similar capability in generating diverse responses. Fourth, compared with BERT-GPT, GPT has worse perplexity, better machine translation scores, and comparable diversity scores.

Figure 2 shows an example of generated responses on the MedDialog-CN test set. The response generated by BERT-GPT is clinically informative and accurate. It prescribes Ebastine and gives detailed instructions of taking this medication. Ebastine is a medication for treating eczema. The patient mentioned that his/her baby has eczema. So this prescription is clinically meaningful. The language quality of the response is also good. It is syntactically and semantically correct and smooth. The response generated by GPT is also good, but

Conversation history: 所谓的湿疹, 孩子从一周岁后起, 到换季的时候, 就周身发痒, 用手一抓破, 就一直是红色的, 水水的, 要过几长时间才会结痂, 留有黑色印迹. (Eczema. Since one year old, whenever changing seasons, the baby feels itchy all over the body. The pimples break when scratched. They are always red and swelling. It takes a long time to scab, leaving a black mark.)
Groundtruth response: 建议到江苏省皮肤病研究所就诊. (I suggest to visit the Dermatology Research Institute of Jiangsu Province.)
Transformer: 需要看到皮肤的形态才能给予具体建议. (I need to see the skin to give detailed suggestions.)
GPT: 如果是过敏性皮肤，建议外用酮康唑乳膏. (If the skin is allergic, can take some Ketoconazole cream.)
BERT-GPT: 你去买一种依巴斯汀一天晚上吃一片试试看. (Can use Ebastine, one tablet every day at bedtime.)

Figure 2: An example of generated responses on the MedDialog-CN test set.

less specific. It believes the baby has a skin allergy issue, but does not pinpoint the exact issue as BERT-GPT does. The response generated by Transformer is less clinically informative. It does not give medical suggestions. But it asks for further information, which is also a valid response.

Figure 3 shows another example. The response generated by BERT-GPT is clinically accurate and concise. The language quality is great. The response generated by GPT is self-conflicting. It says "if there is no abnormality at the throat, you can take a laryngoscope test; if abnormal, you should take a laryngoscope test", which is semantically inconsistent. The response generated by Transformer prescribes two repetitive laryngoscope tests, which is clinically insensible.

5.2 Transfer to Other Datasets

In this section, we study how to use the models pretrained on MedDialog-CN to improve the performance on low-resource dialogue generation tasks where the dataset size is small. The target task is generating medical dialogues related to COVID-19 on the small-sized CovidDialog-Chinese (Yang et al., 2020) dataset. We finetune the MedDialog-pretrained models on CovidDialog-Chinese, and use the finetuned models to generate COVID-19-related dialogues.

5.2.1 Data

We use a Chinese dialogue dataset about COVID-19: CovidDialog-Chinese (Yang et al., 2020), for the experiments. This dataset has 1,088 patient-doctor dialogues about COVID-19, with 9,494 utterances and 406,550 tokens (Chinese characters)

Conversation history: 声音嘶哑, 几年了, 不时发作. (My voice is broken. In the past few years, it happens oftentimes.)
Groundtruth response: 手术方式: 门诊表麻电子喉镜下摘除(小息肉)或住院全麻支撑喉镜下显微镜下摘除(大息肉). (Surgery choices: outpatient, surface anesthetic, remove small polyps using electrolaryngoscope; inpatient, general anesthesia, remove large polyps using laryngoscope and microscope.)
Transformer: 建议到当地儿童医院进行喉镜检查, 然后到比较好的医院进行喉镜检查, 以便确诊和选择治疗. (I suggest to take a laryngoscope test at a local Children's hospital. Then take a laryngoscope test at a better hospital, to confirm and select the best way for treatment.)
GPT: 你好, 声音嘶哑原因较多, 不知你的是喉部还是咽喉部? 如果喉部没有异常, 建议做电子喉镜检查, 如果有异常, 应该做喉镜检查. *声带是指哪些部位的病变? (Hello. Broken voice can be caused by many things. Is your problem at the throat or pharynx? If there is no abnormality at the throat, you can take a laryngoscope test. If abnormal, you should take a laryngoscope test.
BERT-GPT: 您好，建议先到当地医院做一个电子喉镜检查. (Hello. I suggest to take a laryngoscope test at a local hospital.)

Figure 3: Another example of generated responses on the MedDialog-CN test set.

# of dialogues	1,088
# of utterances	9,494
# of tokens	406,550
Avg. # of utterances in a dialogue	8.7
Max # of utterances in a dialogue	116
Min # of utterances in a dialogue	2
Avg. # of tokens in an utterance	42.8
Max # of tokens in an utterance	2,001
Min # of tokens in an utterance	1

Table 6: Statistics of the CovidDialog-Chinese dataset.

in total. Duplicated and incomplete dialogues were removed. The dialogues are multi-turn. The average number of utterances in a dialogue is 8.7. The utterances are reasonably long. The average number of tokens in an utterance is 42.8. Table 6 shows the statistics of this dataset.

5.2.2 Experimental settings

We split the CovidDialog-Chinese dataset into a training set, a validation set, and a test set with a ratio of 0.8:0.1:0.1. The split is based on dialogues. The split statistics are summarized in Table 7. Most hyperparameter settings follow those in Section 4.1, except the following: in optimization, the batch size was set to 8. We evaluate the trained models using automatic metrics including perplexity, NIST-4 (Doddington, 2002), BLEU-2, 4 (Pa-

Split	# Dialogs	# Utterances	# Pairs
Train	870	7,844	3,922
Validation	109	734	367
Test	109	916	458

Table 7: CovidDialog-Chinese split statistics.

Conversation history: 刚刚我坐公交车时候车里后面有人脱下口罩打喷嚏, 我有戴口罩, 普通口罩, 公交车里面还开着空调, 有没有危险? 和风险? (Just now, when I was taking a bus, someone took off his facial mask and sneezed. I was wearing an ordinary mask. The air conditioner in the bus was on. Is this risky?)

Groundtruth response: 你们当地有没有确诊病例? (Are there confirmed cases in your city?)

Transformer without pretraining on MedDialog: 感谢您的信任, 病情资料我已详细阅读. (Thank you for your trust. I have read your medical record.)

Transformer with pretraining on MedDialog: 不舒服吗? (Are you feeling unconforable?)

GPT without pretraining on MedDialog: *我已, 我们在有是不是否么感染. 根据你现有的情况.

GPT with pretraining on MedDialog: 有可能性, 没有风险, 多饮水, 别太担心. (It is likely. No risk. Drink more water. Don't worry too much.)

BERT-GPT without pretraining on MedDialog: 就是没有风险, 可以保险点, 但不一定要全部戴口罩. (No risk. Try to be careful, but it is not necessary to wear mask.)

BERT-GPT with pretraining on MedDialog: 你的年纪?目前的呼吸功能好吗?有没有胸部 CT 的报告? (How old are you? Is your breath OK? Do you have chest CT report?)

Figure 4: An example of generated responses about COVID-19.

Conversation history: 不发烧, 就是锁骨中间不舒服, 有点想咳嗽, 躺下就会感觉好很多, 站起来就更想咳嗽一点, 没有痰, 没去过武汉, 没接触过武汉来的人. (No fever. It is kind of uncomfortable in the middle of clavicle. Have the feeling to cough. Feel much better when lying down. When I stand up, the feeling to cough becomes stronger. No phlegm. Didn't visit Wuhan or contact people from Wuhan.)

Groundtruth response: 发病前 14 天有没有接触湖北地区或者是疫情高发省份来的人? 或者是你所在的小区有没有疑似或者确诊病例? 发热咳嗽咽痛病人最好排除下冠状病毒. (Have you contacted people from Hubei or other provinces that have severe epidemic? Does your community have suspected or confirmed cases? If you have fever, cough, sore throat, it's better to get a test.)

Transformer without pretraining on MedDialog: 你好, 我已经仔细研读了您的资料, 请您现在您现有没有接触史吧. (Hello. I have carefully read your record. Do you have a history of exposure?)

Transformer with pretraining on MedDialog: 请问您有咽喉疾病病史吗? (Do you have medical history of throat disease?)

GPT without pretraining on MedDialog: *你好, 不会的症状, 如果, 我就是新型冠状病毒感染有没有接触到感冒.好, 如果不客气, 也不会有感染, 我就是新型冠状病毒感染的症状?

GPT with pretraining on MedDialog: 从你提供的检查结果来看, 没有发现明显异常. 不发烧不咳嗽不腹泻, 不腹泻, 不考虑是新冠肺炎, 除非有新型冠状病毒感染, 还有说明书, 如果仅仅是普通的, 都没问题. (I didn't see abnormality in the test results that you provided. If no fever, no cough, no diarrhea, no diarrhea, then it is unlikely to be COVID-19, unless infected by coronavirus. And manual. If ordinary, then there is no problem.)

BERT-GPT without pretraining on MedDialog: 咳嗽吗? 你好. 咳嗽考虑支气管炎, 建议口服莲花清瘟颗粒. (Do you cough. Hello. If having cough, it may be bronchitis. You can take some Lianhuaqingwen pills orally.)

BERT-GPT with pretraining on MedDialog: 感冒可能性大, 不过有些病毒或者细菌感染患者出现这种状况比较常见, 建议多喝水, 必要时去医院检查一下. (It is more likely to be cold. But people infected by virus or bacteria may have such conditions as well. Drink more water. Visit the hospital if necessary.)

Figure 5: The second example of generated responses about COVID-19.

pineni et al., 2002b), METEOR (Lavie and Agarwal, 2007), Entropy-4 (Zhang et al., 2018), and Dist-1, 2 (Li et al., 2015). We also perform human evaluation. We randomly select 100 dialog examples and ask 5 undergraduate and graduate students to rate the generated responses in terms of informativeness, relevance, and human-likeness. Informativeness is about whether a response contains sufficient medical information such as explanations of diseases and suggestions for treatment. Relevance is about whether the content of a response matches with that of the conversation history. Human-likeness is about whether a response sounds like a human. The ratings are from 1 to 5. The higher, the better. The ratings from different annotators are averaged as the final results.

5.2.3 Results

Table 8 shows the automatic evaluation results achieved by different models, with or without pretraining (PT) on MedDialog-CN. From this table, we make the following observations. First, on
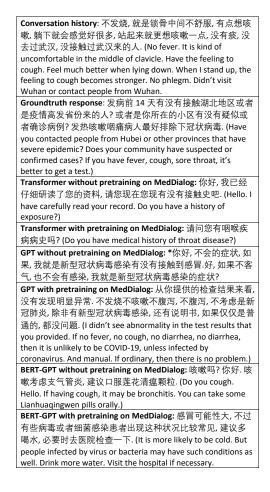
Transformer, pretraining on MedDialog-CN improves results on all metrics. This demonstrates that pretraining on MedDialog-CN can improve performance on low-resource medical dialog generation tasks. Second, on GPT, pretraining on MedDialog-CN improves 5 of the 8 metrics. On BERT-GPT, pretraining on MedDialog-CN improves half of metrics. The reason that improvement on GPT and BERT-GPT is not as significant as that on Transformer is probably because these two models are already pretrained using other corpora. Therefore the value of pretraining on MedDialog-CN is diminishing. However, it is still useful to pretrain on MedDialog-CN to adapt these two models to the medical dialog domain.

Table 9 shows the human evaluation results on the test set of CovidDialog-Chinese. From this table, we can see that on all models, pretraining on MedDialog-CN improve relevance, informativeness, and human-likeness. This further demon-

	Perplexity	NIST-4	BLEU-2	BLEU-4	METEOR	Entropy-4	Dist-1	Dist-2	Avg. Len
Transformer, no PT	53.3	0.39	5.7%	4.0%	13.5%	7.9	5.5%	29.0%	19.3
Transformer, PT	13.7	**0.50**	**7.8**%	**4.7**%	**16.0**%	8.0	7.6%	36.3%	22.0
GPT, no PT	22.1	0.43	6.2%	4.0%	13.9%	**9.0**	5.9%	38.7%	35.0
GPT, PT	**8.9**	0.40	7.0%	4.0%	14.8	8.7	7.4%	39.7%	28.9
BERT-GPT, no PT	10.8	0.36	4.6%	2.8%	12.2%	8.5	7.9%	39.5%	21.6
BERT-GPT, PT	10.2	0.33	5.0%	2.7%	11.2%	8.4	**8.6**%	**43.3**%	21.4

Table 8: Automatic evaluation results on the CovidDialog-Chinese test set.

	Transformer		GPT		BERT-GPT		Groundtruth
	No PT	PT	No PT	PT	No PT	PT	
Relevance	2.25	2.68	1.82	2.74	2.65	2.93	3.42
Informativeness	2.06	2.40	1.72	2.53	2.37	2.77	3.26
Human-likeness	2.57	3.29	1.80	3.20	3.16	3.44	3.78

Table 9: Human evaluation results on the CovidDialog-Chinese test set.

	Transformer	GPT	BERT-GPT	Groundtruth
	No-PT vs PT	No-PT vs PT	No-PT vs PT	vs BERT-GPT
Relevance	0.006	0.008	0.004	0.003
Informativeness	0.014	0.004	0.003	0.004
Human-likeness	0.009	0.001	0.031	0.036

Table 10: Significance tests on human evaluation results.

strates the effectiveness of pretraining. We perform significance tests between different methods based on the double-sided Student's t-test. The results are shown in Table 10. As can be seen, in most cases, the p-value is less than 0.015, demonstrating high statistical significance. For Transformer, GPT, and BERT-GPT, using pretraining (PT) on MedDialog-CN achieves significantly better performance than not using pretraining (No-PT).

Figure 4 shows an example of generating a doctor's response given the utterance of a patient. As can be seen, models pretrained on MedDialog-CN perform better than their unpretrained counterparts. For example, the response generated by GPT without pretraining on MedDialog-CN is not understandable by human. With pretraining on MedDialog-CN, it generates a much better response which gives medical advice. Figure 5 shows another example. Similarly, without MedDialog-pretraining, the response generated by GPT is not readable. With pretraining, the generated response is smooth and clinically informative.

6 Conclusions and Future Works

To facilitate the research and development of medical dialogue systems that can potentially assist in telemedicine, we build large-scale medical dialogue datasets – MedDialog – which contain 1) a Chinese dataset with 3.4 million conversations between patients and doctors, 11.3 million utterances, 660.2 million tokens, covering 172 specialties of diseases, and 2) an English dataset with 0.26 million conversations, 0.51 million utterances, 44.53 million tokens, covering 96 specialties of diseases. To our best knowledge, they the largest of their kind. We pretrain Transformer, GPT, and BERT-GPT on MedDialog-CN. The results show that the generated dialogues by these pretrained models are clinically meaningful and human-like. We use transfer learning to apply these pretrained models for low-resource dialogue generation. On a COVID-19 dialogue generation task where the dataset is small, human evaluation and automatic evaluation show that models pretrained on MedDialog-CN can effectively improve the quality of generated responses.

For future work, we will annotate medical entities in our datasets. Such annotations can facilitate the development of goal-oriented medical dialog systems.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

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*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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*, pages 228–231.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Chia-Wei Liu, Ryan Lowe, Iulian V 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. *arXiv preprint arXiv:1603.08023*.

Reena L Pande and Michael Morris. 2015. Leveraging remote behavioral health interventions to improve medical outcomes and reduce costs. *Am J Manag Care*, 21(2):e000–e000.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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, 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.

Richard Wootton, John Craig, and Victor Patterson. 2017. *Introduction to telemedicine*. CRC Press.

Qingyang Wu, Lei Li, Hao Zhou, Ying Zeng, and Zhou Yu. 2019. Importance-aware learning for neural headline editing. *arXiv preprint arXiv:1912.01114*.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis.

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.

Wenmian Yang, Guangtao Zeng, Bowen Tan, Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, Qingyang Wu, Zhou Yu, et al. 2020. On the generation of medical dialogues for covid-19. *arXiv preprint arXiv:2005.05442*.

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 *NeurIPS*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.