摘 要

在社会面新冠感染风险日益提高的今天，因各种原因出现疑似新冠的症状或者是有着较大感染风险的人们会产生一定医疗咨询方面的焦虑，相对地，医疗系统的压力也会在短期内迅速变大。开辟一个线上的医疗咨询平台可以在一定程度上缓解当前的社会压力与民众焦虑。但与此同时，线上咨询的专业性需要得到保障。因此，本报告的目的是使用专业医疗咨询资料建立一个可以提供 COVID-19 相关咨询的医疗对话模型 CoCoGPT(COVID-19 Consultant GPT)。报告中使用并完善了一个中文医疗对话数据集 CovidDialog-Chinese，其中包含了关于 COVID-19 的医患之间的对话记录。报告中利用这一数据集对一个基于 BERT-GPT 的模型进行训练。由于数据集 CovidDialog-Chinese 相对较小，所以模型首先在对话数据集和其他大规模文本上进行预训练，再在 CovidDialog 任务上进行微调。报告中还将一个基于 Transformer 的模型用相同的方法进行训练，并与 CoCoGPT 进行性能对比。报告中使用了多种评价指标对模型性能进行评估。实验结果表明，CoCoGPT 能够生成质量较高的 COVID-19 相关医学回答。

目　　录

1 任务背景

自新冠肺炎病毒出现以来，疫情形式和相关政策都在随着时间的推移不断变化。身处 2022 年底的现在，我们可以观察到疫情的防治政策发生了较大的改变。我们会更加关注自己及身边人的身体状况，并且由于可能出现的医疗资源的紧缺，在身体出现不适时，我们会尝试自行根据具体的症状推断是否感染了新冠肺炎病毒，并得到相应的医疗建议。因此，为了在一定程度上缓解医疗咨询方面的压力，提供线上智能医疗咨询服务，本报告中提出了一种基于 BERT-GPT 的新冠肺炎相关咨询对话生成模型 CoCoGPT(COVID-19 Consultant GPT)，希望通过这一模型减轻医护人员的压力，进一步提高医疗系统的效率并缓解人们面对新冠疫情时产生的焦虑。

为了全面地构建新冠肺炎相关的知识图谱，便于 CoCoGPT 模型的搭建与发展，本报告中使用了一个中文医疗对话数据集 CovidDialog-Chinese，其中记录了大量医生和患者之间关于新冠肺炎和其他传染病的对话数据。本报告在这一对话数据集上训练基于 BERT-GPT 的对话生成模型以搭建 CoCoGPT，并在同一数据集上训练基于单独 Transformer 的模型与前者进行性能上的对比。Transformer 是一种将对话历史作为输入，生成回答的由编码器 (Encoder) 和解码器 (Decoder) 组成的架构。BERT-GPT 也是一种编码器-解码器型的架构，其中预训练的 BERT 被用于编码层对对话历史进行编码，GPT 被用于解码层进行回答的生成。本报告中采用了多种评估方式对模型结果进行评估，并证明了 CoCoGPT 拥有一定的诊断和建议能力，可以生成具有可参考性和可读性的对话，且性能优于单独的 Transformer-based 对话生成模型。

2 相关工作

纵观过往，诸多工作已经被开展来推动医学对话系统领域的进步。Laranjo(2018) 等人曾对这一领域的发展历史进行了一个全面的回顾 [1]。Lucas(2017)、Philip(2017)、Tanaka(2017) 等人提出了许多方法对系统中步骤或状态的有序序列进行预定义来推动对话的生成 [2][3][4]。Rhee(2014)、Ireland(2016)、Fitzpatrick(2017) 等人使用了预定义的模板对对话历史中的信息进行提取并制定一定的规则利用上述模板中收集到的信息生成相应的回答 [5][6][7]。这些方法严重依赖于相关知识的知识工程操作，并且在快速适应诸如新冠肺炎对话生成之类对时效敏感的新任务时有着许多困难与局限性。

因此，基于神经网络的数据驱动医学对话生成模型开始在一些任务中被研究与应用。Wei(2018) 等人提出了一个任务导向的对话系统 [8]，其可以通过强化学习自动给出医疗诊断结果。这一系统在分析病患自己给出的报告之外，还会与病患进行对话来收集额外的症状信息以进行深入的判断。Xu(2019) 等人提出了一种基于知识路由的相关性对话系统 [9]，其将医学知识图谱融入系统对话管理的主题迁移过程中。Xia(2021) 等人开发了一个基于强化学习的对话系统用于自动给出诊断 [10]，并提出了一种基于生成式对抗网络的使用策略梯度

迭代方法的框架来对这一强化学习模型进行优化。但在他们的工作中，神经网络模型是从开始时就在小规模的医学对话数据集上训练的，因此容易导致模型的过拟合。

3 模型介绍

3.1 Transformer

从对话历史生成回答是一种典型的 SequencetoSequence(Seq2Seq) 任务，而 Transformer 正是一种用于 seq2seq2 任务的编码器-解码器型的架构 [11]，其结构如图1所示。和基于循环神经网络的 seq2seq 模型 (LSTM、GRU 等) 不同的是，Transformer 避免了对序列中字符的循环计算，而是使用了一种不仅可以捕捉字符之间的相关性关系，还易于进行高效率的平行计算的自注意力机制。自注意力机制计算了不同字符对之间的相关性系数并利用此系数求出不同词嵌入的加权和来得到注意力表征。Transformer 是由许多构造块堆叠而成的，每块中都含有一个自注意力层和一个前馈神经网络层。层归一化和残差连接被应用在每两个子层之间。当一个序列被输入后，模型就会使用由上述构造块组成的编码器获取序列中每个字符的表征。此后解码器会接受这些表征作为输入，并解码得到输出字符的序列。为了解码第 i 个字符，解码器首先使用自注意力对已解码得到的序列 $y_i, ..., y_i - 1$ 进行编码，然后计算出上述编码结果和相应的原输入序列间的注意力。随后这些注意力表征会被输入到一个前馈神经层中并多次重复以上步骤。最终得到的表征会被输入到一个线性层中以预测下一个字符。Transformer 的权重参数是通过最大化和输入序列相关的对应输出序列的条件概率来进行学习的。

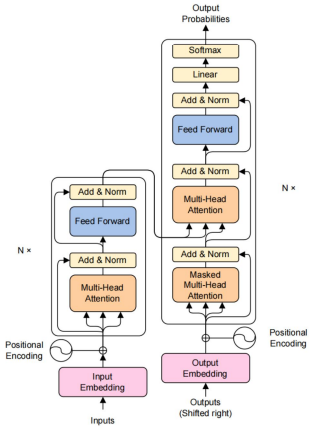

图 1: Transformer 结构

3.2 GPT

GPT 是一种基于 Transformer 的语言模型 (LM)[12]，其结构如图2所示。和根据一个输入序列定义对应输出序列条件概率分布的 Transformer 模型不同，GPT 定义的是单个序列的边缘概率分布。对于一个序列 $x_1, ..., x_n$，LM 定义该序列的概率为：

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

通过这一概率表示基本实现了通过历史序列预测出下一个字符。在 GPT 中，$p(x_i|x_1, ..., x_{i-1})$ 是使用 Transformer 来定义的。其首先使用了多个自注意力层和前馈神经层对序列 $x_1, ..., x_{i-1}$ 进行编码，然后利用 $x_1, ..., x_{i-1}$ 的编码结果对 x_i 进行预测。在此过程中，权重参数是通过不断计算序列中的字符最大似然来进行学习的。GPT-2 是 GPT 的延申，它对原 GPT 模型的调整包括将层归一化移至每个子模块的输入处并在最后的自注意力模块之后添加了一个额外的层归一化。为了对输入字符序列进行更好的表征，模型中使用了字节对编码 (BPE)。

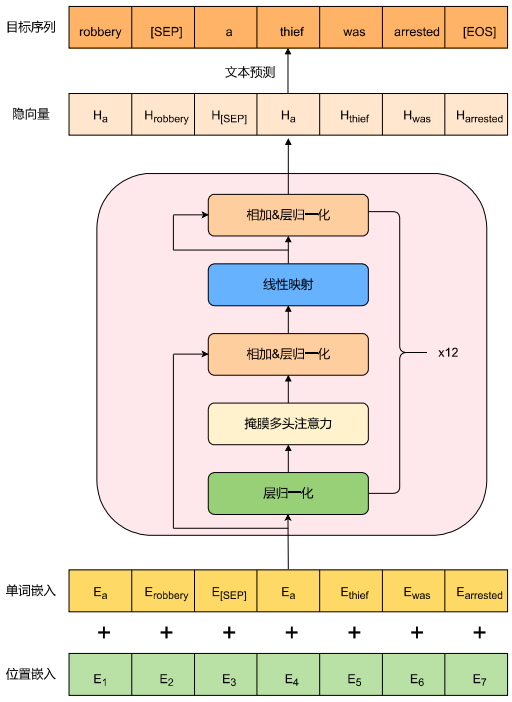

图 2: GPT 结构

3.3 BERT-GPT

BERT-GPT 是一种用于对话生成的模型，使用了预训练的 BERT 对对话历史进行编码并将 GPT 用于生成回答。其中 GPT 的目标是学习一个 Transformer 的解码器用于文本生成，BERT 的目标是学习一个 Transformer 的编码器对文本进行表示。BERT 的模型结构是一个多层的双向 Transformer 编码器。在 GPT 中，每个字符只能对其前面的文本产生单向注意力，而 BERT 中使用的 Transformer 则具有双向自注意力，弥补了 GPT 的缺陷。为了对编码器进行训练，BERT 对一定比例的输入字符进行了随机 mask 操作，然后通过将与这部分字符相关的最终隐状态向量输入至一个对词汇输入进行归一化的 softmax 层中来对这些字符进行预测。由于 Transformer 在对某一字符进行表示时综合考虑了其上下文的内容，因此它拥有比只能对上文进行考虑的 GPT 更强的表征能力。因此在对话生成中，对于输入的对话历史，我们可以使用更强大的预训练 BERT 模型而不是 GPT 对其进行编码以获取字符表征。由 BERT 得到的表征将会被输入 GPT 中以生成回答。

BERT-GPT-Chinese 是一个在中文语料库中进行预训练的 BERT-GPT 模型。在 BERT-GPT-Chinese 中使用的 BERT 编码器被设置为 BERT Chinese，一个用于中文文本的大规模预训练 BERT 模型。在 BERT-GPT-Chinese 中使用的 GPT 解码器有着和 BERT 相同的结构，但额外使用了下三角形式的掩码用于自回归文本生成。这一解码器使用了中文 BERT 参数进行初始化，并接着在一个大规模多领域中文语料库中使用最大似然估计 (MLE) 的方法进行预训练。上述部分最终构成一个由双向 Transformer(BERT-based) 作为编码器，单向 Transformer(GPT-based) 作为解码器，且两者之间用注意力机制相连的模型。用于预训练的中文语料库是从用于自然语言处理的大规模中文语料库 (Large Scale Chinese Corpus for NLP) 收集得到的，其中包括以下数据集：包含 104M 的文章的中文维基数据集，从 63000 个不同数据源得到的两百五万篇新闻文章的中文新闻数据集，包含 493 个不同领域中共计一百五十万个问答对的维基百科 QA 数据集，包含四百一十万条评论和两万八千个主题的社区 QA 数据集。这些数据集的总大小为 15.4GB。本报告在 CovidDialog-Chinese 数据集上对 BERT-GPT-Chinese 模型进行微调，得到用于生成中文新冠肺炎咨询对话的模型 CoCoGPT。

4 实验

4.1 数据集

CovidDialog-Chinese 数据集中包含 1088 段关于新冠肺炎以及其他肺炎的咨询记录，共计 9494 条语句。本报告中在不对数据集中中文语句进行词分割的情况下直接使用汉字构建模型，每个汉字都会被视作一个字符，字符总个数为 406550。数据集中每段咨询记录中语句个数的平均值、最大值和最小值分别是 8.7,116 和 2。每条语句中字符个数的平均值、最大值和最小值分别是 42.8，2001 和 1，如表1所示。每段咨询记录包括以下部分：(1) 病患的

情况描述。(2) 病患和医生之间的对话。(3) 医生给出的诊断和治疗建议。病患的情况描述中包括病患目前患的病、该病的详细病情、病患需要从医生处得到的帮助、病情的持续时间、所用药物和过敏情况、既往病史。这段描述被用作病患一方的第一条语句。上述数据是从一个提供线上医疗服务的平台 haodf.com 上爬取的，其中重复和不完整的对话都会被移除。

表 1: CovidDialog-Chinese 数据集相关数据

	CovidDialog-Chinese
咨询记录	1088
语句	9494
字符	406550
单段咨询记录中语句平均数量	8.7
单段咨询记录中语句最大数量	116
单段咨询记录中语句最小数量	2
单条语句中字符平均数量	42.8
单条语句中字符最大数量	2001
单条语句中字符最小数量	1

4.2 实验方法

本报告中将中文数据集根据对话按 8：1：1 的比例分为训练集、验证集和测试集，如表2所示。

表 2: CovidDialog-Chinese 数据集划分

划分	咨询记录	语句
训练集	870	7844
验证集	109	734
测试集	109	916

模型中使用的超参数都是在验证集上进行微调的。当验证集的损失停止衰减时，训练过程停止。在 BERT-GPT 的微调过程中，源序列和目标序列的最大长度都被设定为 400。其中编码器和解码器的结构和 BERT 中的相似，都是具有 12 层结构的 Transformer，隐藏状态的大小为 768。模型使用的优化器是学习率为 1e-4 的随机梯度下降 (SGD)。对于和 CoCoGPT 进行性能对比的 Transformer 模型，本报告中使用的是 HuggingFace 提供的封装模型并遵循其默认超参数设定。在两个模型的解码过程中，都使用了 k=50 的集束搜索算法。在对模型生成性能进行评估时，本报告中使用困惑度、NIST-4、BLEU-4、Meteor 等指标评价生成结果的相关性，并使用 Entropy-4、Dist-1、Dist-2 等指标评价生成结果的丰富性。BLEU、Meteor 和 NIST 是机器翻译领域常用的评价指标，它们通过匹配 n-gram 模型比较生成结果和真实结果的相似度。困惑度是用于评估生成结果的质量和通顺性的。Entropy 和 Dist 用于评估生成结果中的词汇广度。模型性能越强，则其困惑度就越低，而其他评价指标

就越高。

4.3 实验结果

表3展示了模型的评估结果。

表 3: CovidDialog-Chinese 数据集上的模型性能

	Transformer	CoCoGPT
困惑度	47.9	15.1
NIST-4	0.39	0.36
BLEU-4	4.0 %	2.5 %
Meteor	17.2 %	12.3 %
Entropy-4	7.1	7.9
Dist-1	6.1 %	7.9 %
Dist-2	25.0 %	34.9 %

由表中数据可以得出，CoCoGPT 这一预训练模型的困惑度显著低于 Transformer，这进一步证明了迁移学习和模型集成的有效性。此外，CoCoGPT 的 Dist 指标也显著高于 Transformer。通过对 CoCoGPT 生成的回答进行观察分析，可以发现它们都比 Transformer 生成的回答更加多样化。CoCoGPT 的 BLEU、Meteor、NIST 指标均略低于 Transformer，这可能是因为 CoCoGPT 生成的回答纳入了更多考虑要素使结果相对复杂，在评估时导致其相关性评价指标有少许下降。因此，综合多种评价指标可以得出，基于 BERT-GPT 的 CoCoGPT 模型在新冠肺炎咨询对话生成任务中有着仅基于 Transformer 的一类生成模型无可比拟的优势。

表4展示了病患给出一定的描述语句进行咨询后模型生成的回答。Transfomer 生成的回答语意和逻辑都较模糊。其在“不会”和“新冠肺炎”两个词之间有一个逗号，导致该语句在引导用户进行“是否感染新冠肺炎”的判断时产生歧义。而 CoCoGPT 生成的回答指出了病患不太可能感染新冠病毒并给出了推理原因，回答语句简明且具有可读性、指示性。从客观事实角度而言，这也是一个合理的回答，因为病患在情况描述中提到其新冠肺炎检查结果为阴性。附录 A 中展示了更多新冠肺炎相关咨询中模型生成的回答。

表 4: 对话生成示例

患者:12 月 10 日下午水银测体温 37.4, 身体没有特别不适, 无相关肺炎接触史. 当晚没服药情况下体温降到 36.8。送去医院做新冠肺炎排查。结果出来后排除了。医生只开了清热消炎宁。这几天在家测的体温。早上和晚上都是 37 度左右. 中午和下午 37.1-37.4 之间。没有超过 37.5。无其他不适。
参考: 发热有很多原因，但不是 COVID-19 所致。
Transformer: 医生给您发来一个提醒。不会, 新冠肺炎。
CoCoGPT: 不像新冠肺炎症状，原因是无阳性结果。

5 总结

本报告中提出了一个基于 BERT-GPT 的用于新冠肺炎医疗咨询的对话生成模型 CoCoGPT，其可以对患者输入的病情描述自动生成诊断和医疗建议。报告中使用了包含大量医患之间与新冠肺炎相关的对话资料的中文数据集 CovidDialog-Chinese，并在该数据集上对在大规模对话数据集和其他语料库上进行过预训练的 BERT-GPT 模型进行训练，得到 CoCoGPT 对话生成系统。与此同时，基于 Transformer 的预训练对话生成模型也在 CovidDialog-Chinese 数据集上进行了训练，并与 CoCoGPT 进行性能对比。最终通过多种评价指标对模型进行评估后得出，CoCoGPT 具有较强的医疗咨询对话生成性能，能够生成具有医学意义和较高文本质量的回答。但是，CoCoGPT 仍存在许多不足之处，其模型不能生成详细的细粒度回答，且在生成多个子句的回答时，难以组织好子句的顺序。在未来，需要对模型结构进行更精细的改造，并且增强其对逻辑顺序的捕捉与判断能力。

附录 A 对话生成示例

患者:12 月 11 日中午水银测体温 39.8, 身体关节疼痛, 曾与新冠确诊者接触过。吃过退烧药后体温降至 37.5. 在家抗原检测结果显示为阳性。没有去医院。在家测的体温均超过 37.5，头部非常疼痛。
参考: 患者高烧且抗原检测为阳性，确定为感染新冠肺炎。
Transformer: 产生新冠肺炎症状，很可能是感染者。
CoCoGPT: 患者会确诊新冠肺炎，因为发烧温度较高，检测阳性。

患者:12 月 11 日中午水银测体温 37.4，但喉咙很疼，伴随干咳。次日开始发低烧，且浑身酸疼，四肢无力，水银测体温 38.1，咳嗽更加厉害. 12 月 10 日曾出过门。
参考: 出现了新冠肺炎早期症状，建议立刻进行检测。
Transformer: 许多新冠肺炎症状，很可能是感染者。
CoCoGPT: 患者可能感染新冠肺炎，有部分症状但未检测。

患者:12 月 1 日下午开始发低烧，咳嗽，腹痛，抗原检测为阴性。次日体温达到 40.1，且肺部十分难受，严重咳嗽，吞咽时有剧痛。全身疼痛，无法下床。11 月在外地出差，接触的人较多。
参考: 患者已产生严重新冠症状。
Transformer: 检测结果较好，症状较强。产生，病毒感染。
CoCoGPT:COVID-19 症状明显，建议重复检测。

参考文献

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

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

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics, 2002.

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

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

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

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

[8] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. Task-oriented dialogue system for automatic diagnosis.

In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(Volume 2: Short Papers), pages 201–207, 2018.

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

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

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

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