On the Generation of Medical Dialogs for COVID-19

Meng Zhou[*], Zechen Li[*], Bowen Tan[†], Guangtao Zeng[*], Wenmian Yang[*], Xuehai He[*], Zeqian Ju[*], Subrato Chakravorty[*], Shu Chen[*], Xingyi Yang[*], Yichen Zhang[*], Qingyang Wu[*], Zhou Yu[◇], Kun Xu[•], Eric Xing[†‡] and Pengtao Xie[*]
UC San Diego[*], CMU[†], Columbia University[◇], Tencent AI Lab[•],
Mohamed bin Zayed University of Artificial Intelligence[‡]
p1xie@eng.ucsd.edu

Abstract

Under the pandemic of COVID-19, people experiencing COVID19-related symptoms have a pressing need to consult doctors. Because of the shortage of medical professionals, many people cannot receive online consultations timely. To address this problem, we aim to develop a medical dialog system that can provide COVID19-related consultations. We collected two dialog datasets – CovidDialog – (in English and Chinese respectively) containing conversations between doctors and patients about COVID-19. While the largest of their kind, these two datasets are still relatively small compared with general-domain dialog datasets. Training complex dialog generation models on small datasets bears high risk of overfitting. To alleviate overfitting, we develop a multi-task learning approach, which regularizes the data-deficient dialog generation task with a masked token prediction task. Experiments on the CovidDialog datasets demonstrate the effectiveness of our approach. We perform both human evaluation and automatic evaluation of dialogs generated by our method. Results show that the generated responses are promising in being doctor-like, relevant to conversation history, clinically informative and correct. The code and the data are available at https://github.com/UCSD-AI4H/COVID-Dialogue.

1 Introduction

During the COVID-19 pandemic, people who are experiencing symptoms similar to those of COVID-19 or were exposed to risk factors have a pressing need to consult doctors. However, medical professionals are highly occupied, who do not have enough bandwidth to provide COVID19-related consultations.

To address this issue, we aim to develop a COVID19-targeted dialog system. We build two medical dialog 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).

While the largest of their kind, these two datasets are still relatively small compared with general-domain dialog datasets. Training complex dialog generation models on small datasets bears high risk of overfitting. To alleviate overfitting in COVID-19 dialog generation, we develop a multi-task learning approach where a masked-token prediction (MTP) (Devlin et al., 2018) task is used to regularize the training of dialog generation models. Our method performs the MTP task and the dialog generation task simultaneously. The MTP loss serves as a regularization term and is optimized jointly with the dialog generation loss. Due to the presence of the MTP task, the dialog generation model is less likely to be biased to the dialog generation task defined on the small-sized training data. We perform experiments on our collected two COVID-19 dialog datasets, where the results demonstrate the effectiveness of our approach. We perform human evaluation and automatic evaluation of dialogs generated by our approach. The results show that the generated responses demonstrate high potential to be doctor-like, relevant to patient history, clinically informative and correct.

The major contributions of this paper are:

- We collect two medical dialog datasets about COVID-19: one in English, the other in Chinese.
- We develop a multi-task learning approach, which uses a masked-token prediction task to regularize the dialog generation task to alleviate overfitting.
- We evaluate our method on the collected COVID-19 dialog datasets and the results demonstrate the

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 886–896
August 1–6, 2021.

effectiveness of our method.

2 Related Works

Several works have studied data-driven medical dialog generation. Wei et al. (2018) proposed a task-oriented dialog 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. (2019) proposed a knowledge-routed relational dialog system that incorporates medical knowledge graph into topic transition in dialog management. Xia et al. proposed an automatic diagnosis dialog system based on reinforcement learning. In these works, the neural models are trained from scratch on small-sized medical dialog datasets, which are prone to overfitting.

3 Datasets

We collected two dialog 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.

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. The conversations are from 582 patients and 117 doctors. Each consultation starts with a short description of the medical conditions of a patient, followed by the conversation between the patient and a doctor.

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. The conversations are from 935 patients and 352 doctors. 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.

	English	**Chinese**
#dialogs	603	1,088
#tokens	90,664	406,550
Average #utterances per dialog	2.0	8.7
Max #utterances per dialog	17	116
Min #utterances per dialog	2	2
Average #tokens per utterance	49.8	42.8
Max #tokens per utterance	339	2,001
Min #tokens per utterance	2	1

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

For both datasets, the dialogs are crawled from openly accessible medical websites whose owners make these dialogs visible to the public. The patients' personal information is de-identified by owners of these websites. We further checked the crawled dialogs to ensure they do not contain private information of patients. Besides, we also manually removed borderline sensitive information, such as specific dates and destinations in patients' travel histories. Experts in privacy and security domains helped to check the final version of shared data and ensured there is no breach of patient privacy or confidentiality.

4 Method

Given a dialog 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 conversation history – the concatenation of all utterances (from both patient and doctor) before t_i. A dialog generation model takes s as input and generates t. The model consists of an encoder which encodes s and a decoder which takes the encoding of s as input and generates t. The size of the CovidDialog datasets is small. Training neural dialog models on these small datasets has high risk of overfitting.

To solve this problem, we develop a multi-task

Split	# dialogs	# utterances	# pairs
Train	482	981	490
Validation	60	126	63
Test	61	122	61

Table 2: English dataset split statistics

	C	R	I	D
Transformer	2.24	2.57	2.53	2.29
GPT-2	2.58	2.91	2.65	3.09
BART	2.61	3.01	2.74	3.42
BART+TAPT	2.65	3.04	2.68	3.38
Ours	**2.83**	**3.16**	**2.88**	**3.47**
Groundtruth	3.26	3.63	3.51	3.55

Table 3: Human evaluation on the CovidDialog-English test set. C, R, I, and D represent correctness, relevance, informativeness, and doctor-likeness respectively. For GPT-2, the "large" version is used.

learning approach, which uses a masked-token prediction (Devlin et al., 2018) task to regularize the dialog generation task. Given the conversation histories in the training set, we encode them using an encoder. Then on top of the encodings, two tasks are defined. One is the dialog generation task, which takes the encoding of a conversation history as input and predicts its corresponding response. The prediction is conducted using a dialog decoder. The other task is masked-token prediction (MTP). In MTP, some percentage of the input tokens are masked at random. The text with masked tokens is fed into the text encoder which learns a latent representation for each token including the masked ones. The task is to predict these masked tokens by feeding the final hidden vectors (produced by the encoder) of the masked tokens into an output softmax operation over the vocabulary. The loss of the MTP task serves as a data-dependent regularizer of the encoder to prevent the encoder from overfitting to the data-deficient dialog generation task. Formally, the method solves the following optimization problem:

$$\mathcal{L}^{(g)}(H, R; W^{(e)}, W^{(g)}) + \lambda \mathcal{L}^{(p)}(H; W^{(e)}, W^{(p)})$$

where H represents the conversation histories and R represents their corresponding responses. $W^{(e)}$, $W^{(g)}$, and $W^{(p)}$ denote the encoder, decoder in the dialog generation task, and prediction head in the MTP task respectively. $\mathcal{L}^{(g)}$ denotes the generation loss and $\mathcal{L}^{(p)}$ denotes the MTP loss. λ is a tradeoff parameter.

5 Experiments

We compare with the following baselines: Transformer (Vaswani et al., 2017), GPT-2 (Radford et al., b), unregularized BART (Liu et al., 2019), unregularized BERT-GPT (Wu et al., 2019), and task adaptive pretraining (TAPT) (Gururangan et al., 2020).

5.1 Experiments on the English Dataset

5.1.1 Experimental Settings

For the English dataset, we split it into a training, a validation, and a test set based on dialogs, with a ratio of 8:1:1. Table 2 shows the statistics of the data split. The hyperparameters were tuned on the validation dataset. Our method and TAPT are both applied to the BART encoder, where the probability for masking tokens is 0.15. If a token t is chosen to be masked, 80% of the time, we replace t with a special token `[MASK]`; 10% of the time, we replace t with a random word; and for the rest 10% of the time, we keep t unchanged. For the regularization parameter λ, we set it to 0.1.

We perform human evaluation of the generated responses. Five medical students are asked to give ratings (from 1 to 5, higher is better) to generated responses in four aspects: 1) Correctness: whether the response is clinically correct; 2) Relevance: how relevant the response is to the conversation history; 3) Informativeness: how much medical information and suggestions are given in the response; and 4) Doctor-like: how the response sounds like a real doctor. The responses are de-identified: annotators do not know which method a response is generated by. The groundtruth response from the doctor is also given ratings (in an anonymous way). Human evaluation was conducted on the test examples in the CovidDialog-English dataset. The ratings from different annotators are averaged.

We also 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 Dist-n (Li et al., 2015) (where $n = 1$ and 2).

5.1.2 Results on the English Dataset

Table 3 shows the human evaluation results. From this table, we make the following observations. **First**, our method outperforms the unregularized BART on all metrics. This demonstrates the effectiveness of our method in alleviating overfitting

	Transformer	GPT-2			BART	BART+TAPT	Ours
		Small	Medium	Large			
Perplexity	263.1	28.3	17.5	18.9	15.3	15.0	**14.9**
NIST-4	0.71	1.90	2.01	**2.29**	1.88	1.89	2.04
BLEU-2	7.3%	9.6%	9.4%	**11.5%**	8.9%	7.7%	8.3%
BLEU-4	5.2%	6.1%	6.0%	**7.6%**	6.0%	3.4%	5.0%
METEOR	5.6%	9.0%	9.5%	**11.0%**	10.3%	9.2%	9.8%
Entropy-4	5.0	6.0	**6.6**	**6.6**	6.5	6.3	**6.6**
Dist-1	3.7%	9.5%	16.6%	13.9%	**16.8%**	11.5%	15.7%
Dist-2	6.4%	22.9%	**36.7%**	31.0%	35.7%	25.3%	33.1%
Avg. Len	40.0	51.3	50.1	54.4	45.4	59.8	50.8

Table 4: Performance on the CovidDialog-English test set.

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?
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.
DialoGPT-large: in brief: 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. regards, dr jnikolla, pulmonologist
BART: 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?
BART+TAPT: if you have been in contact with someone who has been diagnosed with covid-19 then you should definitely get tested. it will make your immune system stronger. hope i have answered your query. let me know if i can assist you further. take care regards, dr. jnikolla
Ours: you do not need to be tested for covid-19 at this time. if you develop fever, cough, shortness of breath and/or difficulty breathing, then you need to consult your pcp and get tested.

Table 5: Generated responses on a test example in the CovidDialog-English dataset.

and improving generalization performance. Our method encourages the encoder to solve an additional MTP task, which reduces the risk of overfitting to the data-deficient response generation task on the small-sized training dialogs. **Second**, our method performs better than BART+TAPT. Our method and TAPT both leverage conversation histories for masked-token prediction. The difference is: TAPT uses these histories to pretrain the encoder while our method uses these histories to regularize the encoder during finetuning. In our method, the encoder is learned to perform the response generation task and MTP task simultaneously. Thus the encoder is not completely biased to the generation task. In TAPT, the encoder is first learned by performing the MTP task, then finetuned by performing the generation task. There is a risk that after finetuning, the encoder is largely biased to the generation task on the small-sized training data, which leads to overfitting. **Third**, our method achieves a doctor-like score that is close to the groundtruth. This indicates that the responses generated by our method have high language quality. The relevance rating of our method is higher than 3, which indicates a good level of relevance between the generated responses and conversation histories. The informativeness rating of our method is better than baselines, but still has a large gap with that of the groundtruth. Additional efforts are needed to improve informativeness, such as incorporating medical knowledge.

Table 4 summarizes the automatic evaluation results achieved by different methods. From this table, we make the following observations. **First**, our method achieves lower (better) perplexity (which is a relatively more reliable metric among various automatic metrics) than the baselines, which further demonstrates the effectiveness of our approach. **Second**, on machine translation metrics including NIST-4, BLEU-2, BLEU-4, and METEOR, the GPT2-large model achieves the highest scores. However, as noted in (Liu et al., 2016), machine translation metrics are not very reliable for evaluating dialog systems. **Third**, on diversity metrics including Entropy-4, Dist-1, and Dist-2, the GPT2-Medium model performs better than other methods. The average length of the generated responses by different methods is close to that of the groundtruth, which is around 50.

Table 5 shows an example of generating a doctor's response given the utterance of a patient. As can be seen, the response generated by our method is more relevant, informative, and human-like, compared with those generated by other baselines. It

	Transformer	GPT-2		BERT-GPT	BERT-GPT-TAPT	Ours
		No MMI	MMI			
Perplexity	53.3	22.1	25.7	10.8	9.3	**9.0**
NIST-4	0.39	0.43	**0.46**	0.36	0.30	0.37
BLEU-2	5.7%	6.2%	**7.2%**	4.6%	5.1%	5.4%
BLEU-4	4.0%	4.0%	**5.4%**	2.8%	2.6%	3.9%
METEOR	13.5%	13.9%	**14.3%**	12.2%	11.9%	13.0%
Entropy-4	7.9	9.0	**9.1**	8.5	7.8	7.9
Dist-1	5.5%	5.9%	3.2%	7.9%	**9.1%**	7.1%
Dist-2	29.0%	38.7%	35.7%	39.5%	**39.7%**	36.6%
Avg Len	19.3	35.0	58.7	21.6	13.9	20.6

Table 6: Performance on the CovidDialog-Chinese test set.

Split	#dialogs	#utterances	#pairs
Train	870	7844	3922
Validation	109	734	367
Test	109	916	458

Table 7: Chinese dataset split statistics

	C	R	I	D
Transformer	1.94	2.09	2.03	2.61
GPT-2	1.72	1.87	1.69	1.78
BERT-GPT	2.15	2.70	2.32	3.02
TAPT	2.27	2.68	2.42	3.11
Ours	**2.87**	**2.77**	**2.49**	**3.19**
Groundtruth	3.11	3.47	3.22	3.71

Table 8: Human evaluation on CovidDialog-Chinese test set. C, R, I, and D represent correctness, relevance, informativeness, and doctor-likeness respectively. Our method and TAPT are based on BERT-GPT. In GPT-2, no maximum mutual information (MMI) is used.

gives correct and informative medical advice such as "if you develop fever, cough, shortness of breath and/or difficulty breathing, then you need to consult your pcp and get tested" and has correct grammar and semantics. In contrast, BART gives clinically incorrect responses such as if someone is a smoker, he or she should be tested for COVID-19. So does BART+TAPT, which incorrectly suggests that getting tested will make the immune system stronger. The responses from GPT2-large and Transformer do not contain any useful medical advice.

5.2 Experiments on the Chinese Dataset

Based on dialogs, we split the Chinese dataset into a training set, validation set, and test set, with a ratio of 8:1:1. Table 7 shows the statistics of the data split. The regularization parameter λ was set to 0.8. Human evaluation was conducted by 5 medical students, on 100 randomly-sampled examples from the test set of CovidDialog-Chinese. The ratings from different annotators are averaged.

5.2.1 Results on the Chinese Dataset

Table 8 shows the human evaluation results. Our method and TAPT are based on BERT-GPT. As can be seen, our approach outperforms unregularized BERT-GPT. This further demonstrates the effectiveness of our approach in alleviating overfitting and improving generalization performance. In addition, our method outperforms TAPT. This further demonstrates that it is more beneficial to perform MTP and dialog generation jointly than separately.

Table 6 summarizes the automatic evaluation results. Our method achieves the lowest (best) perplexity among all methods. Our method outperforms unregularized BERT-GPT and BERT-GPT-TAPT on machine translation metrics as well. GPT2-MMI achieves the highest scores on machine translation metrics. BERT-GPT-TAPT performs better than other methods on diversity metrics.

6 Conclusions

In this paper, we make the first attempt to develop dialog generation models about COVID-19. We first collected two datasets – CovidDialogs – which contain medical conversations between patients and doctors about COVID-19. To alleviate the risk of overfitting, we develop a multi-task learning approach, which uses a masked-token prediction task to regularize the dialog generation model. Human evaluation and automatic evaluation results demonstrate the effectiveness of our proposed method in alleviating overfitting and generating clinically meaningful and linguistically high-quality dialogs about COVID-19.

Acknowledgement

This work was supported by gift funds from Tencent AI Lab and Amazon AWS.

Broader Impact

Dialog systems developed using the collected data in this work should be used very cautiously, under the guidance and supervision of physicians and with approval from the Food and Drug Administration. These dialog systems have the potential to provide timely and accessible COVID-19 consultations to the general public, especially to those who are underserved medically. However, clinical consultation is mission-critical. If the dialog systems make clinical errors, they may cause negative health issues to users. Therefore, these dialog systems should be used as assistants to physicians, rather than operating independently without human supervision. The collected dialogs are from public medical forums, which may be largely different from the patient-doctor dialogue in clinics and hospitals. Such a bias should be paid attention to when using this dataset.

References

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

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

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

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. 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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

David Ireland, Christina Atay, Jacki Liddle, Dana Bradford, Helen Lee, Olivia Rushin, Thomas Mullins, Dan Angus, Janet Wiles, Simon McBride, et al. 2016. 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.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231.

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

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

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

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

I. Loshchilov and F. Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

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

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.

Pierre Philip, Jean-Arthur Micoulaud-Franchi, Patricia Sagaspe, Etienne De Sevin, Jérôme Olive, Stéphanie Bioulac, and Alain Sauteraud. 2017. 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.

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

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

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

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

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

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

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

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

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7346–7353.

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

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

Appendix

A Related Works

A.1 Medical Dialog Generation

Many works have been devoted to developing medical dialog 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 dialog generation.

A.2 Self-supervised Learning for NLP

Self-supervised learning (SSL) aims to learn meaningful representations of input data without using human annotations. It creates auxiliary tasks solely using the input data and forces deep networks to learn highly-effective latent features by solving these auxiliary tasks. In NLP, various auxiliary tasks have been proposed for SSL, such as next token prediction in GPT (Radford et al., a), masked token prediction in BERT (Devlin et al., 2018), text denoising in BART (Lewis et al., 2019), and so on. These models have achieved substantial success in learning language representations. The GPT model (Radford et al., a) is a language model (LM) based on Transformer (Vaswani et al., 2017). Unlike Transformer which defines a conditional probability on an output sequence given an input sequence, GPT defines a marginal probability on a single sequence. In GPT, the conditional probability of the next token given the historical sequence is defined using the Transformer decoder. The weight parameters are learned by maximizing the likelihood on the sequence of tokens. 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. To train the encoder, BERT masks some percentage of the input tokens at random, and then predicts those masked tokens by feeding the final hidden vectors (produced by the encoder) corresponding to the masked tokens into an output softmax over vocabulary. BERT-GPT (Wu et al., 2019) is a model used for sequence-to-sequence modeling where a pretrained BERT is used to encode the input text and GPT is used to generate the output text. 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 is learned to reconstruct the original text. ALBERT (Lan et al., 2019) uses parameter-reduction methods to reduce the memory consumption and increase the training speed of BERT. It also introduces a self-supervised loss which models inter-sentence coherence.

B Datasets

Table 9 shows an example of the English dataset.

C Experiments

C.1 Baselines

We compare the following baselines.

- **Transformer**. A conversation history is fed into the Transformer (Vaswani et al., 2017) encoder and the encoding is fed into the Transformer decoder to generate the corresponding response. The weights of the encoder and decoder are initialized randomly.
- **GPT-2**. Given a dialog history s and a ground-truth response $t = x_1, \cdots, x_n$, a GPT-2 model (Radford et al., a) is trained to maximize the following probability: $p(t|s) = p(x_1|s) \prod_{i=2}^{n} p(x_i|s, x_1, \cdots, x_{i-1})$, where conditional probabilities are defined by the Transformer decoder. For experiments on CovidDialog-English, the GPT-2 model is pretrained on English Reddit dialogs (Zhang et al., 2019). For experiments on CovidDialog-Chinese, the GPT-2 model is pretrained on Chinese chatbot corpus[1].
- **Unregularized BART** (Liu et al., 2019). This approach is the same as Transformer, except that the encoder and decoder are initialized using the pretrained BART (Lewis et al., 2019). The encoder and decoder are finetuned on CovidDialog-

[1] https://github.com/codemayq/chinese_chatbot_corpus

Description of patient's medical condition: I have a little fever with no history of foreign travel or contact. What is the chance of Covid-19?
Dialog **Patient**: Hello doctor, I am suffering from coughing, throat infection from last week. At that time fever did not persist and also did not feel 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.

Table 9: An exemplary 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.

	Transformer	BERT-GPT	Ours	TAPT	GPT-2
GPU	TITAN Xp	GeForce RTX 2080	GeForce GTX 1080Ti	GeForce GTX 1080Ti	TITAN Xp
Num. of GPUs	1	1	1	1	1
Runtime	105	230	488	240	27

Table 10: Computing infrastructure and runtime (seconds per epoch) on CovidDialog-Chinese

English. During finetuning, no self-supervised regularization is used.

- **Unregularized BERT-GPT.** This approach is the same as Transformer, except that the encoder is initialized using pretrained BERT and the decoder is initialized using pretrained GPT-2. BERT and GPT-2 are both pretrained on large-scale Chinese corpus (Cui et al., 2019). The encoder and decoder are finetuned on CovidDialog-Chinese. During finetuning, no self-supervised regularization is used.
- **Task adaptive pretraining (TAPT)** (Gururangan et al., 2020). In this approach, given the Transformer encoder pretrained using BART/BERT on large-scale external corpora, it is further pretrained by predicting masked tokens on the input conversation histories in the CovidDialog datasets (without using output responses). Then the encoder is finetuned by predicting the responses from conversation histories. Similar to our method, TAPT also performs masked-token prediction (MTP) on conversation histories. The difference is: TAPT performs the MTP task and the generation task sequentially while our method performs these two tasks jointly.

C.2 Experimental Settings

C.2.1 Experimental Settings on the English Dataset

For GPT-2, we used three variants (Zhang et al., 2019) with different sizes: small, medium, and large, with 117M, 345M, and 762M weight parameters respectively. Maximum mutual information was not used. We used the Adam (Kingma and Ba, 2014) optimizer for the Transformer model and the AdamW (Loshchilov and Hutter, 2017) optimizer for other models. For all methods except TAPT, we used the optimizer with linear learning rate scheduling, setting the initial learning rate as 4e-5 and the batch size as 4. We perform TAPT for 100 epochs, setting the initial learning rate as 1e-4 and the batch size as 256. The objective for dialog generation 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 during decoding.

Among the automatic evaluation metrics, 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 lexical diversity of generated responses. For perplexity, the lower, the

	Transformer	BERT-GPT	Ours	TAPT	GPT-2
Num. of epochs	30	2	2	2	8
Validation loss	3.17	1.90	2.10	2.08	2.90
Validation perplexity	32.74	6.68	8.14	7.98	18.94

Table 11: Validation performance on CovidDialog-Chinese

Transformer	90M
BERT-GPT	203M
GPT-2	81M

Table 12: Number of weight parameters of each model on CovidDialog-Chinese

	GPU	Runtime
Transformer	GeForce GTX 1080 Ti $\times$ 4	72
GPT-2	GeForce GTX 1080 Ti $\times$ 4	252
BART	GeForce GTX 1080 Ti $\times$ 4	180
Ours	Tesla P100-PCIE-16GB $\times$ 1	270
TAPT	Tesla P100-PCIE-16GB $\times$ 1	150

Table 13: Computing infrastructure and runtime (seconds per epoch) on the CovidDialog-English dataset

better. For other metrics, the higher, the better. As noted in (Liu et al., 2016), while automatic evaluation is useful, they are not completely reliable. Among these metrics, perplexity is generally considered to be more reliable than others.

C.2.2 Experimental Settings on the Chinese Dataset

The hyperparameters were tuned on the validation set. We stop the training procedure when the validation loss stops to decrease. Our method and TAPT are both applied to the BERT encoder in BERT-GPT, where the probability of masking tokens is 0.15. The encoder and decoder structures in BERT-GPT are similar to those in BERT, which is a Transformer with 12 layers and the size of the hidden states is 768. Network weights are optimized with stochastic gradient descent with a learning rate of 1e-4. In the finetuning of BERT-GPT, the max length of the source sequence and target sequence was set to 400. During decoding for all methods, beam search with $k = 50$ was used.

For GPT-2, we used the DialoGPT-small (Zhang et al., 2019) 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. For Transformer, we used the HuggingFace implementation[2] and followed their default hyperparameter settings. We evaluated the models using perplexity, NIST-4, BLEU-2, 4, METEOR, Entropy-4, and Dist-1, 2.

C.2.3 Additional Details about Human Evaluation

In human evaluation on CovidDialog-Chinese, we randomly select 100 examples. Each example includes a conversation history, groundtruth response, and responses generated by different methods. When presented to annotators, the groundtruth and responses generated by different methods are de-identified (given a response, annotators do not know which method generated this response) and randomly shuffled for different examples. The ratings from different annotators are averaged. In human evaluation on CovidDialog-English, we perform evaluation on all test examples.

C.3 Additional Analysis on Experimental Results

Additional analysis of results in Table 3 1) Pretrained models including GPT-2 and BART perform better than Transformer. This further demonstrates the effectiveness of pretraining. 2) BART performs better than GPT-2, though GPT-2 achieves better scores on machine translation metrics. This is in accordance with the results in (Liu et al., 2016) that machine translation metrics are not good for evaluating dialogue generation.

Additional analysis of results in Table 4 1) Pretrained models including GPT-2 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. 2) BART achieves lower perplexity than GPT-2

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

	Transformer	GPT-2	TAPT	BART	Ours
Num. of epochs	100	5	5	5	5
Validation loss	8.02	3.06	2.88	2.84	2.87
Validation perplexity	260.30	21.50	17.74	17.28	17.56

Table 14: Validation performance on CovidDialog-English

Transformer	36M
GPT-2	768M
BART	406M

Table 15: Number of weight parameters of each model on CovidDialog-English

models. This is probably because BART is pretrained on a much larger and more diverse corpus than GPT-2, which enables BART to better model the language. 3) GPT2-large performs better than BART on machine translation metrics including NIST, BLEU, and METEOR. This is probably because GPT2-large is pretrained on dialogue data and therefore tends to generate n-grams that are more related to dialogues. 4) On diversity-related metrics including Entropy and Dist, BART is on par with GPT-2 models.

Additional analysis of results in Table 8 1) Pretrained BERT-GPT works better than unpretrained Transformer. Though pretrained, GPT-2 is not as good as Transformer. The possible reason is the training corpora of GPT-2 is daily dialogues, which has a large domain shift from medical dialogues. The performance gap between BERT-GPT and Groundtruth is larger than that between BART and Groundtruth, despite the number of Chinese training dialogues is larger than that of English training dialogues. This indicates that it is more challenging to develop COVID-19 dialogue systems on Chinese. One major reason is the Chinese dialogues are more noisy than the English ones, with a lot of incorrect grammar, abbreviations, semantic ambiguities, etc.

Additional analysis of results in Table 6 1) Pretrained models including GPT-2 and BERT-GPT achieve lower perplexity than Transformer. This further demonstrates the effectiveness of transfer learning. 2) GPT2-MMI achieves better scores than other methods on machine translation metrics, which is consistent with the results on the CovidDialog-English dataset. 3) BERT-GPT-TAPT achieves better Dist scores than other methods. We manually checked the generated responses by BERT-GPT-TAPT. Indeed, they are more diverse than others. 4) Maximum mutual information (MMI) does not have a clear efficacy in improving the quality of generated responses.

D Computing infrastructure, runtime, validation performance, number of weight parameters, implementation details

D.1 On Chinese CovidDialog

The computing infrastructure and runtime (seconds per epoch) on CovidDialog-Chinese is shown in Table 10. The validation performance is shown in Table 11. The number of weight parameters of each model on CovidDialog-Chinese is shown in Table 12.

We use PyTorch to implement all models. The version of Torch is 1.4.0 (or above). The python package "Transformers[3]" is 2.1.1 for GPT-2 and 2.8.0 (or above) for Transformer and BERT-GPT. When testing, we calculate NIST-n (Doddington, 2002), BLEU-n (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) using NLTK[4] with version 3.5, and calculate Entropy-n (Zhang et al., 2018) and Dist-n (Li et al., 2015) based on the scripts in DialoGPT[5]. We use Gradient Accumulation in PyTorch to enlarge the mini-batch size to 32. Gradient Accumulation is a mechanism of PyTorch, which splits a large batch into smaller batches. The computation on smaller batches is executed sequentially. We set the number of gradient accumulation as 4 so that the mini-batch size is $8 * 4 = 32$.

D.2 On English CovidDialog

Table 13 shows the computing infrastructure and runtime (seconds per epoch) on the CovidDialog-English dataset. The number of epochs and validation performance of each model on CovidDialog-English are shown in Table 14. The number of weight parameters of each model on CovidDialog-English is shown in Table 15.

[3]https://github.com/huggingface/transformers
[4]https://www.nltk.org/
[5]https://github.com/microsoft/DialoGPT