

从传统自然语言处理到大语言模型电子病历生成技术的探索

牛宇翔 葛珊衫 王力华

(首都医科大学附属北京友谊医院 北京 100050)

〔摘要〕 目的/意义 探讨并展望大语言模型应用于病历文书生成的效果和关键技术，以提升临床工作效率。方法/过程 综述病历文书生成技术发展历程，从传统自然语言处理方法到深度学习方法，再到大语言模型创新应用，并探讨关键技术路线。结果/结论 未来研究方向主要包括基于上下文学习的病历文书生成、基于检索增强生成的病历文书生成及基于混合专家模型的病历文书生成技术。

〔关键词〕 大语言模型；病历生成；自然语言处理；医疗信息化

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2025.03.014

Exploration and Research of Electronic Medical Record Generation Technology from Traditional Natural Language Processing to Large Language Model

NIU Yuxiang, GE Shanshan, WANG Lihua

Capital Medical University, Beijing Friendship Hospital, Beijing 100050, China

〔Abstract〕 **Purpose/Significance** To discuss and look forward to the effect and key technologies of applying large language model to medical record generation, so as to improve the efficiency of clinical work. **Method/Process** The paper reviews the development of medical record generation technology, from traditional natural language processing (NLP) methods to deep learning methods, and then to innovative applications of large language models (LLM), and discusses the key technical routes. **Result/Conclusion** The future research direction mainly includes the generation of medical records based on context learning, the generation of medical records based on retrieval-augmented generation and the generation of medical records based on mixture of experts.

〔Keywords〕 large language models (LLM); medical record generation; natural language processing (NLP); healthcare informatization

1 引言

随着信息化技术的发展，医院医疗质量和效率得到明显提升，但门诊量持续增长使医护人员面临高强度工作压力。以首都医科大学附属北京友谊医院

(以下简称北京友谊医院) 2024 年第 2 季度门诊量为例，3 院区日均门诊量约 1.2 万人次，共 400 余个出诊单元，每个出诊单元（4 小时）接诊 30~40 例患者，每例患者仅分配 6~8 分钟诊疗时间。此外，医生每日需投入大量时间和精力书写病历，每份约 50~100 字，录入需 1.5~2 分钟，患者等待就诊的时间进一步延长，暴露出传统信息化手段在应对日益增长的就诊需求和提升临床工作效率方面的不足。

近年来，大语言模型（large language models,

〔修回日期〕 2024-12-03

〔作者简介〕 牛宇翔，助理工程师，发表论文 1 篇；通信作者：王力华。

LLM) 为病历文书自动化生成提供新方向, 其具备出色的语言理解和生成能力, 可帮助医生高效生成结构化、标准化病历, 提升医疗记录的准确性与质量。但当前病历文书生成面临挑战, 主要源于医患对话不规范: 一是患者受文化水平、知识背景、理解和表达能力限制, 难以清晰、有条理地表达, 复杂对话甚至出现先后表述不一致的情况, 影响模型识别效果, 需提升对于此类对话的处理能力; 二是医患对话中存在大量打断和插话现象, 影响完整表达; 三是部分科室患者病情严重或年龄较大, 需家属陪诊, 对话人数超过两人, 需提升该场景下语音识别能力。本文探讨病历文书生成技术的发展历程、LLM 的创新应用, 分析各种技术的优缺点, 并针对以上问题提出 3 个关键技术方向: 基于上下文学习的病历文书生成、基于检索增强生成的病历文书生成、基于混合专家模型 (mixture of experts, MoE) 的病历文书生成, 最后展望未来发展趋势。

2 传统自然语言处理病历文书生成技术

2.1 基于抽取式自动摘要

传统方法多采用抽取式自动摘要, 从大量医学信息中提取关键句或片段, 生成结构化病历文书。例如, 何剑虎等^[1]使用自动摘要工具从非结构化病历诊断中抽取重要句子。刘伟伟等^[2]通过结构化访谈识别医生在解释患者病史时认为关键的语句。李明等^[3]结合标注训练语料和定制语义模式, 精准提取病历和摘要中的关键信息。Velupillai S 等^[4]聚焦从病历文本中辨识诊断性陈述, 基于标注语料库对其学习并分类, 确保摘要质量。Moen H 等^[5]开发语义相似性建模方法, 从潜在主题中选取代表性句子, 生成语义丰富的摘要。Liang J 等^[6]提出抽取式自动摘要, 专为电子病历设计, 显著提升系统性能, 并对其他临床自然语言处理 (natural language processing, NLP) 任务产生积极影响。

抽取式自动摘要应用于病历文书生成的优势在于: 实现简便, 计算资源需求低; 保持医学文本原意和语境, 语法忠实度高; 快速提取关键信息, 提升病历处理效率。但局限性在于: 摘要流畅性欠

缺, 可读性有待提升; 逻辑连贯性可能不足; 易产生信息冗余, 影响文本质量。

2.2 基于生成式自动摘要

随着深度学习技术的发展, 生成式自动摘要在病历文书生成中得到广泛应用, 尤其是序列到序列 (seq2seq) 方法的成功尝试, 推动了病历生成技术的进步。例如, Zhang Y 等^[7]将 X 射线报告的“发现”和“背景”部分融入生成模型, 为生成评估报告提供关键信息。Mac Avaney S 等^[8]提出本体感知的生成式临床摘要模型, 将报告本体纳入编码过程, 生成更精确的摘要。Shing H C 等^[9]构建融合抽取和生成的管道式框架, 通过注重召回率的抽取器提取关键句子, 再用生成器去除冗余信息, 提高病历文书质量。姜会珍等^[10]研发诊室听译机器人, 结合语音识别和知识图谱生成结构化病历, 并通过卷积神经网络实现个性化病历生成, 极大提高了医生的记录效率。Searle T 等^[11]提出预训练指导信号生成模型, 通过抽取与生成结合、专注临床术语, 生成的摘要更贴近实际需求, 提高了文书的质量和临床实用性。

3 大语言模型技术

3.1 常见模型

近年来, 以预训练语言模型^[1]为代表的大语言模型^[12]技术不断发展。ELMo、BERT、GPT-3 等 LLM 在多个任务上取得显著成果, 获得学术界和工业界广泛关注。尤其是 GPT-3.5、GPT-4 驱动的 ChatGPT 对话系统, 凭借出色的语言理解与生成能力, 在各行业成功应用。国内也探索通过智能语音技术与医院信息系统深度融合, 实现病历语音录入、超声报告助手、语音随访等智能化应用, 积累了丰富的病历文书生成经验。

3.2 性能分析

LLM 具有 Transformer 架构^[13]和不断扩展的数据规模及计算能力, 逐渐普及和广泛应用^[14]。尽管 GPT-4^[15]和 PaLM^[16]等大模型仍为闭源, 仅能通过黑箱接口访问, 但开源模型如 Llama-2^[17]已可

供定制和研究。自 BERT^[18] 和 GPT-2^[19] 问世，学术界和工业界逐步采用“预训练-微调”模式：先基于海量通用数据预训练，随后利用特定领域数据进行微调，适应具体任务。ChatGPT 还采用指令微调技术，响应特定指令并生成符合人类偏好的响应方法^[20]。在临床医疗领域，该技术已被用于开发如 Med-PALM^[21] 医学问答模型和 DoctorGLM^[22] 中文问诊模型，展示了巨大潜力。

面对计算资源限制时，为高效实现领域适应，业界采用前缀微调^[23] 和低秩适应^[24] 等技术。这些技术仅需调整少量参数即可实现高效的领域适应，其中低秩适应技术在医学问答和放射学报告摘要生成等关键领域已证明其出色性能^[25]。根据现有文献实验对比^[6-7,10,12,21]，不同病历文书生成方法在各项性能指标上表现不一。传统 NLP 方法处理速度较快，但准确性和可扩展性较弱；优点是可解释性较高，且不依赖大数据。相比之下，深度学习方法在准确性和可扩展性方面有所提升，但速度稍慢，可解释性较低，且更依赖数据。LLM 在准确性和可扩展性方面表现最佳，能生成高质量病历文书，但处理速度较慢，且需极大的数据量支持；此外，其可解释性仍是较大挑战。这些对比反映了各方法在不同场景下的适用性和局限性。病历文书生成方法性能对比，见表 1。

表 1 病历文书生成方法性能对比

特性	传统方法	深度学习方法	大语言模型方法
准确性	低	中等	高
速度	快	中等	慢
可扩展性	低	中等	高
可解释性	高	低	低
数据依赖	低、中等	高	极高

4 基于大语言模型的病历生成

尽管 LLM 在众多 NLP 任务中展现了卓越性能，并在医疗问答^[21]、在线问诊^[22]、心理咨询^[26] 等临床医学应用中取得了显著成效，但在病历文书生成方面尚显不足。以下 3 个方向有助于充分发挥 LLM 在病历文书生成方面的潜力。

4.1 基于上下文学习

上下文学习^[25] 是一种领域特定任务适应方法，无需大规模调整模型权重即可完成病历文书生成。第 1 步：选择基础能力较强的 LLM 作为基座，如通用文本数据预训练的 Baichuan2-Chat（参数规模：70 亿）或基于医学领域文本微调的 HuatuoGPT2（参数规模：70 亿）。第 2 步：精心标注数个“输入、输出（病历文书）”数据对，作为上下文学习示例数据，帮助模型理解病历文书的格式和内容要求。第 3 步：构建上下文学习提示模板（如“你是一个临床医生，参照例子根据医患对话内容撰写病历文书/n 输入：<示例输入>/n 输出：<示例文书>/n 输入：<医学文本>/n 输出：”），模板与示例数据结合使用，增强模型对临床信息的理解和生成能力。第 4 步：利用提示模板和示例数据进行性能测试，使用自动指标（BERTScore、BLEU、ROUGE）或临床医生评价生成效果。

该应用面临的挑战主要包括两方面。一方面，模型需理解并保持长文本的上下文信息，医患对话中重要信息可能出现在不同部分，这对模型上下文保持能力提出了较高要求。潜在解决方案：分段处理，将长文本分割成多个段落，逐段处理后拼接结果，以提高全局上下文保持能力；层次化编码，长文本分为若干层级，每个层级处理局部上下文，最后合并成整体表示，增强上下文理解。另一方面，病历文书中含大量医学术语和专业表述，模型需准确生成并遵循格式规范。潜在解决方案：领域特定微调，医学领域数据微调，增强模型对医学术语的理解和生成能力；知识融合，集成医学知识到模型中，提供额外背景知识支持；模板化生成，使用模板化文本生成方式，确保生成文书符合医学文档格式规范。

4.2 基于检索增强生成

医院内丰富的病历文书资源是宝贵的知识财富。临床医生撰写病历时，能凭借自身经验调动记忆中的相关知识。有效利用历史病历文书，将极大提升病历自动生成的精准度和可靠性。基于检索增强生成的病历文书生成方法，融合 LLM 的生成能力

和检索式信息抽取的知识获取能力。在生成过程中引入信息检索步骤，增强模型对特定领域知识的理解和应用，生成更准确、丰富的病历文书。第 1 步：通过 Elastic 字词匹配或 Faiss、Milvus 向量匹配搭建信息检索系统，快速检索与输入病历相关的片段。第 2 步：将检索信息与输入病历结合，形成提示指令，输入 LLM，生成更准确的病历文书。为实现高效、准确的信息检索，需优化检索性能、确定合适的检索粒度。选择恰当的检索算法、索引结构和数据库，细致调整检索器参数，可提升检索效率和准确性。根据病历文书生成任务的需求和特点，选择适宜检索粒度，可在检索精确度和效率间取得平衡。采用混合搜索、动态向量更新和上下文感知搜索等先进技术，实施查询分解和提示工程优化，可进一步提升检索性能和效果。综合考虑以上因素，检索增强生成技术能更精准地辅助 LLM，生成准确、丰富、紧密相关的病历文书，极大提高临床文档质量和医疗决策效率，为医疗专业人员提供强有力的工具，也为患者提供更精准、个性化的医疗服务。

该应用难点包括两方面。一方面，模型需预测何时使用检索内容，何时直接基于学习生成，否则检索性能不佳会影响生成质量。潜在解决方案包括：相关性判别，对检索出病历与已有病历进行细粒度相似度比对，不满足阈值时不予采信；置信度预测，构建小参数模型，基于病历内容预测生成置信程度，置信度高时采用检索生成通道。另一方面，检索相关病历参考时，需确保患者隐私和数据安全。潜在解决方案包括：数据匿名，在检索和生成过程中对敏感数据匿名化处理；根据科室和医生职责严格设置访问权限，确保数据安全。

4.3 基于 MoE

基于 MoE^[27] 的病历文书生成技术结合深度学习和混合专家模型。在 LLM 中引入混合专家机制，针对不同科室、不同病种的病历文书，调用不同专家的知识。具体而言，首先构建多个专家网络，每个专家在特定子任务上训练，如诊断说明、治疗建议或患者历史摘要；其次开发门控网络动态决定输入样本路由至哪个专家网络，实现最优任务分配；最

后设计集成框架，有效整合专家输出，生成最终病历文书。

该方法优势在于：MoE 技术通过专家网络专业化细分，确保病历文书生成各环节获得最专业的支持；门控网络智能路由机制，优化任务分配，提升病历生成效率和准确性；集成框架确保不同专家知识有机融合，生成病历文书既全面又专业。其挑战在于构建多个特定科室、子领域表现优异的专家网络。潜在解决方案包括：为子领域准备专门数据集，如诊断数据集、治疗建议数据集等，确保专家网络训练针对性；利用领域模型进行迁移学习，缩短训练时间并提高性能。另一个挑战是如何有效整合多个专家网络的输出，确保生成的病历文书全面、准确且符合临床需求。其潜在解决方案包括：加权融合策略，使用加权平均或投票机制整合专家输出，根据专家网络置信度动态调整权重；内容一致性检查，实施内容一致性检查机制，确保专家网络输出不冲突或重复。

5 基于大语言模型生成门诊病历应用效果

北京友谊医院 2023 年 5 月完成 LLM 生成病历内部开发和内测。上线前针对 LLM 生成病历效果进行阶段性人工评价：对 20 段真实门诊对话进行识别、摘要、生成。将 20 份病历文本和门诊录音交由 5 名医生在完整性、真实性、行文规范、书写位置、主观评价 5 个维度独立评价，满分 10 分。整体主观评价总平均分为 8.19 分，5 个维度均超过 8 分，见表 2。

表 2 大语音模型生成病历效果阶段性人工评价（分）

医生	整体主观评价	少写内容	多写内容	行文规范	位置错误
医生 A	8.25	8.08	8.95	7.65	8.80
医生 B	8.70	8.80	8.95	9.20	9.40
医生 C	8.40	8.30	8.05	8.25	8.55
医生 D	7.20	9.95	9.80	7.05	9.55
医生 E	8.40	8.55	8.70	8.70	9.80
总平均分	8.19	8.74	8.89	8.17	9.22

2023 年 8 月至今，在北京友谊医院内分泌科、消化内科、心内科等科室试用，反馈效果良好，达到

预期。病历质量大幅提升, 门诊看诊效率提升 15%, 单个患者问诊时间节省 80%, 门诊病历录入效率提升 15%, 患者就医满意度和获得感大幅提升。

6 结语

随着人工智能技术的不断进步, 尤其是 LLM 领域的突破, 病历文书生成技术迎来崭新时代。本文介绍了病历文书生成技术的演进, 从抽取式摘要传统方法到 LLM 的应用, 并探讨了关键技术的创新点和应用潜力。LLM 技术, 如上下文学习、检索增强生成以及 MoE 技术, 为病历文书生成带来全新的视角和解决方案。这些技术优化了病历生成流程, 丰富了病历内容。随着技术日益成熟和应用不断深化, LLM 将在病历文书生成领域扮演更加关键的角色, 如提升医疗文档编写的效率与品质, 为临床决策提供丰富、精准信息支持, 以及优化医生工作流程, 为患者带来高质量的医疗服务体验, 实现医疗服务质量整体提升。

作者贡献: 牛宇翔负责研究设计、论文撰写; 葛珊衫负责关键技术研究; 王力华负责提供指导。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 何剑虎, 王德健, 赵志锐, 等. 大语言模型在医疗领域的前沿研究与创新应用 [J]. 医学信息学杂志, 2024, 45 (9): 10-18.
- 刘伟伟, 王立军, 庞娟, 等. 基于自然语言处理的肿瘤专科病历质控系统建设 [J]. 医学信息学杂志, 2024, 45 (2): 77-81, 91.
- 李明, 李昱熙, 戴廉, 等. 医生对医疗人工智能使用意愿性以及影响因素分析 [J]. 医学信息学杂志, 2024, 45 (5): 26-31.
- VELUPILLAI S, KVIST M. Fine-grained certainty level annotations used for coarser-grained e-health scenarios [C]. New Delhi: Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, 2012.
- MOEN H, PELTONEN L M, HEIMONEN J, et al. Comparison of automatic summarisation methods for clinical free

- text notes [J]. Artificial intelligence in medicine, 2016, 67 (2): 25-37.
- LIANG J, TSOU C H, PODDAR A. A novel system for extractive clinical note summarization using EHR data [C]. Minneapolis: The 2nd Clinical Natural Language Processing Workshop, 2019.
- ZHANG Y, DING D Y, QIAN T, et al. Learning to summarize radiology findings [C]. Brussels: The Ninth International Workshop on Health Text Mining and Information Analysis, 2018.
- MACAVANEY S, SOTUDEH S, COHAN A, et al. Ontology-aware clinical abstractive summarization [C]. Paris: The 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019.
- SHING H C, SHIVADE C, POURDAMGHANI N, et al. Towards clinical encounter summarization: learning to compose discharge summaries from prior notes [EB/OL]. [2024-10-28]. <https://arxiv.org/abs/2104.13498>.
- 姜会珍, 胡海洋, 马琰, 等. 基于医患对话的病历自动生成技术研究 [J]. 中国数字医学, 2021, 16 (10): 36-40.
- SEARLE T, IBRAHIM Z, TEO J, et al. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models [J]. Journal of biomedical informatics, 2023, 141 (5): 104358.
- 谢甲琦, 李政. 基于预训练语言模型的公众健康问句分类 [J]. 医学信息学杂志, 2021, 42 (12): 33-36, 43.
- BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [C]. Vancouver: Advances in Neural Information Processing Systems 33, 2020.
- ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report [EB/OL]. [2024-10-28]. <https://arxiv.org/abs/2303.08774>.
- CHOWDHURY A, NARANG S, DEVLIN J, et al. Palm: scaling language modeling with pathways [J]. Journal of machine learning research, 2023, 24 (240): 1-113.
- TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine-tuned chat models [EB/OL]. [2024-10-28]. <https://arxiv.org/abs/2307.09288>.

(下转第 101 页)

践范式，具有重要的推广应用价值。

作者贡献：徐肖攀负责研究设计、论文撰写与修订；卢虹冰负责研究设计；张国鹏、戎军艳、常小红、杜鹏、张文立、刘天帅负责课程群教学实施、教学效果数据收集与整理；刘洋负责论文审核与修订。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 刘洋, 徐肖攀, 戎军艳, 等. 新工科和新医科背景下医学大数据与人工智能课程体系建设研究 [J]. 医疗卫生装备, 2023, 44 (5): 90 – 93.

2 牛秋月, 王晓鹏. 新医科背景下以人工智能为导向的医学院校计算机基础课程体系构建初探 [J]. 中国教育技术装备, 2023 (13): 57 – 60.

3 王沛若, 蔡巧玲, 高静, 等. 智慧医学在临床医学教育中的应用与展望 [J]. 继续医学教育, 2021, 35 (9): 1 – 10.

4 许静, 童睿, 崔巍. 混合式教学模式在临床医学教学中的改革探讨 [J]. 医学教育研究与实践, 2020, 28 (1): 3 – 10.

5 夏丽霞, 王荣, 林征, 等. 大数据视角下智能护理决策支持系统数据平台构建研究 [J]. 中国数字医学, 2022, 17 (3): 55 – 62.

6 朱妍昕, 曹洪欣. 本科医学信息学课程改革 [J]. 解放军

军医院管理杂志, 2021, 28 (9): 898 – 899.

7 汪瑾. 智慧医学在临床医学教育中的应用与展望 [J]. 精品生活, 2022, 22 (24): 84 – 86.

8 吕亚兰, 侯筱蓉. 医学生信息检索课程混合式教学实践与效果评价 [J]. 中国高等医学教育, 2023 (7): 67 – 68.

9 王金, 王秋杰, 赵文龙, 等. 教育数字化转型背景下智慧医学人才培养探索 [J]. 医学信息学杂志, 2024, 45 (10): 1 – 6.

10 张岩, 顾悦. 实践教学赋能生物信息学拔尖人才培养 [J]. 医学信息学杂志, 2024, 45 (11): 95 – 103.

11 何云景, 王平, 吴晓惠. OBE 教育模式实施的核心逻辑 [J]. 上海建桥学院学报, 2017 (3): 24 – 26.

12 刘俊, 潘超. OBE 结合 PDCA 教学模式在高校“新工科”教育中的应用 [J]. 湖北理工学院学报, 2021, 37 (3): 65 – 68.

13 刘晶, 周满英. OBE 视角下医学信息素养课程 BOPPPS 教学模式研究与实践 [J]. 中国中医药图书情报杂志, 2022, 46 (3): 59 – 62.

14 郭建鹏, 张娟. O – PIPAS: 一种以生为本的高校翻转课堂教学模式 [J]. 中国高等教育评论, 2018 (1): 179 – 192.

15 郭建鹏. 翻转课堂教学模式: 变式 – 统一 – 再变式 [J]. 中国大学教学, 2021 (6): 77 – 86.

16 郭建鹏, 翻转课堂教学模式: 变式与统一 [J]. 中国高教研究, 2019 (6): 8 – 14.

(上接第 83 页)

17 DEVLIN J, CHANG M W, LEE K, et al. BERT: pre – training of deep bidirectional transformers for language understanding [C]. Minneapolis: Conference on the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

18 RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2024 – 10 – 28]. <https://openai.com/index/better-language-models/>.

19 WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero – shot learners [C]. Online: International Conference on Learning Representations, 2022.

20 WEI J, BOSMA M, ZHAO V Y, et al. Large language models encode clinical knowledge [J]. Nature, 2023, 620 (7972): 172 – 180.

21 XIONG H, WANG S, ZHU Y, et al. DoctorGLM: fine – tuning your Chinese doctor is not a herculean task [EB/OL]. [2024 – 10 – 28]. <http://arxiv.org/abs/2304.01097>.

22 LI X L, LIANG P. Prefix – tuning: optimizing continuous prompts for generation [C]. Online: The 59th Annual Meeting of the Association for Computational Linguistics and

the 11th International Joint Conference on Natural Language Processing, 2021.

23 HU E J, SHEN Y, WALLIS P, et al. Lora: low – rank adaptation of large language models [C]. Online: International Conference on Learning Representations, 2022.

24 LAMPINEN A, DASGUPTA I, CHAN S, et al. Can language models learn from explanations in context [C]. Abu Dhabi: Findings of the Association for Computational Linguistics: EMNLP, 2022.

25 CHEN Y, XING X, LIN J, et al. Soulchat: improving LLMs’ empathy, listening, and comfort abilities through fine – tuning with multi – turn empathy conversations [C]. Singapore: Findings of the Association for Computational Linguistics: EMNLP, 2023.

26 ZHOU Y, LEI T, LIU H, et al. Mixture – of – experts with expert choice routing [J]. Advances in neural information processing systems, 2022, 35 (2): 7103 – 7114.

27 赵鑫, 李军毅, 周昆, 等. 大语言模型 [EB/OL]. [2024 – 10 – 28]. <https://llmbook-zh.github.io>.