CDialog: A Multi-turn Covid-19 Conversation Dataset for Entity-Aware Dialog Generation

Deeksha Varshney[†*], **Aizan Zafar**[†*], **Niranshu Kumar Behra**[†] **Asif Ekbal**[†]
[†]Department of Computer Science and Engineering,
Indian Institute of Technology Patna, India
{1821cs13,aizan_1921cs17,niranshu_1901cs39,asif}@iitp.ac.in

Abstract

The development of conversational agents to interact with patients and deliver clinical advice has attracted the interest of many researchers, particularly in light of the COVID-19 pandemic. The training of an end-to-end neural based dialog system, on the other hand, is hampered by a lack of multi-turn medical dialog corpus. We make the very first attempt to release a high-quality multi-turn Medical Dialog dataset relating to Covid-19 disease named *CDialog*, with over 1K conversations collected from the online medical counselling websites. We annotate each utterance of the conversation with seven different categories of medical entities, including diseases, symptoms, medical tests, medical history, remedies, medications and other aspects as additional labels. Finally, we propose a novel neural medical dialog system based on the CDialog dataset to advance future research on developing automated medical dialog systems. We use pre-trained language models for dialogue generation, incorporating annotated medical entities, to generate a virtual doctor's response that addresses the patient's query. Experimental results show that the proposed dialog models perform comparably better when supplemented with entity information and hence can improve the response quality.

1 Introduction

Currently, telemedicine is absolutely appropriate in reducing the risk of COVID-19 among healthcare providers and patients due to the diversion of medical resources as millions of people around the world have experienced delays in diagnosis and treatment. Conversational agents (Gopalakrishnan et al., 2019; Zhao et al., 2020; Wu et al., 2018; Reddy et al., 2019) have been proved to be effective in carrying on a natural conversation and understanding the meanings of words to respond with a coherent dialog. It has been also effective in providing support to complete several tasks such as booking a ticket (Liao et al., 2020), getting reservations (Wei et al., 2018), etc. In medical domain, (Zeng et al., 2020; Liu et al., 2020; Li et al., 2021; Wei et al., 2018; Xu et al., 2019) have come up with standard techniques to model medical dialogs which reduces face-to-face consultations, resulting in reduced costs and helps the patient get quicker medical treatment. However, medical dialog systems are more difficult to implement than the standard task-oriented dialog systems (TDSs) as there are several other professional phrases / formal medical expressions that are frequently conveyed while communicating (Shi et al., 2020).

A significant effort has recently been undertaken to collect medical dialog data for research on medical dialog systems (Shi et al., 2020; Zeng et al., 2020; Liao et al., 2020; Liu et al., 2020; Yang et al., 2020). They all, however, have some limitations: (i) A comprehensive diagnosis and treatment procedure is lacking. (ii) Labels are not fine-grained enough. Prior research has typically provided a single poorly graded label for the entire utterance, which may mislead model training and/or lead to erroneous assessment. Furthermore, the scale of the medical entities involved is limited. (iii) Dialog length is limited to an average of 2 turns only. From Figure 1, it can be seen that the original CovidDialog corpus (Yang et al., 2020) has a dialog with only one turn and the patient and doctors utterances are also too lengthy having all the information together at one place. We attempt to split this dialog to make it more suitable for dialog settings by separating and pairing the doctors' and patients' utterances at appropriate points. For example, the first sentence of the patient's query (c.f Q) from Figure 1, is chosen as the first utterance (c.f X_1) for the multi-turn dialog as shown on the right. To maintain the dialog flow, we include generic utterances by doctors as the second utterance such as *"Yes sure, please state your concern."* (c.f X_2). We

[*]Equal Conrtibution

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11373 - 11385
December 7-11, 2022

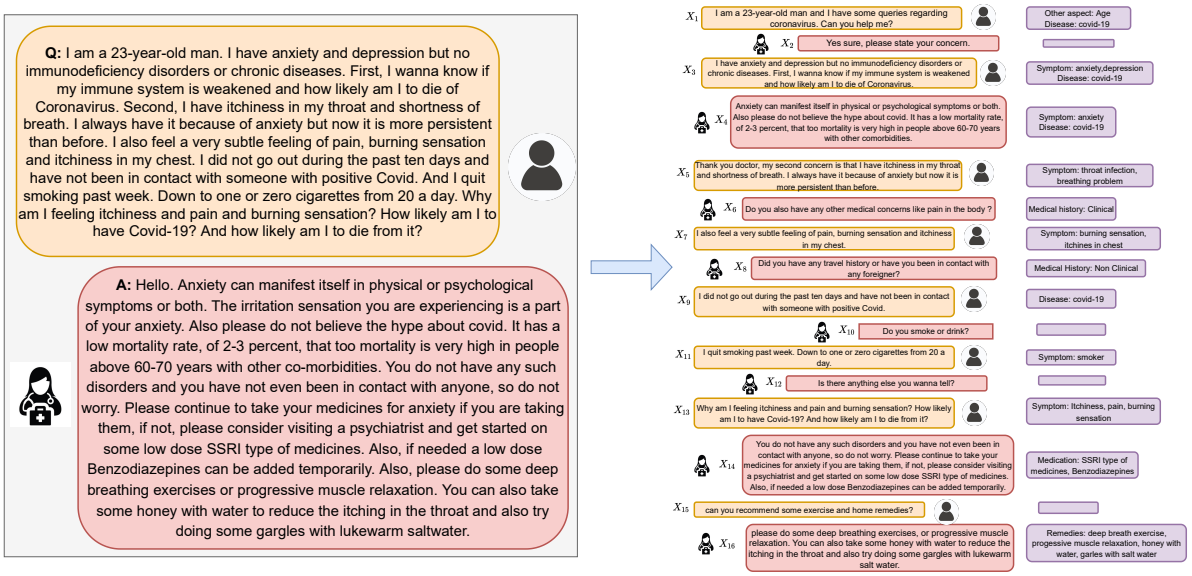

Figure 1: Sample conversation from the CDialog dataset. Sample on left side is from existing CovidDialog dataset (Yang et al., 2020). We have extended this to a multi-turn dialog with eight turns along with entity information. Right side shows such extended samples.

also include appropriate sentences from doctor's response (c.f A), as subsequent utterances (c.f X_4) which comprehends to patient's utterance (c.f X_3) at that point.

Further, we also assign fine-grained medically relevant categories to these utterances. For example, for the third utterance in Figure 1, there are two different kinds of categories: informing symptom status *(Symptoms: anxiety, depression)* and inquiring diseases *(Disease: Covid-19)*.

To address the issue of lack of medically relevant dialog data, we create *CDialog*, a multi-turn Medical Dialog dataset pertaining to Covid-19 disease. As indicated in Table 1, our dataset has the following advantages over the existing conversational datasets. First, our dataset is the largest Covid-19 related dialogue dataset with highest average number of dialogue turns, and thus more suitable for training neural conversation models. Second, CDialog is informational and diversified, with 12 types of diseases and 253 types of entities, which is far more representative of an actual medical consultation scenario. Furthermore, to gain a better grasp of the response generation task, we compare a number of cutting-edge models on CDialog by using popular pre-trained language models like BERT (Devlin et al., 2019) and GPT (Radford et al., 2019). Moreover, we create a medical entity-aware dialog system that makes use of entity-level knowledge. According to the experimental results, combining entity information with dialog history in the generation process improves the response quality.

Our current work makes the following contributions:

1. We build and release *CDialog*, a multi-turn medical dialog dataset related to Covid-19. CDialog has around 1K conversations and with more than 7K utterances annotated with seven types of medical entities, giving it a credible standard for evaluating the medical consultation capabilities of dialog systems.
2. On the CDialog dataset, we present several baselines for response generation and propose techniques for utilizing the relevant medical dialog entities in the medical dialog system.
3. We conduct rigorous experiments, including quantitative and qualitative evaluation, to evaluate a number of cutting-edge pre-trained models for medical dialog generation. Empirical evaluation demonstrates that annotated entities as auxiliary information significantly improves the response quality.

2 Related Work

For dialog generation, sequence-to-sequence models (Vinyals and Le, 2015; Sutskever et al., 2014) are very popular. Shang et al. (2015) proposed a recurrent neural network (RNN) based encoder-decoder architecture for short text conversations. Li et al. (2016); Xing et al. (2017); Zhao et al. (2017); Tao et al. (2018) developed models to help improve the performance of traditional dialog systems us-

Dataset	#Domain	#Diseases	#Dialogs	#Utterances	Avg. Dialog length	#Entities
DX(Dxy) (Xu et al., 2019)	Pediatrics	5	527	2,816	5.26	46
COVID-EN (Yang et al., 2020)	COVID-19	1	603	1,232	2.00	-
MedDialog-EN (Zeng et al., 2020)	Diabetes, elderly problems, pain management, etc [1]	96	260,000	510,000	2.00	-
CDialog (ours)	**COVID-19 & related symptoms**	**12**	**1,012**	**7,982**	**8.00**	**253**

Table 1: Comparison of our corpus to other medical dialog corpora. Statistics include the number of dialogs, disease types, utterances, entity types and average dialog length.

ing extra features such as topic of the conversation, different objective function. Serban et al. (2016, 2017); Xing et al. (2017); Zhang et al. (2019) proposed a number of models for efficiently selecting the conversational context in multi-turn conversation system.

Recent work by (Zhang et al., 2020a) using pre-trained language models has demonstrated captivating performance on generating responses that make sense under the conversation contexts while also carrying out specific content to keep the conversation going by fine-tuning GPT-2 (Radford et al., 2019) in different sizes on social media data. Among all accessible pre-trained language models, BERT is commonly utilised in the medical domain, as several models, such as BioBERT (Lee et al., 2020), Clinical-BERT (Alsentzer et al., 2019), and so on are implemented using the data from a specific domain.

Information extraction (Zhang et al., 2020b), relation prediction (Du et al., 2019; Lin et al., 2019; Xia et al., 2021), and slot filling (Shi et al., 2020) are some of the recent tasks performed on medical data. In medical domain, the use of a reinforcement learning framework in dialog systems (Wei et al., 2018) has encouraged dialog management strategy learning. Further (Xu et al., 2019) increased the rationality of medical conversation decision-making by including external probabilistic symptoms into a reinforcement learning framework. Liao et al. (2020); Xia et al. (2020) used hierarchical reinforcement learning for automatic disease diagnosis. These RL systems, on the other hand, solely learn from tabular data containing the existence of symptoms, ignoring the importance of other key information such as symptom features, tests, and treatment. Furthermore, (Ferguson et al., 2009; Wong et al., 2011; Gatius and Namsrai, 2012; Liu et al., 2016a) constructed early end-to-end medical dialog systems on large scale Chinese medical dialog corpora.

Wei et al. (2018) released the first dataset for

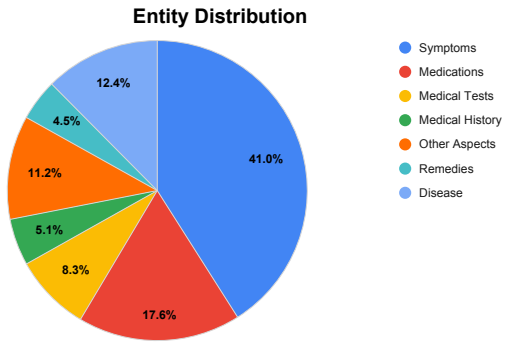

Figure 2: Entity distribution in the CDialog dataset

medical diagnosis, although it only includes structured user goal data rather than natural language dialog. Xu et al. (2019) released a simple dataset named DX with 527 real language dialogs. Recently, (Zeng et al., 2020) released a high-quality unlabelled medical dialogue dataset named MedDialog in Chinese and English covering more than 50 diseases. Although, MedDialog corpora contains the highest number of dialogs, they do not cover dialogs on Covid-19 and have an average dialogue length of only 2. Furthermore, (Shi et al., 2020) released a general-domain medical dialog corpus containing 2K labelled data and 100K unlabeled data, but in the form of individual utterances rather than the entire dialog. MedDG (Liu et al., 2020) compared to the previous corpora involved more diseases, entities, dialogs, and the utterances to alleviate the issue of data scarcity. Li et al. (2021) also released a high quality knowledge-aware medical conversation dataset (KaMed) from ChunyuDoctor, a large online Chinese medical consultation platform. Similar to previous datasets, (Shi et al., 2020; Liu et al., 2020; Li et al., 2021) did not focus on Covid-19 disease.

We create and release a multi-turn dialog dataset named *CDialog* which contains 1K English consultations between patients and doctors along with medical entity annotated utterances. Finally, we propose an entity-aware neural medical conversa-

tion model that generates appropriate responses by utilizing the annotated entities.

3 Resource Creation

In this section, we describe the details of resource creation.

3.1 CDialog Dataset

We extend the CovidDialog dataset (Yang et al., 2020) with the dialogs from the diseases which are the symptoms of Covid-19 and named it as Ext-CovidDialog which now contains approximately 10K dialogs. The motivation for extending the dataset comes from the fact that a conversation about Covid-19 can benefit from the conversations about fever, cough, cold, and other symptoms of Covid-19. We used online platforms of health service consultations such as icliniq.com and heathcaremagic.com to crawl data for fever, cough, etc. We extended the dialog length of 1K dialogs (from 2 to 8) using the dialogs from Ext-CovidDialog (contains $\sim$ 10K dialogs) and also annotated them with several medical entities. The resulting dataset is named as *CDialog* which is finally our proposed dataset for this work.

Our motivation is within the scope of building a conversational system that would engage in online conversation with the users. While developing an automated conversational system, generating longer responses is often a problem for the deep learning models. Hence, we have manually broken this longer utterance into multiple turns. We interacted with the medical experts in our university hospital to ensure that such splitting does not distort the crucial health-related information, rather we added generic statements in order to maintain the flow of the conversation.

3.1.1 Construction Details

As shown in Figure 1, we show a sample of the created and annotated conversation from the CDialog dataset. The average number of utterances in the crawled data (Ext-CovidDialog) is 2.0 per conversation, and the average number of tokens in an utterance is 103. As a result, this conversation is more akin to a question-and-answer session, with the patient describing their problem in detail and the doctor thoroughly answering each question. We aim to convert this question-and-answer (c.f Figure 1 left) setup into a multi-turn human-like conversation format (c.f Figure 1 right). For this, we first view the patient query (c.f Q in Figure 1) as a combination of individual sentences such that each sentence represents some meaningful intent. Then, we choose an appropriate sentence to start the conversation. For each chosen sentence from the patient's query, we search for its significant response in the doctor's answer (c.f A in Figure 1). We have introduced/modified the dialogs in between as needed to ensure that all dialogs are continuously readable and do not go out of context. Because medical data annotation involves annotators with proficient medical knowledge, the annotation cost is high. We employ four annotators with relevant medical expertise. Before beginning the annotation process, we explained the annotation guidelines (c.f Appendix B) using a few examples from the dataset to the annotators. We observe a Fleiss' kappa (Fleiss, 1971) score of 0.85 among the annotators denoting good agreement between them for the task of converting single turn dialogs into multi turn dialogs.

Medical Entity Annotation: We choose the following seven different kinds of entities for annotation after consulting with domain experts: *Diseases* such as allergic conjunctivitis, allergic cough, bacterial conjunctivitis, and so forth; *Symptoms* such as pneumonia, body ache, cough and so on; *Medication* such as anti-allergic tablets, betadine gargle solution, hydroxychloroquine and so on; *Medical Tests*, such as x-rays, etc; *Medical history*, which may be "clinical" or "non-clinical"; *Remedies* such as gargle, exercise, and so on; and *other factors* such as age, nature of pain, duration, and location. As a result, we have 253 entities consisting of 25 different medical tests, 87 different symptoms, 138 different medications, 12 different diseases, 2 different medical histories, 10 unique remedies and 4 other aspects. The distribution of entities in the CDialog dataset is depicted in Figure 2. It shows the proportion of entities in each of the seven categories. Each utterance of the conversation is labeled separately using the seven entity categories, as shown in the right side of Figure 1. The annotation process involved four annotators with relevant medical backgrounds. They begin by discussing the creation of an annotation template. Each participant annotates a small portion of the data and reports the confusing utterance. We summarize our observations and then revise the annotations once more. We observe a Fleiss' kappa (Fleiss, 1971) score of 0.89 between annotators denoting great agreement between them for the entity annotation

task.

More details on the platform and annotators payment can be found in the Appendix B.

3.1.2 Dataset Statistics and Comparision to Existing Dataset

As a result of the annotation process as described in Section 3.1.1, the *CDialog* dataset contains 1012 English consultations about Covid-19 and Covid-related symptoms, such as allergic conjunctivitis, allergic cough, bacterial conjunctivitis, and so forth, which aids in building the multi-turn dialog generation model. The total amount of tokens is 1,085,204 and the total number of utterances is 7,982. The average, maximum, and minimum number of utterances are 8.0, 48, and 2, respectively. The average, maximum, and minimum number of tokens in an utterance are 136, 5313, and 2, respectively. The dataset statistics is shown in Table 5 in the Appendix A.

We compare our proposed *CDialog* dataset to the other publicly available datasets in Table 1 and observe that only three out of the many available datasets as mentioned in Section 2 are in English. When compared to these datasets, we find that the average dialogue length in *CDialog* is eight, indicating that it is more conversational in nature, and our dataset is the largest, focusing solely on Covid-19 with entity annotation for developing entity-aware language models.

4 Methodology

4.1 Task Definition

The goal of a medical dialog system is to provide context-consistent and medically inclined responses based on conversation histories. Formally, given the history of conversations between doctor and patient comprising of K utterances, $X = X_1, X_2, ..., X_i, .., X_K$, where X_i is either a doctor's or a patient's utterance. Each utterance is tagged with an entity set $E = e_1^1, ..., e_s^1, ...e_1^K, ..., e_s^K$, where s is the total number of entities associated with an utterance, X_i. The response generation task is to generate $Y = y_1, y_2, ..., y_M$ with M words given the set of previous K utterances with entity set e_s^K. The architecture is shown in Figure 3.

4.2 Entity-aware Dialog Model

Since generative models are inapplicable to our dataset's annotated entity labels, we present entity-aware models that make use of the supplementary entity knowledge. In this method, the entity set after the dialog history is directly concatenated as new input text and then used to encourage the models for generating the relevant responses.

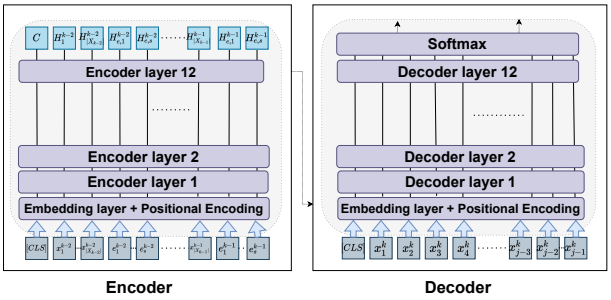

Figure 3: Model architecture

4.2.1 Model Description

To generate contextualized utterance representation for the input sequences, we use the BioBERT_BASE (Lee et al., 2020) pre-trained model (Cased: hidden-1024, heads-16, layer-8, 1M parameters). The context utterances are concatenated with the current user utterance to form a single input utterance. The following is the flattened token sequence for the input utterance combined with the associated entity set:

$$[CLS], x_1^{k-2}, ..., x_{|X_{k-2}|}^{k-2}, e_1^{k-2}, ..., e_s^{k-2}, [SEP], x_1^{k-1}, ..., x_{|X_{k-1}|}^{k-1}, e_1^{k-1}, ..., e_s^{k-1}[SEP] \quad (1)$$

where the $[CLS]$ token is inserted at the start of the sequence to indicate the beginning of the sentence. The $[SEP]$ token denotes the end of a sentence and distinguishes one sequence from the next. Each token is first embedded through three layers (Token, Segment, and Position). The hidden states are obtained by feeding the respective vectors obtained from these three embedding layers into the BioBERT encoder. Furthermore, the hidden vector for each i-th word in the input utterance is denoted as H_i^{k-1}. The bidirectional nature of BioBERT ensures joint conditioning on both the left and right contexts of a token. Then, using a BioBERT decoder, we generate the doctor's response, $Y = y_1, y_2, ..., y_M$, using the words from the gold response, X_k = (x_1^k,x_2^k,.....,$x_{|X_k|}^k$) every time. The decoder predicts each word, y_j, conditioned on x_1^k,...,x_{j-1}^k,

$H_1^1, ..., H_1^{k-1}, ..., H_{|H_{k-1}|}^{k-1}, H_{e,1}^{k-1}, ..., H_{e,s}^{k-1}$

$$Q_j^k = BioBERT\_decoder(H^k) \quad (2)$$

$$P(y_j^k) = \text{softmax}(Q_j^k) \quad (3)$$

4.2.2 Training Loss

The decoder loss is the cross-entropy between the output distribution $P(y_j^k)$ and the reference distribution, T_j, denoted as

$$Loss = -\sum T_j log(P(y_j^k))) \quad (4)$$

5 Experimental Setup

This section describes the baseline models and evaluation metrics. Implementation details can be found in the Appendix C.

5.1 Baselines

We use the following baseline models:

1. **GPT-2 (Radford et al., 2019)**: It is a language model based on Transformer pretrained on Reddit dialogs, in which the input sequence is passed through the model to generate conditional probability on the output sequences.

2. **DialogGPT$_{finetune}$ (Zhang et al., 2020a)**: The model was trained using 147 million Reddit chats and is based on the OpenAI GPT-2 architecture. We begin by concatenating all dialog turns within a dialogue session into a long text that is terminated by the end-of-text token.

3. **BERT (Devlin et al., 2018)**: This model makes use of Transformer attention mechanism which learns contextual relations between the words (or, sub-words) in a text. BERT as an encoder is used to encode the input and BERT as a decoder is used to generate relevant output.

4. **BART (Lewis et al., 2019):** In this model a bidirectional encoder is used for encoding the input sequences and the appropriate response is generated using a left-to-right decoder.

5. **BioBERT (Lee et al., 2020)**: BioBERT is a model similar to BERT aside from that it has been pre-trained on a large biomedical corpus. It outperformed BERT and other state-of-the-art models in several tasks of biomedical text analysis. We use BioBERT both as the encoder and decoder.

The entity set after the dialogue history is directly concatenated as new input text in **BERT-Entity, BART-Entity, and BioBERT-Entity** and then used to stimulate the models to produce the relevant responses.

5.2 Evaluation Metrics

5.2.1 Automatic Evaluation

We evaluate our models on test set, using the following standard metrics. The BLEU (Papineni et al., 2002) score computes the amount of word overlap with the words from the ground truth response. ROUGE-L (Lin, 2004) measures the longest matching sequence of words between the candidate and the reference summary using longest common sub sequence method. Perplexity (PPL) is computed to learn how well the system learns to model the dialog data. We also compute *unigram* F1-score [2] between the predicted sentences and the ground truth sentences. Embedding-based metrics [3] (Liu et al., 2016b) such as Greedy Matching, Vector Extrema and Embedding Average are an alternative to word-matching-based metrics. These metrics assign a vector to each word in order to comprehend the desired sense of the predicted sentence, as described by the word embedding.

5.2.2 Human Evaluation

To evaluate the quality of generated responses from a human point of view, we randomly select 50 dialogs from each model developed using the CDialog dataset and analyze the predicted responses with the assistance of three human evaluators. For each example, we provide the responses (generated by models and ground-truth by humans) to our annotators. Human raters are post-graduates in science and linguistics with annotation experience for text mining tasks. We also had our model outputs validated by a doctor with a postgraduate degree in medicine. The important medical information was found to be retained in the responses. To assess the accuracy of our model predictions, we employ the following metrics: (i) Fluency: It is a measure of sentence's grammatical correctness. (ii) Adequacy: This metric is used to determine whether the generated response is meaningful and relevant to the conversation history. (iii) Entity Relevance (ER): This metric is used to determine whether or not a response contains the correct medical entities.

The scale runs from 1 to 5. The higher the number, the better. For the fluency metric, the ratings refer to incomprehensible, disfluent, non-native, good and flawless English, respectively. Similarly, for the adequacy metric these correspond to none,

[2] https://github.com/facebookresearch/ParlAI/blob/master/parlai/core/metrics.py

[3] https://github.com/Maluuba/nlg-eval

Models	PPL	F1%	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	Embedding Average	Vector Extrema	Greedy Matching
GPT-2	55.45	9.43	0.145	0.044	0.018	0.009	0.108	0.820	0.355	0.630
DialogGPT$_{finetune}$	52.34	9.89	0.148	0.048	0.019	0.009	0.109	0.832	0.359	0.637
BERT	38.48	10.01	0.147	0.045	0.021	0.012	0.124	0.851	0.360	0.640
BART	25.14	11.82	0.161	0.059	0.029	0.017	0.139	0.855	0.368	0.644
BioBERT	**22.67**	15.68	0.204	0.100	0.066	0.051	0.174	0.862	0.401	0.663
BERT-Entity	38.40	10.36	0.150	0.049	0.020	0.010	0.123	0.849	0.353	0.637
BART-Entity	25.92	11.81	0.168	0.061	0.032	0.020	0.138	0.854	0.362	0.643
BioBERT-Entity	22.97	**17.60**	**0.217**	**0.126**	**0.094**	**0.078**	**0.191**	**0.865**	**0.404**	**0.667**

Table 2: Automatic evaluation results for the baseline and suggested model on CDialog dataset. BERT-Entity, BART-Entity, and BioBERT-Entity: BERT, BART and BioBERT based models with the entities concatenated with the input sequences, respectively.

Models	Fluency	Adequacy	Entity Relevance	Kappa
BERT	2.65	1.80	2.59	0.87
BART	3.31	2.18	1.92	0.86
BioBERT	3.60	2.31	2.00	0.85
BERT-Entity	2.71	1.84	1.69	0.81
BART-Entity	3.16	2.33	2.06	0.82
BioBERT-Entity	**3.55**	**2.86**	**2.33**	0.82

Table 3: Human assessment results for the baseline and proposed model on the CDialog datasets. The bolded values represent the best value.

little meaning, much meaning, most meaning and all meaning, respectively. The ratings from the various annotators are averaged and shown in Table 3. We compute the Fleiss' kappa (Fleiss, 1971) score to measure the inter-annotator agreement.

6 Results and Analysis

Table 2 and Table 3 show the automatic and human evaluation results of baselines and the proposed models.

6.1 Automatic Evaluation

Table 2 shows the results using automatic evaluation metrics on the *CDialog* dataset. On most metrics, we see that BioBERT-Entity outperforms Bert-Entity and BART-entity models[4], demonstrating the effectiveness of incorporating medical entities with biomedical embeddings as additional learning signals for improving the task of medical dialog generation. Overall, we observe that entity based models tends to perform better and capture majority of the entities present in the dialog. On CDialog, BioBERT-Entity yields a significant performance improvement by a margin of around 12.25% in F1 score, and 52.94% in BLEU-4 on the test set when compared to the strongest baseline, BioBERT. Apart from word overlapping based metrics, we also notice significant improvement in embedding based metrics denoting efficient decoding using relevant entity information. Comparison to more baseline models can be found in Appendix D.1.

[4]We did a t-test (Lehmann and Romano, 2006) with the null hypothesis between proposed (BioBERT-Entity) and best baseline(BioBERT) (and BART and BERT with and without entity). For both settings the p-value was less than 0.001, indicating that the proposed methods significantly outperform the baselines.

6.2 Human Evaluation Results

Table 3 shows the result of human evaluation. Entity based models outperform the baseline models on fluency, adequacy, and medical entity relevance, demonstrating consistency with automatic evaluation results. All of the kappa values are greater than 0.75, indicating that the annotators are in good agreement.

In Table 4, we present a few example conversations as predicted by the entity based *BioBERT-Entity*, *BART-Entity* and vanilla *BioBERT* and *BART* models on the test set from CDialog corpus. As seen in the first example, BioBERT-Entity correctly decodes the response by utilising the context information and provides counselling to the patient. In the same example, we may note that BioBERT-Entity, as opposed to models without entity information like BioBERT, generates a more adequate response by utilizing the entity *"diabetes"*.

6.3 Error Analysis

Using the generated responses, we conduct a thorough examination of our proposed model and classify the errors it encounters into the following categories:

1. **Generic Response**: We see cases of generic responses by the doctor, such as *"would you like to video or text chat with me?"* in about 4% of our predicted responses by all the models leading to reduced medical entity presence. The reason being that there for many samples in the training

	Conversation
Patient	hello doctor, i have type 1 diabetes, does that mean i can't get the virus just mildly ? am i likely to have severe complications ? **Symptom**: diabetes
Doctor	diabetes + covid-19. because you are a type 1 diabetic it means that you are at greater risk for complications if you contract covid-19. **Symptom**:*diabetes*, **Disease**:*Covid-19*
BioBERT-Entity	yes, it doesn't mean diabetic. but most cases of diabetes are painless. they grow very rapidly. don't change, however they have some risk and the complications are serious.
BioBERT	any other medical causes should be considered quickly. i understand your situation and would try to help.
BART-Entity	most likely you are having a diabetes disease.
BART	have you consulted any physician ?
Patient	hi doctor, my 5 year old son woke up not feeling well. i measured his temperature and mouth reading is 38.3 and armpit is 38.2. **Symptoms**:*fever*, **Disease**:*Covid-19*, **Other aspect**:*age*
Doctor	what more symptoms does he show ?
Patient	he says his head is sore and his neck is sore. am scared it may be covid - 19 ? **Symptoms**:*head sore , neck sore*, **Disease**: *covid - 19*
Doctor	arrange testing. stay home, provide fluids and ibuprofen after checking his temperature. **Symptoms**:*fever*, **Medications:** *ibuprofen*, **Remedies:** *stay home , provide fluids*
Patient	what to do if he does not get well ?
Doctor	if he does not look well to you, please arrange a virtual or in person appointment with your pediatrician.
Patient	are there any precautions i should follow ?
Doctor	avoid contact with others, use good hand washing and consider testing for influenza and coronavirus. would you like to video or text chat with me? **Medical Test:** *influenza test, covid test*, **Remedies:** *use hand sanitizer*
BioBERT-Entity	ok, you can send a picture without the difference.
BioBERT	no problem, glad to be of help . be safe and avoid hand washing or dusty hands.
BART-Entity	please call your doctor as soon as possible. if he develops some signs of covid - 19, he should be examined and tested as soon as possible.
BART	your welcome

Table 4: Case Study: Examples of predictions from our proposed models on the test set. We attempt to predict Doctor's responses based on the sequence of Patient-Doctor-Patient utterances. The corresponding sets of medical entities are bolded.

data where such responses are present to maintain proper information flow and leading to a reasonable conversation.

2. **Non-Fluency**: We observe around 5% cases of non fluency such as "if you were feeling ?", "yes, we can think you give me?" mostly for BERT and BART models. The reason for this is that these models do not take into account medical entities because they are not trained on biomedical data, which leads to inconsistency in responses since they miss important medical terms while predicting responses.

3. **Inadequacy**: The model sometimes fails to predict correct responses for patient utterances having a large set of context utterances. For example in Table 4, we may observe in the second sample that since the conversation history comprises of more than six utterances. The model fails to keep track of the previous information and hence generates an inadequate or a generic response.

4. **Incorrect entity prediction**: In around 10% cases, the model predicts some irrelevant medical entities resulting in contextually incorrect responses. For example, **Patient**: *i am experiencing nasal congestion, sneezing (unaffected by: recent exposure to allergens, exposure to secondhand smoke), sore throat, itchy eyes, ear pressure, nasal drainage, post nasal drip, eye irritation, runny nose, and watery eyes*; **Doctor**: *i think it is itching/congestion. with the itching could be seasonal allergies would consider benadryl 1/2 to 1 tab at bedtime and zyrtec during the day. itching is pretty specific for allergies?*; **Predicted Response:** *hi, also called urti-allergy. have you taken any medicines?* As can be seen, the predicted response missed all of the entities mentioned in the patient's utterance. However, the reason could be that because many entities were mentioned in the utterance, the model was confused and mentioned "urti-allergy" which is also very close to the mentioned symptoms.

More details on the performance of baseline models on these errors can be found in Appendix E.

7 Conclusion

In this paper, we have created an enriched multi-turn medical dialog corpora with manually labeled medical entities. The dataset is typically constructed for the purpose of developing an efficient medical dialog system, with an average dialog length of 8. To facilitate effective conversation understanding and generation, we propose an entity-aware neural conversational model for medical dialog generation. The evaluation results on two benchmark medical datasets show that a BERT-based model with biomedical embeddings and relevant medical entities can successfully generate correct and informative responses.

In the future, we aim to use a medical knowledge graph generated using a UMLS database

to provide domain knowledge into medical dialogues and model the relationship between different medical entities. The codes and dataset used to replicate our findings are available at https://github.com/deekshaVarshney/CDialog; https://www.iitp.ac.in/ãi-nlp-ml/resources.htmlCDialog.

8 Ethical Declaration

All of the datasets used in this study are freely available to the public which are collected from public websites. We followed the policies for using those data and did not violate any copyright issues. The dataset used in this paper is solely for academic research purposes. In a real-world application, medical dialogue systems could be used to counsel patients and collect data for diagnosis. Even if the agent makes a few minor mistakes during the process, doctors will eventually take over in the end. Annotation was done by a dedicated team of people who work full-time. Dataset is medically verified by the health department of our institute. We are not disturbing any health related information and only adding generic statements in order to maintain the flow of the conversation. We further got the data collection and annotation process reviewed by our university review board.

9 Limitations

Detailed cases of limitations by our model is described in Section 6.3. Modelling medical entities is a challenging task in dialog generation. We aim to further investigate this task in the future.

10 Acknowledgement

We would like to thank the reviewers for their constructive comments. Authors gratefully acknowledge the support from the projects "Percuro-A Holistic Solution for Text Mining", sponsored by Wipro Ltd; and "Sevak-An Intelligent Indian Language Chabot", sponsored by Imprint 2, SERB, Government of India.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Du, Mingqiu Wang, Linh Tran, Gang Li, and Izhak Shafran. 2019. Learning to infer entities, properties and their relations from clinical conversations. *arXiv preprint arXiv:1908.11536*.

George Ferguson, James Allen, Lucian Galescu, Jill Quinn, and Mary Swift. 2009. Cardiac: An intelligent conversational assistant for chronic heart failure patient heath monitoring. In *2009 AAAI Fall Symposium Series*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Marta Gatius and Tsetsegkhand Namsrai. 2012. A conversational system to assist the user when accessing web sources in the medical domain. In *The Fifth International Conference on advances in computer-human interactions*, pages 160–164. Citeseer.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Erich L Lehmann and Joseph P Romano. 2006. *Testing statistical hypotheses*. Springer Science & Business Media.

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*.

Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. 2021. Semi-supervised variational reasoning for medical dialogue generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 544–554.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Kangenbei Liao, Qianlong Liu, Zhongyu Wei, Baolin Peng, Qin Chen, Weijian Sun, and Xuanjing Huang. 2020. Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning. *arXiv preprint arXiv:2004.14254*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5033–5042.

Chaochun Liu, Huan Sun, Nan Du, Shulong Tan, Hongliang Fei, Wei Fan, Tao Yang, Hao Wu, Yaliang Li, and Chenwei Zhang. 2016a. Augmented lstm framework to construct medical self-diagnosis android. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 251–260. IEEE.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016b. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *arXiv preprint arXiv:2010.07497*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Revanth Gangi Reddy, Danish Contractor, Dinesh Raghu, and Sachindra Joshi. 2019. Multi-level memory for task oriented dialogs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3744–3754.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7(8).

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586.

Xiaoming Shi, Haifeng Hu, Wanxiang Che, Zhongqian Sun, Ting Liu, and Junzhou Huang. 2020. Understanding medical conversations with scattered keyword attention and weak supervision from responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8838–8845.

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.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4418–4424.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

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.

Wilson Wong, John Thangarajah, and Lin Padgham. 2011. Health conversational system based on contextual matching of community-driven question-answer pairs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2577–2580.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2018. Global-to-local memory pointer networks for task-oriented dialogue. In *International Conference on Learning Representations*.

Yuan Xia, Chunyu Wang, Zhenhui Shi, Jingbo Zhou, Chao Lu, Haifeng Huang, and Hui Xiong. 2021. Medical entity relation verification with large-scale machine reading comprehension. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3765–3774.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1062–1069.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

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*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: A large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020a. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020b. Mie: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

Yufan Zhao, Wei Wu, and Can Xu. 2020. Are pre-trained language models knowledgeable to ground open domain dialogues? *arXiv preprint arXiv:2011.09708*.

A Dataset statistics

Table 5 presents the dataset statistics for the proposed CDialog dataset. The dataset is split into 80:10:10 ratio for preparing the training, test and validation sets.

We conduct several experiment to show the effectiveness of the annotation of entities. They are described as follows. Since, we have broken the longer utterances into short utterances, having extra information in the form of entity annotation is clearly useful. This is already demonstrated by our experiments in Table 2, by building models both with and without entities. The results clearly show improvement in performance for models with entity. Similarly, we conduct an additional experiment with the Ext-CovidDialog dataset and observed that with the entities there is no improvement in the model. Hence, showing that for shorter utterances the entity annotation is more useful. Results on Ext-CovidDialog: **BioBERT** - *F1-score*: 0.222; **BioBERT + Entity** - *F1-score*: 0.211

Statistics	CDialog
#Conversations	1,012
#Utterances	7,982
#Tokens	1,085,204
Average # Utterances	8
Maximum # Utterances	48
Minimum # Utterances	2
Average # Tokens	136
Maximum # Tokens	5,313
Minimum # Tokens	2

Table 5: Dataset statistics

B Annotation Details

Annotation Guideline: Given a query from patient and an answer from doctor, the task is to convert it into a multi-turn dialog by selecting sentences from the query-answer pair such that they

form a sensible multi-turn conversation. Each turn in the conversation contains an utterance by the patient and a response by the doctor. Figure 4, shows an overview of the pipeline for creating the multi-turn dialog data.

1. For each sample query-answer pair, we employ two annotators, one who produces utterances for the patient and one who acts as a doctor and selects relevant sentences as responses. This configuration has several advantages over using a single annotator to serve as both a patient and a doctor such as when two annotators chat about a passage, their dialogue flow is natural and when one annotator responds with a vague response, the other can raise a flag, which we use to identify bad workers.

2. Both the acting patient and doctor sees the original query and answer and also the conversation that happened until now i.e utterances and response from previous turns.

3. While framing a new utterance for starting the conversation, we want annotators to see the longer query and mostly pick the first sentence as their utterance and modify accordingly to begin the conversation. For example, as shown in Figure 1, the annotator picks the " I am a 23-year-old man" sentence from Q and adds "and I have some queries regarding coronavirus. Can you help me?" in order to start the conversation.

4. While responding, we want the annotator to look into the longer answer (c.f. A in Figure 1) and pick the appropriate sentence as the doctor's utterance and we further ask them to sometime respond with only generic sentences such as *Is there anything else you wanna tell?* (c.f X_{12}), *Yes sure, please state your concern.* (c.f X_2) to generate a natural conversation.

5. For medical entity annotation, seven empty columns are provided to choose the relevant medical term for the different categories as defined in Section 3.1.1. For example in Figure 1, for utterances X_4, the relevant medical entities to be annotated are *Symptom*: Anxiety; *Disease*: Covid-19. The annotators were also asked to remove any names to anonymize the data.

Annotators details: The annotators are regular employees (paid monthly as per university norms) at the rate of 35k/month. The annotators have been employed in our research group and they have been working on similar projects since the last three years.

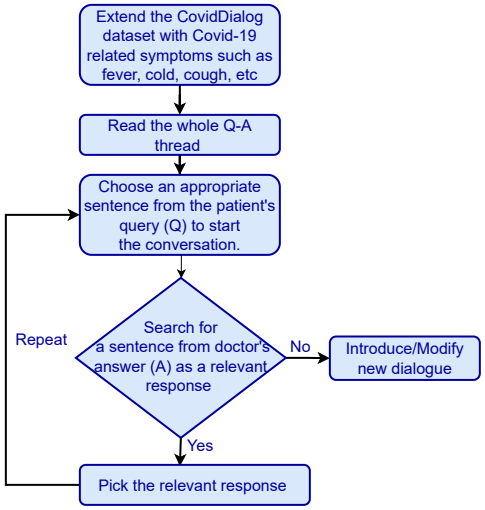

Figure 4: Construction and annotation pipeline of CDialog Dataset

C Implementation Details

All the experiments are implemented using Pytorch framework. BART and BioBERT had hidden size of 1024 while BERT had hidden size of 512. The number of layers is set to 2, 12 and 6 for BERT, BART and BioBERT model respectively. For all the three model BERT, BART and BioBERT number of parameters were 96764928, 457762816 and 360749056 respectively. We use grid search to get the optimal hyperparameter values. We use the AdamW optimizer with learning rate fixed to 0.0005 and the beam size set to 1, while decoding the responses. We choose the best model when the loss on the validation set does not decrease further. We use the GeForce GTX 1080 Ti as the computing infrastructure. Each model is trained up to 30 epochs. After three runs with different random seeds for each method, the variances of the results are at most 1e-4, and they have no impact on the trend.

D Results

D.1 Automatic Evaluation

We also compare our proposed approaches with LSTM based state-of-the-art models such as Seq2Seq (Vinyals and Le, 2015), HRED (Serban

et al., 2015) and VHRED (Serban et al., 2017). Seq2Seq obtains a F1-score of 5.20 and BLEU-4 score of 0.001 on test set of our proposed CDialog dataset. HRED obtains a F1-score of 5.67 and a BLEU-4 score of 0.003 with an embedding average, extrema and greedy score of 0.611, 0.302, 0.542 respectively. VHRED obtains F1-score of 6.11 and a BLEU-4 score of 0.003 with an embedding average, extrema and greedy score of 0.621, 0.304, 0.552 respectively.

E Error Analysis

Performance of baseline models on Inadequacy and Incorrect entity prediction.

1. **Inadequacy**: The prediction by baseline models BART and BioBERT models is shown in Table 4. As can be seen, the baseline models also struggle to maintain track of information, resulting in an insufficient or generic response.

2. **Incorrect entity prediction**: For the example shown in 6.3, 4-th point, the performance of baseline models is as follows: *BERT*: have you been recently? please send for any more information. i have read your query in detail. *BART*: do you have family history? *BioBERT*: not allergy. if you have already taken antibiotics, it may help. did you have any other contact with a doctor? It can be noted that the baseline models perform even worse than the models with entities in terms of retaining relevant clinical information in the predicted response.