MedDialog: A Large-scale Medical Dialogue Dataset

Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, Pengtao Xie
UC San Diego
PENGTAOXIE2008@GMAIL.COM

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 a large-scale medical dialogue dataset – MedDialog – that contains 1.1 million conversations between patients and doctors and 4 million utterances. To our best knowledge, MedDialog is the largest medical dialogue dataset to date. The dataset is 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 cost. 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 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 to physicians. In additional to practicing face-to-face medicine which already makes physicians highly occupied, physicians need to provide remote consultations in telemedicine, 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 interests in developing artificial intelligence (AI) methods to assist in telemedicine. In particular, medical dialogue systems are being developed to server as

---

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

"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 accordingly.

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

To address the limitations of existing datasets, we build a large-scale medical dialogue dataset that contains 1.1 million patient-doctor consultations and 4 million utterances. It covers 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, it is the largest medical dialogue dataset to date. The data is open to the public.

2. Dataset

The MedDialog dataset contains 1,145,231 consultations between patients and doctors. The total number of utterances is 3,959,333: 2,179,008 from doctors and 1,780,325 from patients. 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. 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. If we combine consecutive utterances from the same person into a single one, there are 3,209,660 utterances: 1,981,844 from doctors and 1,227,816 from patients. The data is crawled from haodf.com[2], which is an online platform of healthcare services, including medical consultation, scheduling appointment with doctors, etc.

The consultations 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.

2.1. Advantages of the dataset

- **Large number of conversations and utterances.** To our best knowledge, MedDialog is the largest medical dialogue dataset. It has about 1.1 million conversations and 4 million utterances.

- **Broad coverage of medical specialities.** consultations are about 29 broad categories of specialties and 172 fine-grained specialties.

---

2. https://www.haodf.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
(Description of 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?)B
医生：考虑睑腺炎或者睑缘炎
(Doctor: Consider blepharitis or blepharitis)
病人：严重吗
(Patient: is it severe?)
医生：目前看，不算严重。用几天药物看看
(Doctor: At present, it is not severe. Try to take the medications for a few days first.)
病人：哦哦
(Patient: OK)
医生：边治疗边交流吧
(Doctor: Let me know how it works.)

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 the patient, (2) dialogue between doctor and patient, and (3) diagnosis and treatment suggestions given by the doctor.

- **Diversity of the patients.** The patients are from 31 provincial-level administrative divisions in China, with different ethics, age, gender, occupation, education, income, etc. Such diversity greatly minimizes population bias in the dataset.

2.2. Limitations of the dataset

- The language is Chinese, which is not easy for non-Chinese-speaking researchers to work on.

- The patients are from China. The dataset may have a bias to the Chinese population.

- The doctors are from China. The medical consultations, diagnosis, and treatment recommendations may be biased to the practice of medicine in China.

3. Conclusions

To facilitate the research and development of medical dialogue systems that can potentially assist in telemedicine, we build a large-scale medical dialogue dataset that contains 1.1 million conversations between patients and doctors and 4 million utterances. The dataset is publicly available and is continuously growing.

References

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

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