Develop Medical Dialogue Systems for COVID-19

Guangtao Zeng   ZENGGUANGTAO98@GMAIL.COM
UC San Diego

Qingyang Wu   WILWU@UCDAVIS.EDU
UC Davis

Yichen Zhang   YIZ037@ENG.UCSD.EDU
UC San Diego

Zhou Yu   JOYU@UCDAVIS.EDU
UC Davis

Eric Xing   EPXING@CS.CMU.EDU
CMU

Pengtao Xie   PENGTAOXIE2008@GMAIL.COM
UC San Diego

Abstract

The pandemic of COVID-19 has caused 1,345,048 infections and 74,565 deaths as of April 6th in 2020. The pandemic also causes burnout of medical professionals and panic of the public. To alleviate the burnout and panic, medical dialogue systems that are able to provide COVID-related consultations to citizens are in great need. In this paper, we develop a baseline approach to generate medical dialogues about COVID-19. We use the GPT2 language model for conversation generation. It takes a conversation history as input and generates the next response. To alleviate overfitting, we pretrain the GPT2 model on a large Chinese corpora and fine-tune it on the small-sized COVID-19 dialogue dataset. Experiments demonstrate that this approach is 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 code is available at https://github.com/UCSD-AI4H/COVID-Dialogue

1. Introduction

Coronavirus disease 2019 (COVID-19) is currently spreading across the world, causing more than 800,000 infections and 40,000 deaths as of April 1st in 2020. Under the pandemic, medical resources in all countries are extremely scarce, and medical staff is facing tremendous work intensity and pressure. What's worse, fear spreads like a virus. Fearful reactions to COVID-19 can put lives at risk. For example, healthy people with mild cough symptoms go to hospital to ask a doctor for advice and get infected in the hospital, which is not only regrettable, but also increase the burden of medical work.

These circumstances motivate us to build a dialogue system that can answer COVID-19 related questions from patients. Currently, there are several online platforms where patients who are concerned that they may be infected by COVID-19 or other pneumonia can consult doctors and doctors can provide advice here. However, compared with the huge number

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 1: 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.

of patients consulting doctors, the number of doctors on the online platform is far from enough. Besides, doctors cannot answer the questions 24 hours without rest. A dialogue system which can answer COVID-19 related questions from patients automatically and real-time will reduce the workload of doctors greatly. Meanwhile, people's panic will be relieved since their questions can be answered any time and their worries can be alleviated in time. Reducing medical pressure, relieving people's anxiety, and effectively assisting medical work, a dialogue system is on demand and has a promising future in application.

In this paper, we make the first attempt to develop a medical dialogue system for COVID-19. We use the CovidDiag-Chinese dataset (Ju et al., 2020) to train the dialogue generation model. The CovidDiag-Chinese dataset contains 399 consultations about COVID-19 and other related pneumonia, having 8440 utterances (in Chinese). We use a high-capacity language model – GPT2 (Radford et al.) – for response generation: given the conversation history, concatenate all utterances in the history into a single sentence; then feed the concatenated sentence into the GPT2 model to generate the next response. Considering that the CovidDiag-Chinese is small in size, which incurs a high risk of overfitting, we use transfer learning to mitigate this risk: first pretrain the GPT2 model on a large Chinese corpora, then fine-tune the pretrained model on the CovidDiag-Chinese dialogues. Experiments demonstrate that this approach is promising. But to build a fully useful COVID-19 dialogue system, more advanced methods are yet to be developed.

2. Dataset

We use the CovidDialog-Chinese (Ju et al., 2020) dataset for model training. It contains 399 consultations about COVID-19 and other related pneumonia, having 8440 utterances. The average, maximum, and minimum number of utterances in a conversation is 15.5, 182, and 2 respectively. The average, maximum, and minimum number of Chinese characters in an utterance is 26.2, 405, and 4 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 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 disease. Figure 1 shows an exemplar consultation. The data is crawled from haodf.com[1], which is an online platform of healthcare services, including medical consultation, scheduling appointment with doctors, etc.

3. Method

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 h_i is the concatenation of all utterances (from both patient and doctor) before d_i. We train a high-capacity language model – GPT2 (Radford et al.) – using these pairs. The models takes h_i as input and generates d_i.

---

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

Conversation History

医生：对于呼吸道的症状，发热，冠状病毒没有特异性表现，有些没有症状。接着要去做血常规，胸部 ct。如果这些都符合新冠的诊断，那还要做核酸检测才能确诊。如果没有接触史，那可能性不大。

(Doctor: Coronavirus has no specific manifestations such as respiratory symptoms or fever. Some patients show no symptoms. They take blood test and chest CT scan. Even if the results of these tests meet the diagnosis criteria of COVID-19, nucleic acid testing is still required to confirm the diagnosis. If you don't have close contact with COVID-19 patients, then you are unlikely to be infected.)

病人：没有去过疫区。嗓子不舒服已经 1 个多月了。没有发烧。去医院检查说是咽炎。开了黄芩胶囊。吃了 4 天不见好。吃药期间感觉前胸灼热感，烧心的感觉，停药就好了。

(Patient: I haven't been to the epidemic area. My throat has been uncomfortable for more than a month. No fever. I went to the hospital and the doctor said it is pharyngitis. The doctor prescribed Huang Qin Capsules. I took it for 4 days, but I didn't feel better. I felt my chest and heart were burning while taking medicine. The burning disappears when I stop taking the medicine.)

医生：你可以配点雷贝拉唑，蒲地兰口服液治疗咽炎。

(Doctor: You can take some Rabeprazole and Pudilan oral solution to treat pharyngitis.)

病人：孕妇可以用吗？

(Patient: Can pregnant women use it?)

医生：孕妇最好少用药。

(Doctor: Pregnant women should take less medication.)

病人：我都不知道是不是医生检查错了。会不会不是咽炎？开的药吃了更严重，停药好些。

(Patient: I don't even know if the doctor made a wrong diagnosis or not. Maybe it's not pharyngitis? It becomes worse when I take the prescribed medications. I feel better when stopping taking the medications.)

医生：那你可能对药不适应，最好不要服用。

(Doctor: You may not be used to these medications. You'd better stop taking them.)

病人：一吃黄芩胶囊等消炎药就前胸灼热感是怎么回事呢？停药就好了。

(Patient: I want to know why I feel burning in the chest when I take anti-inflammatory drugs such as Huang Qin capsules? I feel better after stopping taking the medications.)

医生：那可能是你对这个药不能适应。

(Doctor: It is probably because you cannot get used to this medication.)

病人：早之前吃过几天阿莫西林胶囊也是这样子灼热，也是停药就好了。但是阿莫西林胶囊我以前吃过不会出现这样的情况。

(Patient: I had been taking Amoxicillin capsules for a few days before and I also felt burning. When I stopped the medication, I felt better. But I didn't have such symptoms when I took Amoxicillin capsules long time ago.)

医生：那不一定是药的关系，可能是病情要这样子。

(Doctor: That is not necessarily caused by medication. Maybe it is because the condition should be like this.)

病人：咋回事？这情况是好还是不好。

(Patient: What is the situation? Will my illness get better?)

Ground-truth Response

医生:如果同时 2 个药都是这样子，那可能是病情要发展，发展这样子的，跟药物不一定有关系。

(Doctor: If this happens for both medications, probably it is not caused by medications, but rather the medical condition itself.)

Generated Response

医生：那有没有症状，暂时没有反应？平时没有症状? 后面要查的。复查的。复查的。ct 胸部 ct 胸部 ct 也要查过来。然后再定期随访。

(Doctor: Are there any symptoms? No symptoms temporarily? Do you have such symptoms before? You should have a follow-up appointment. Follow-up appointment. Follow-up appointment. CT chest CT chest CT should be taken as well. Have regular follow-ups.)

Figure 2: An example of generated response of doctor by our model.

The size of the CovidDialog-Chinese dataset is small. Directly training the GPT2 model on this small dataset would result in poor generalization on unseen data. To solve this problem, we utilize transfer learning, which pretrains the GPT2 model on a large Chinese corpora, then fine-tunes the pretrained model on the CovidDialog-Chinese dataset. For the encoder in GPT2, we set it to Chinese BERT (Cui et al., 2019), which is a large-scale pretrained BERT language model. For the decoder in GPT2, we let it have the same structure as BERT (Devlin et al., 2018) but apply lower-triangular mask for autoregressive text generation. We initialize the decoder with BERT's weights. Then we pretrain the decoder with an maximum likelihood estimation (MLE) objective on a large-scale multi-domain Chinese corpus. The resulting model consists of a bidirectional Transformer (Vaswani et al., 2017) 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[2], including the following datasets: Chinese Wikipedia, 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. Given the pretrained model, we fine-tune it on the $\{(s_i, t_i)\}$ pairs in the CovidDialog-Chinese dataset with an MLE objective.

4. Experiments

We split the dialogues into training set, validation set, and test set, with a ratio of 0.8, 0.1, and 0.1. We stop the fine-tuning on the training set when the perplexity on the validation set starts to increase. The max length of the source sequence and target sequence is set to 400. The encoder and decoder structures are similar to those in BERT, which is 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.

With pretraining, the perplexity on the test set is 12.1. In contrast, without pretraining, the perplexity on the test set is 2241964.1. This comparison demonstrates the effectiveness of transfer learning. Figure 2 shows an example of generated response of doctor by our model. The generated response is asking for the symptoms of the patient, asking the patient to have follow-up appointments, and suggesting the patient to take CT scan.

5. Conclusions

In response to the burnout of medical professionals and panic of the public, it is necessary to develop medical dialogue systems that are able to provide consultations to people for COVID-19 related topics. We develop a COVID-19 dialogue system as a baseline for the research community to benchmark with. We train a GPT2 language model which takes conversation history as input and generates the next response. To avoid overfitting, we pretrain the GPT2 model on a large Chinese corpora, then fine-tune the pretrained model on the small-sized COVID-19 dialogue dataset. Experiments demonstrate that this approach is promising in generating meaningful dialogues about COVID-19. But to have a fully useful

---

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

and deployable COVID-19 dialogue system, more advanced dialogue generation methods are needed.

References

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.

Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, and Pengtao Xie. Coviddialog: Medical dialogue datasets about covid-19. *https://github.com/UCSD-AI4H/COVID-Dialogue*, 2020.

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

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.