On the Generation of Medical Dialogues for COVID-19

Wenmian Yang[1*], Guangtao Zeng[1*†], Bowen Tan[2*], Zeqian Ju[1†], Subrato Chakravorty[1], Xuehai He[1], Shu Chen[1†], Xingyi Yang[1], Qingyang Wu[3], Zhou Yu[3], Eric Xing[2], Pengtao Xie[1]

UC San Diego[1], CMU[2], UC Davis[3]

PENGTAOXIE2008@GMAIL.COM

Abstract

Under the pandemic of COVID-19, people experiencing COVID19-related symptoms or exposed to risk factors have a pressing need to consult doctors. Due to hospital closure, a lot of consulting services have been moved online. Because of the shortage of medical professionals, many people cannot receive online consultations timely. To address this problem, we aim to develop a medical dialogue system that can provide COVID19-related consultations. We collected two dialogue datasets – CovidDialog – (in English and Chinese respectively) containing conversations between doctors and patients about COVID-19. On these two datasets, we train several dialogue generation models based on Transformer, GPT, and BERT-GPT. Since the two COVID-19 dialogue datasets are small in size, which bear high risk of overfitting, we leverage transfer learning to mitigate data deficiency. Specifically, we take the pretrained models of Transformer, GPT, and BERT-GPT on dialog datasets and other large-scale texts, then finetune them on our CovidDialog datasets. Experiments demonstrate that these approaches are promising in generating meaningful medical dialogues about COVID-19. But more advanced approaches are needed to build a fully useful dialogue system that can offer accurate COVID-related consultations. The data and code are available at https://github.com/UCSD-AI4H/COVID-Dialogue

1. Introduction

As of May 8th in 2020, the COVID-19 pandemic has killed 272,778 people out of 3,910,738 infected cases. People who are experiencing symptoms (e.g., fever, cough) similar to those of COVID-19 or were exposed to risk factors such as close contact with infected cases have a pressing need to consult doctors, largely because of the panic over this unknown new disease. However, under the pandemic situation, coming to hospitals is dangerous and has high risk of suffering cross-infection. Cross-infection refers to the fact that many people visiting hospitals at the same time and infected individuals will spread coronavirus to healthy ones. To prevent spreading of the coronavirus, many non-urgent clinics and hospitals have been closed physically and encourage people to consult doctors through telemedicine services (e.g., phone calls, video conferencing). However, medical professionals are highly occupied by taking care of the infected patients and have very thin bandwidth to deal with the surging requests of consultations related to COVID-19. As a result, many people could not receive timely advice for effectively dealing with their medical conditions.

---

0. *Equal contribution
0. †The work was done during internship at UCSD.

To address the large imbalance between the surging need of consultations from citizens and the severe shortage of medical professionals available to provide online consultation services, it is highly valuable to develop intelligent dialogue systems which act as "virtual doctors" to provide COVID-related consultations to people. These "virtual doctors" can greatly ease the burden of human doctors and timely address the concerns of the public.

To facilitate the research and development of COVID19-targeted dialogue systems, we build two medical dialogue datasets that contain conversations between doctors and patients, about COVID-19 and other pneumonia: (1) an English dataset containing 603 consultations, 1232 utterances, and 90664 tokens (English words); (2) a Chinese dataset containing 1088 consultations, 9494 utterances, and 406550 tokens (Chinese characters).

On these two datasets, we train several dialogue generation models based on Transformer (Vaswani et al., 2017), GPT (Radford et al., a; Zhang et al., 2019), and BERT-GPT (Wu et al., 2019; Lewis et al., 2019). Transformer is an encoder and decoder architecture which takes the conversation history as inputs and generates the response. Self-attention is used to capture the long-range dependency among tokens. GPT is a language model based on the Transformer decoder. When generating a response, GPT predicts the next token using its context including the already decoded tokens in this response and the conversation history. BERT-GPT is an encoder-decoder architecture as well where the pretrained BERT (Devlin et al., 2018) is used to encode the conversation history and GPT is used to decode the response. The small size of CovidDialog datasets incurs high risk of overfitting, if directly training the large-sized neural models on CovidDialog. To alleviate this risk, we take the pretrained weights of these models on large-scale dialogue dataset and other corpus and finetune the weights on CovidDialog. Experiments demonstrate that the models trained on CovidDialog datasets are promising in generating clinically meaningful consultations about COVID-19. The datasets and code are publicly available at https://github.com/UCSD-AI4H/COVID-Dialogue

The rest of the papers are organized as follows. Section 2 and 3 present the datasets and methods. Section 4 gives experimental results. Section 5 reviews related works and Section 6 concludes the paper.

2. Dataset

In this section, we present two collected datasets – CovidDialog-English and CovidDialog-Chinese – which contain medical conversations between patients and doctors about COVID-19 and other related pneumonia. The statistics of these two datasets are summarized in Table 1.

2.1. The English Dataset

The CovidDialog-English dataset contains 603 English consultations about COVID-19 and other related pneumonia, having 1,232 utterances. The number of tokens (English words) is 90,664. The average, maximum, and minimum number of utterances in a conversation is 2.0, 17, and 2 respectively. The average, maximum, and minimum number of tokens in an utterance is 49.8, 339, and 2 respectively. Each consultation starts with a short description of the medical conditions of a patient, followed by the conversation between the patient

Description of medical condition

I have a little fever with no history of foreign travel or contact. What is the chance of Covid-19?

Dialogue

Patient: Hello doctor, I am suffering from coughing, throat infection from last week. At that time fever did not persist and also did not felt any chest pain. Two days later, I consulted with a doctor. He prescribed Cavidur 625, Montek LC, Ambrolite syrup and Betaline gargle solution. Since then throat infection improved and frequent cough also coming out. Coughing also improved remarkably though not completely. From yesterday onwards fever is occuring (maximum 100-degree Celcius). I have not come in touch with any foreign returned person nor went outside. In our state, there is no incidence of Covid-19. Please suggest what to do?

Doctor: Hello, I can understand your concern. In my opinion, you should get done a chest x-ray and CBC (complete blood count). If both these are normal then no need to worry much. I hope this helps.

Patient: Thank you doctor. After doing all these I can upload all for further query.

Doctor: Hi, yes, upload in this query only. I will see and revert to you.

Patient: Thank you doctor. As per your guidelines, I have made one test. Due to city shutdown, I could not able to make an x-ray of chest. Fever is coming every 12 hours. But cough does not persist currently. The only problem is a fever. I am uploading the blood examination report. Please advice.

Doctor: Hi, I can understand your concern. I have gone through the report you have attached. (attachment removed to protect patient identity). Your total count is on the higher side of normal along with low hemoglobin. So start Azithromycin 500 mg 1-0-0, Tablet Dolo 650 1-0-1. Once your fever and cough subside, start an Iron tablet for low hemoglobin. I hope this helps.

Patient: Thank you doctor. How many tablets of Azithromycin to be taken? Is it in the morning? I want to make you know that I have already taken Azithromycin five days before. Should she start it? And also can I start an Iron tablet today itself?

Doctor: Hi, Yes, you can take iron tablet from today onwards. And no need for Azithromycin now as you have already taken it. Drink plenty of fluids orally and keep yourself hydrated. Do warm water gargles and steam inhalation four to five times a day.

Figure 1: An exemplar consultation in the CovidDialog-English dataset. It consists of a brief description of the patient's medical conditions and the conversation between the patient and a doctor.

Description of medical conditions and history

疾病: 肺炎相关
(Disease: Pneumonia related)
病情描述: 咳嗽从 2 月开始到现在，有痰，怀疑是不是新冠状病毒？
(Description of medical condition: Cough from February to now. There is sputum. Am I infected by novel coronavirus?)
希望获得的帮助: 能不能医好这咳嗽病？
(Help needed: How to cure the cough?)
患病多久: 半年内
(Hong long the condition has been: Less than 6 months)
用药情况：
17：盐酸氨溴索片，头孢克肟片，肺力咳合剂；
21：盐酸氨溴索胶囊，头孢克肟片，喉疾灵胶囊，盐酸依匹斯汀胶囊；
23：盐酸依匹斯汀胶囊，盐酸克林霉素棕榈酸酯分散片，盐酸氨溴索口服溶液，喉疾灵胶囊；
26：肺力咳合剂，盐酸氨溴索胶囊，盐酸克林霉素棕榈酯分散片，复方甲氧那明胶囊，乙酰半胱氨酸颗粒；
(Medications:
17: Ambroxol hydrochloride tablets, Cefixime tablets, Feilike mixture;
21: Ambroxol hydrochloride capsules, Cefixime tablets, Houjiling capsules, and Epinastine hydrochloride capsules;
23: Epinastine hydrochloride capsules, Clindamycin hydrochloride palmitate dispersible tablets, Ambroxol hydrochloride oral solution, Houjiling capsules;
26: Feilike Mixture, Ambroxol Hydrochloride Capsules, Clindamycin Palmitate Dispersible Tablets, Compound Methionine Capsules, Acetylcysteine Granules;)
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。是新冠活不了这么久的
(Doctor: Thank you for your trust. I have read the medical information in detail. If it was because of the coronavirus, you couldn't have lived that long.)
病人: 但是我吃了很多药都不见好，怎么办？
(Patient: But I took a lot of medications and it didn't work, what should I do?)
医生: 用点抗过敏药，比如西替利嗪
(Doctor: Use some antiallergic drugs, such as cetirizine.)
病人: 西替利嗪口服液？
(Patient: Cetirizine liquid?)
医生: 片
(Doctor: Tablets.)
病人: 我要不要去检查一下核酸？
(Patient: Should I get a RT-PCR test?)
医生: 不要浪费国家的资源了，不可能的
(Doctor: Impossible to be COVID. No need to do the test.)
病人: 我现在一动就会出很多汗，呼吸喉咙有点痒，咳痰。
(Patient: I sweat a lot when I move. The throat is itchy. There is sputum.)
医生: 你家几个人和你一样症状？
(Doctor: How many people in your family have the same symptoms as you?)
病人: 没有
(Patient: None)
……

Diagnosis and suggestions

病情摘要及初步印象: 过敏性咳嗽，焦虑抑郁状态
 (Summary of the condition and initial impressions: Allergic cough, anxiety and depression.)
总结建议: 西替利嗪片口服，必要时心理科就诊。
 (Summary of recommendations: Take cetirizine tablets orally, if necessary, see a psychologist.)

Figure 2: An exemplar consultation in the CovidDialog-Chinese dataset. It consists of (1) description of medical conditions and history of the patient, (2) dialogue between doctor and patient, and (3) diagnosis and treatment suggestions given by the doctor.

ON THE GENERATION OF MEDICAL DIALOGUES FOR COVID-19

	English	Chinese
Number of dialogues	603	1088
Number of utterances	1,232	9,494
Number of tokens	90,664	406,550
Average number of utterances in a dialogue	2.0	8.7
Maximum number of utterances in a dialogue	17	116
Minimum number of utterances in a dialogue	2	2
Average number of tokens in an utterance	49.8	42.8
Maximum number of tokens in an utterance	339	2001
Minimum number of tokens in an utterance	2	1

Table 1: Statistics of the English and Chinese dialogue datasets about COVID-19.

and a doctor. Figure 1 shows an example. The original dialogues are crawled from online healthcare forums, including icliniq.com[1], healthcaremagic.com[2], and healthtap.com[3].

2.2. The Chinese Dataset

The CovidDialog-Chinese dataset contains 1,088 Chinese consultations about COVID-19 and other related pneumonia, having 9,494 utterances. In this work, we develop models directly on Chinese characters without performing word segmentation. Each Chinese character in the text is treated as a token. The total number of tokens in the dataset is 406,550. The average, maximum, and minimum number of utterances in a conversation is 8.7, 116, and 2 respectively. The average, maximum, and minimum number of tokens in an utterance is 42.8, 2001, and 1 respectively. Each consultation consists of three parts: (1) description of patient's medical condition and history; (2) conversation between patient and doctor; (3) (optional) diagnosis and treatment suggestions given by the doctor. In the description of the patient's medical condition and history, the following fields are included: present disease, detailed description of present disease, what help is needed from the doctor, how long the disease has been, medications, allergies, and past diseases. This description is used as the first utterance from the patient. Figure 2 shows an exemplar consultation. The data is crawled from haodf.com[4], which is an online platform of healthcare services, including medical consultation, scheduling appointments with doctors, etc. Duplicated and incomplete dialogues were removed.

3. Methods

In this section, we present several well-established and state-of-the-art methods for dialogue generation. 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. The

---

1. https://www.icliniq.com/en_US/
2. https://www.healthcaremagic.com/
3. https://www.healthtap.com/
4. https://www.haodf.com/

size of the CovidDialog datasets is small. Directly training neural models on these small datasets would result in poor generalization on unseen data. To solve this problem, we utilize transfer learning, which pretrains the neural models on large corpus, then finetunes the pretrained models on the CovidDialog datasets.

3.1. Transformer

Generating response t from the conversation history s is a typical sequence-to-sequence (seq2seq) (Sutskever et al., 2014) modeling problem. Transformer (Vaswani et al., 2017) is an encoder-decoder architecture for sequence-to-sequence (seq2seq) modeling. Different from seq2seq models (Sutskever et al., 2014) that are based on recurrent neural networks (e.g., LSTM (Hochreiter and Schmidhuber, 1997), GRU (Chung et al., 2014)) which model a sequence of tokens via a recurrent manner and hence is computationally inefficient. Transformer eschews recurrent computation and instead uses self-attention which not only can capture the dependency between tokens but also is amenable for parallel computation with high efficiency. Self-attention calculates the correlation among every pair of tokens and uses these correlation scores to create "attentive" representations by taking weighted summation of tokens' embeddings. Transformer is composed of a stack of building blocks, each consisting of a self-attention layer and a position-wise feed-forward layer. Residual connection (He et al., 2016) is applied around each of the two sub-layers, followed by layer normalization (Ba et al., 2016). Given the input sequence, an encoder, which is a stack of such building blocks, is applied to obtain a representation for each token. Then the decoder takes these representations as inputs and decodes the sequence of output tokens. To decode the i-th token, the decoder first uses self-attention to encode the already decoded sequence $y_1, \cdots, y_{i-1}$, then performs input-output attention between the encodings of $y_1, \cdots, y_{i-1}$ and those of the input sequence. The "attentive" representations are then fed into a feed-forward layer. The three steps are repeated for multiple times. Finally, the representation is fed into a linear layer to predict the next token. The weight parameters in Transformer is learned by maximizing the conditional likelihood of output sequences conditioned on the corresponding input sequences.

3.2. GPT

The GPT model (Radford et al., a) is a language model (LM) based on Transformer. Different from Transformer which defines a conditional probability on an output sequence given an input sequence, GPT defines a marginal probability on a single sequence. Given a sequence of tokens $x_1, \cdots, x_n$, an LM defines a probability on the sequence:

$$p(x_1, \cdots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_1, \cdots, x_{i-1}), \tag{1}$$

which basically predicts the next token based on the historical sequence. In GPT, $p(x_i | x_1, \cdots, x_{i-1})$ is defined using the Transformer decoder, which first uses a stack of self-attention and feed-forward layers (each followed by layer normalization) to encode $x_1, \cdots, x_{i-1}$, then predicts x_i from the encodings of $x_1, \cdots, x_{i-1}$. The weight parameters are learned by maximizing the likelihood on the sequence of tokens. GPT-2 (Radford et al., b) is an extension

of GPT, which modifies GPT by moving layer normalization to the input of each sub-block and adding an additional layer normalization after the final self-attention block. Byte pair encoding (BPE) (Sennrich et al., 2015) is used to represent the input sequence of tokens.

Pretrained GPT models for dialogue generation   DialoGPT (Zhang et al., 2019) is a GPT-2 model pretrained on English Reddit dialogues. The dataset is extracted from comment chains in Reddit from 2005 till 2017, comprising 147,116,725 dialogue instances with 1.8 billion tokens. Given a dialogue history S and a ground-truth response $T = x_1, \cdots, x_n$, DialoGPT is trained to maximize the following probability

$$p(T|S) = p(x_1|S) \prod_{i=2}^{n} p(x_i|S, x_1, \cdots, x_{i-1}) \tag{2}$$

where the conditional probabilities are defined by the Transformer decoder. A maximum mutual information (MMI) (Li et al., 2015) scoring function is used to penalize generated responses that are bland. We finetune DialoGPT on our CovidDialog-English dataset for generating English COVID-19 dialogues. GPT2-chitchat[5] is a GPT-2 model pretrained on Chinese Chatbot Corpus[6] which contains about 14M dialogues and 500k-Chinese-Dialog[7] which contains 500K Chinese dialogues. The training strategy of GPT2-chitchat is the same as that of DialoGPT. We finetune GPT2-chitchat on our CovidDialog-Chinese dataset for generating Chinese COVID-19 dialogues.

3.3. BERT-GPT

BERT-GPT (Wu et al., 2019) is a model used for dialogue generation where pretrained BERT is used to encode the conversation history and GPT is used to generate the responses. While GPT focuses on learning a Transformer decoder for text generation purposes, BERT (Devlin et al., 2018) aims to learn a Transformer encoder for representing texts. BERT's model architecture is a multi-layer bidirectional Transformer encoder. In BERT, the Transformer uses bidirectional self-attention, whereas in GPT every token can only attend to context to its left. To train the encoder, BERT masks some percentage of the input tokens at random, and then predict those masked tokens by feeding the final hidden vectors (produced by the encoder) corresponding to the mask tokens into an output softmax over the vocabulary. Since BERT leverages context to both the left and the right for representing a token, it presumably has better representation power than GPT which only leverages context to the left. In dialogue generation, for the given conversation history, instead of using GPT for obtaining the representation, we can use a more powerful pretrained BERT to encode it. The BERT encoding of the conversation history is fed into GPT to generate the response.

In BERT-GPT, the pretraining of the BERT encoder and the GPT decoder is conducted separately, which may lead to inferior performance. Auto-Regressive Transformers (BART) (Lewis et al., 2019) has a similar architecture as BERT-GPT, but trains the BERT encoder and GPT decoder jointly. To pretrain the BART weights, the input text is corrupted randomly, such as token masking, token deletion, text infilling, etc., then the network

---

5. https://github.com/yangjianxin1/GPT2-chitchat

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

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

is learned to reconstruct the original text. BART is pretrained on the data used in (Liu et al., 2019), consisting of 160Gb of news, books, stories, and web texts.

Pretrained BERT-GPT models for dialogue generation   BERT-GPT-Chinese (Wu et al., 2019) is a BERT-GPT model pretrained on Chinese corpus. For the BERT encoder in BERT-GPT-Chinese, it is set to the Chinese BERT (Cui et al., 2019), which is a large-scale pretrained BERT language model on Chinese texts. For the GPT decoder in BERT-GPT, it has the same architecture as BERT but applies lower-triangular mask for autoregressive text generation. The decoder is initialized with Chinese BERT's weights. Then the decoder is pretrained with a maximum likelihood estimation (MLE) objective on a large-scale multi-domain Chinese corpus. The resulting model consists of a bidirectional Transformer as the encoder, a unidirectional Transformer as the decoder, and an attention mechanism to connect them. The Chinese corpus used for pretraining is collected from the Large Scale Chinese Corpus for NLP[8], including the following datasets: Chinese Wikipedia which contains 104M articles, News which contains 2.5 million news articles from 63,000 sources, Baike QA which is a wiki question answering (QA) dataset with 1.5 million QA pairs from 493 different domains, and Community QA which contains 4.1 million comments and 28 thousand topics. The total size of these datasets is 15.4 GB. We finetune BERT-GPT-Chinese on the CovidDialog-Chinese dataset for Chinese COVID-19 dialogue generation. For English COVID-19 dialogue generation, we finetune the pretrained BART model on the CovidDialog-English dataset.

4. Experiments

4.1. Experiments on the English Dataset

4.1.1. Experimental Settings

For the English dataset, we split it into a training, a validation, and a test set based on dialogues, with a ratio of 8:1:1. Table 2 shows the statistics of the data split. The size of the vocabulary (number of unique English words) was set to x. The hyperparameters were tuned on the validation dataset. For all methods, we used the Adam (Kingma and Ba, 2014) optimizer with linear learning rate scheduling, setting the initial learning rate as 4e-5 and the batch size as 4. The objective is the cross entropy loss with label smoothing where the factor was set to 0.1. For pretrained models, we finetune them on the CovidDialog-English dataset for 5 epochs, while for the un-pretrained Transformer, we train it for 50 epochs. We set a checkpoint at the end of every epoch and finally take the one with the lowest perplexity on validation set as the final model. In response generation, for all models, we use beam search with beam width of 10 as our decoding strategy. For DialoGPT (Zhang et al., 2019), we used three variants with different sizes: DialoGPT-small, DialoGPT-medium, DialoGPT-large, with 117M, 345M and 762M weight parameters respectively. Maximum mutual information was not used.

We performed automatic evaluation, using metrics including perplexity, NIST-n (Doddington, 2002) (where $n = 4$), BLEU-n (Papineni et al., 2002) (where $n = 2$ and 4), METEOR (Lavie and Agarwal, 2007), Entropy-n (Zhang et al., 2018) (where $n = 4$), and

---

8. https://github.com/brightmart/nlp_chinese_corpus

ON THE GENERATION OF MEDICAL DIALOGUES FOR COVID-19

Split	#Dialogues	# Utterances	# (Source, Target) Pairs
Train	482	981	490
Validation	60	126	63
Test	61	122	61

Table 2: English dataset split statistics

	Transformer	DialoGPT-small	DialoGPT-medium	DialoGPT-large	BART
Perplexity	263.1	28.3	17.5	18.9	**15.3**
NIST-4	0.71	1.90	2.01	**2.29**	1.88
BLEU-2	7.3%	9.6%	9.4%	**11.5**%	8.9%
BLEU-4	5.2%	6.1%	6.0%	**7.6**%	6.0%
METEOR	5.6%	9.0%	9.5%	**11.0**%	10.3%
Entropy-4	5.0	6.0	**6.6**	**6.6**	6.5
Dist-1	3.7%	9.5%	16.6%	13.9%	**16.8**%
Dist-2	6.4%	22.9%	**36.7**%	31.0%	35.7%
Avg. Len	40.0	51.3	50.1	54.4	45.4

Table 3: Performance on the CovidDialog-English test set.

Dist-n (Li et al., 2015) (where $n = 1$ and 2). BLEU, METEOR, and NIST are common metrics for evaluating machine translation. They compare the similarity between generated responses and the ground-truth by matching n-grams. NIST is a variant of BLEU, which weights n-gram matches using information gain to penalize uninformative n-grams. Perplexity is used to measure the quality and smoothness of generated responses. Entropy and Dist are used to measure the lexical diversity of generated responses. For perplexity, the lower, the better. For other metrics, the higher, the better.

4.1.2. RESULTS

Table 3 summarizes the results achieved by different methods. From this table, we make the following observations. First, pretrained models including DialoGPT and BART in general perform better than un-pretrained Transformer. This demonstrates the effectiveness of transfer learning, which leverages external large-scale data to learn powerful representations of texts. Second, BART achieves lower perplexity than DialoGPT models. This is probably because BART is pretrained on a much larger and more diverse corpus than DialoGPT, which enables BART to better model the language. Third, DialoGPT-large performs better than BART on machine translation metrics including NIST, BLEU, and METEOR. This is probably because DialoGPT-large is pretrained on dialogue data and therefore tends to generate n-grams that are more related to dialogues. Fourth, on diversity-related metrics including Entropy and Dist, BART are on par with DialoGPT models. Note that the comparison between different architectures is not totally fair since they are pretrained on different corpus. Due to the lack of computing resources, we are not able to make a fair comparison by training these architectures on the same corpus. We will leave such a study

9

Conversation history
Patient: i have all the symptoms except fever, i went to medicross and dr said i can get tested if i want to i'm not sure if i should. she gave me antibiotics klacid xl 500mg, she said i can take it if i feel worse i'm worried it will make immune system bad?

Groundtruth response
in brief: antibiotic i don't recommend antibiotics for a simple viral upper respiratory tract infection unless examination revealed signs of acute bronchitis or sinusitis. they are not effective for viral infections like covid 19 with no bacterial lung involvement either. if you've been exposed to someone with covid 19 or or if you or someone you were exposed to travelled to a region where it was endemic, get tested would you like to video or text chat with me?

Generated responses
BART: hello and welcome to 'ask a doctor' service. i have reviewed your query and here is my advice. yes, you should get tested for covid-19. get tested if you have fever, cough or shortness of breath. if you are a smoker or have been in contact with someone with covid, get tested. would you like to video or text chat with me?

DialoGPT-large: do you have access to a telephonic consult, or do you need to go in person? would you like to video or text chat with me? i can understand your concern. i have reviewed your query and here is my advice. yes, telephonically would be best. would you be happy to chat with you? i have answered your query. let me know if i can assist you further.

DialoGPT-medium: in brief: fever, dry cough, shortness of breath, fever can be from strep throat (antibiotics for strep), or from a strep infection, or from some other cause such as allergies, irritants, or irritants. would you like to video or text chat with me?

DialoGPT-small: in brief: yes. if you feel worse you should go to a doctor. would you like to video or text chat with me? let me know if i can assist you with any further questions.

Transformer: hello and welcome to 'ask a doctor' service. i have reviewed your query and here is my advice. i have reviewed your query. let me know if i can assist you further.

Figure 3: An example of generated English responses.

to the future. The average length of the generated responses by different methods is close to that of the ground-truth, which is around 50.

Figure 3 shows an example of generating a doctor's response given the utterance of a patient. As can be seen, the response generated by BART is more relevant, informative, and human-like, compared with those generated by other baselines. BART's response suggests the patient to get tested for COVID-19 since the patient stated that "I have all the symptoms except fever". This response gives correct and informative medical advice: "get tested if you have fever, cough, or shortness of breath", "if you are a smoker or have been in contact with someone with covid, get tested". The response is human-like, with correct grammar

and semantics. It begins with a welcome opening, then provides medical advice, and finally offers to further discuss via video. In contrast, the response generated by DialoGPT-large is not informative. It does not provide any useful medical advice. The response generated by DialoGPT-medium is informative, but not very relevant. The patient has no fever, but this response focuses on talking about the causes of fever. Similar to DialoGPT-large, the responses generated by DialoGPT-small and Transformer are uninformative.

4.2. Experiments on the Chinese Dataset

4.2.1. EXPERIMENTAL SETTINGS

Split	# Dialogues	# Utterances	# (Source, Target) Pairs
Train	870	7844	3922
Validation	109	734	367
Test	109	916	458

Table 4: Chinese dataset split statistics

Based on dialogues, we split the Chinese dataset into a training set, validation set, and test set, with a ratio of 8:1:1. Table 4 shows the statistics of the data split. The vocabulary size (number of unique Chinese characters) was set to 13317. The hyperparameters were tuned on the validation set. We stop the training procedure when the validation loss stops to decrease. For DialoGPT, we used the DialoGPT-small architecture where the number of layers in the Transformer was set to 10. The context size was set to 300. The embedding size was set to 768. The number of heads in multi-head self-attention was set to 12. The epsilon parameter in layer normalization was set to 1e-5. Network weights were optimized with Adam, with an initial learning rate of 1.5e-4 and a batch size of 8. The Noam learning rate scheduler with 2000 warm-up steps was used. In the finetuning of BERT-GPT, the max length of the source sequence and target sequence was set to 400. The encoder and decoder structures are similar to those in BERT, which is a Transformer with 12 layers and the size of the hidden states is 768. The network weights are optimized with stochastic gradient descent with a learning rate of 1e-4. For Transformer, we used the HuggingFace implementation[9] and followed their default hyperparameter settings. During decoding for all methods, beam search with $k = 50$ was used. We evaluated the models using perplexity, NIST-4, BLEU-2, 4, METEOR, Entropy-4, and Dist-1, 2.

4.3. Results on the Chinese Dataset

Table 5 summarizes the results. From this table, we make the following observations. First, pretrained models including DialoGPT and BERT-GPT achieve lower perplexity than Transformer. This further demonstrates the effectiveness of transfer learning. Second, DialoGPT-MMI achieves better scores on machine translation metrics, which is consistent with the results on the CovidDialog-English dataset. Third, BERT-GPT achieves much better Dist scores than other methods. We manually checked the generated responses by BERT-GPT. Indeed, they are more diverse than others. Fourth, maximum mutual

---

9. https://github.com/huggingface/transformers

	Transformer	DialoGPT, no MMI	DialoGPT, with MMI	BERT-GPT
Perplexity	53.3	37.0	42.1	**10.2**
NIST-4	0.39	0.41	**0.45**	0.30
BLEU-2	5.7%	5.5%	**6.2**%	4.5%
BLEU-4	4.0%	3.1%	**4.5**%	2.7%
METEOR	**13.5**%	13.4%	13.3%	11.1%
Entropy-4	7.9	**9.4**	**9.4**	8.3
Dist-1	5.5%	2.2%	1.6%	**8.7**%
Dist-2	29.0%	25.3%	29.3%	**40.5**%
Avg Len	37.5	43.3	55.6	18.7

Table 5: Performance on the CovidDialog-Chinese test set.

Patient: 去医院，被检查的工作人员戴一次性手套碰了下手腕，停好车就去洗了手。我这会有被感染新冠病毒吗? (In the hospital, my waist was touched by a nurse wearing gloves. I washed my hands after parking. Did I have the risk of being infected by coronavirus?)

Groundtruth: 不会的，不用担心。(Not likely. Do not worry.)

BERT-GPT: 感染可能性非常小。(The likelihood of being infected is very small)

DialoGPT, MMI: 还需要是就也就感染没有其他 37.在发来，干咳。还有点乏力一般在家吃感觉。没有的，吃感冒了，就好。我这种清瘟胶囊

DialoGPT, no MMI: 我就是疫情况不会不会有没有

Transformer: 看不建议你紧张也没有。注意率很多喝点。

Figure 4: An example of generated Chinese responses.

information (MMI) does not have a clear efficacy in improving the quality of generated responses.

Figure 4 shows an example of generating a doctor's response given the utterance of a patient. The response generated by BERT-GPT matches with the ground-truth, both of which indicate that the patient has low risk of being infected. The responses generated by other methods are not understandable.

5. Related Works

Many works have been devoted to developing medical dialogue systems. Please refer to (Laranjo et al., 2018) for a comprehensive review. Some methods (Lucas et al., 2017; Philip et al., 2017; Tanaka et al., 2017) predefine a sequence of steps or states which are used to guide the conversation. Other methods (Rhee et al., 2014; Ireland et al., 2016; Fitzpatrick

et al., 2017) use predetermined templates to extract information from the conversation history and use rules to generate responses from the filled slots in the templates. These methods rely heavily on knowledge engineering and are difficult to be quickly adapted to a new and time-sensitive task such as COVID-19 dialogue generation.

Data-driven medical dialogue generation based on neural networks has been investigated in several works. Wei et al. (Wei et al., 2018) proposed a task-oriented dialogue system to make medical diagnosis automatically based on reinforcement learning. The system converses with patients to collect additional symptoms beyond their self-reports. Xu et al. (Xu et al., 2019) proposed a knowledge-routed relational dialogue system that incorporates medical knowledge graph into 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. In these works, the neural models are trained from scratch on small-sized medical dialogue datasets, which are prone to overfitting.

6. Conclusions

In this work, we make the first attempt to develop dialogue systems that can provide medical consultations about COVID-19. To achieve this goal, we first collected two datasets – CovidDialog – which contain medical conversations between patients and doctors about COVID-19. Then on these datasets, we train dialogue generation models based on pre-trained Transformer, DialoGPT, and BERT-GPT on large-scale dialogue datasets and other corpus. Experimental results show that these trained models are promising in generating clinically meaningful and linguistically high-quality consultations for COVID-19.

References

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

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*, 2019.

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

George Doddington. 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, 2002.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2): e19, 2017.

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

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

David Ireland, Christina Atay, Jacki Liddle, Dana Bradford, Helen Lee, Olivia Rushin, Thomas Mullins, Dan Angus, Janet Wiles, Simon McBride, et al. Hello harlie: enabling speech monitoring through chat-bot conversations. In *Digital Health Innovation for Consumers, Clinicians, Connectivity and Community-Selected Papers from the 24th Australian National Health Informatics Conference, HIC 2016, Melbourne, Australia, July 2016.*, volume 227, pages 55–60. IOS Press Ebooks, 2016.

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

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 2018.

Alon Lavie and Abhaya Agarwal. 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, 2007.

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, 4:51, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

Pierre Philip, Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne De Sevin, Jérôme Olive, Stéphanie Bioulac, and Alain Sauteraud. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Scientific reports*, 7 (1):1–7, 2017.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. b.

Hyekyun Rhee, James Allen, Jennifer Mammen, and Mary Swift. Mobile phone-based asthma self-management aid for adolescents (masmaa): a feasibility study. *Patient preference and adherence*, 8:63, 2014.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

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

Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PloS one*, 12(8), 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 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, 2018.

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

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. 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, 2019.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1810–1820, 2018.

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