·方法学研究·

面向真实世界的知识挖掘与知识图谱补全研究（四）：真实世界数据标注平台搭建及基于预训练语言模型的自动化抽取方法探索

阎思宇[1#]，谭杰骏[2#]，朱海锋[2]，黄　桥[1]，王诗淳[1,3]，马文昊[1,3]，石涵予[4]，王永博[1]，任相颖[1]，胡文斌[2]，靳英辉[1]

1. 武汉大学中南医院循证与转化医学中心（武汉 430071）
2. 武汉大学计算机学院（武汉 430072）
3. 武汉大学第二临床学院（武汉 430071）
4. 武汉大学弘毅学堂（武汉 430072）

【摘要】目的 探索搭建真实世界数据标注平台，并比较检索增强生成式技术（retrieval augmented generation，RAG）结合大语言模型，及预训练语言模型的预训练–微调方法的真实世界数据提取效果。**方法** 以真实世界电子病历数据中的膀胱癌病理记录为例，搭建真实世界数据标注平台，并基于平台标注数据比较 RAG 结合 GPT-3.5，及基于 BERT、RoBERTa 模型的预训练–微调方法自动化抽取膀胱癌癌症分型、分期的效果。**结果** 全训练集微调的预训练–微调模型抽取效果优于 RAG 结合大模型的方法与小样本微调的预训练–微调模型，RoBERTa 模型效果总体优于 BRET 模型，但这些方法的抽取效果均有待提升。在测试集中，使用全训练集微调的 RoBERTa 模型抽取膀胱癌分型、T 分期、N 分期的 F1 值分别为 71.06%、50.18%，73.65%。**结论** 预训练语言模型在处理临床非结构化数据方面具有应用潜力，但现有方法在信息抽取效果上仍有提升空间。未来工作需进一步优化模型或训练策略，以加速数据赋能。

【关键词】真实世界数据；电子病历；标注平台；预训练语言模型；检索增强生成；大语言模型；病理记录；膀胱癌

【中图分类号】TP 181；R 737.14　　**【文献标识码】**A

Research on real-world knowledge mining and knowledge graph completion (IV): construction of a real-world data annotation platform and exploration of automatic extraction method based on pre-trained language models

YAN Siyu[1#], TAN Jiejun[2#], ZHU Haifeng[2], HUANG Qiao[1], WANG Shichun[1,3], MA Wenhao[1,3], SHI Hanyu[4], WANG Yongbo[1], REN Xiangying[1], HU Wenbin[2], JIN Yinghui[1]

1. Center for Evidence-Based and Translational Medicine, Zhongnan Hospital of Wuhan University, Wuhan 430071, China

2. School of Computer Science, Wuhan University, Wuhan 430072, China

DOI: 10.12173/j.issn.1004-5511.202409004

[#] 为共同第一作者

基金项目：国家自然科学基金面上项目（82174230）；武汉大学中南医院青年交叉学科专项基金（ZNQNTC2023006）

通信作者：胡文斌，博士，教授，博士研究生导师，Email：hwb@whu.edu.cn

靳英辉，博士，副教授，硕士研究生导师，Email：jinyinghui0301@163.com

3. The Second Clinical College of Wuhan University, Wuhan 430071, China
4. HongYi Honor College of Wuhan University, Wuhan 430072, China
#Co-first authors: YAN Siyu and TAN Jiejun
Corresponding authors: HU Wenbin, Email: hwb@whu.edu.cn; JIN Yinghui, Email: jinyinghui0301@163.com

【Abstract】Objective To explore the construction of a real-world data annotation platform, and compare the real-world data extraction performance of retrieval augmented generation (RAG) combined with large language models and pre-training fine-tuning methods for pre-trained language models. Methods Taking the pathological records of bladder cancer in the real world electronic medical record data as an example, a real-world data annotation platform was built. Based on the platform annotation data, the effects of automatic extraction of cancer typing and staging of bladder cancer using RAG combined with GPT-3.5, and the pre-training fine tuning method based on BERT and RoBERTa models were compared. Results The extraction effects of the pre-training and fine-tuning model based on the fine-tuning of the full-training set were better than that of RAG combined with large model method and pre-training and fine-tuning model with the few-shot fine-tuning, and the effects of RoBERTa model were generally better than that of BERT model, but the extraction effects of these methods needs to be improved totally. The F1 scores for extracting bladder cancer typing, T staging, and N staging in the test set, using the RoBERTa model fine-tuned with the entire training set, were 71.06%, 50.18%, and 73.65% respectively. Conclusion Pre-trained language models have the application potential in processing clinical unstructured data, but there is still room for improvement in the information extraction effect of existing methods. Future work requires further optimization of models or training strategies to accelerate data empowerment.

【Keywords】Real-world data; Electronic medical records; Annotation platform; Pre-trained language model; Retrieval augmented generation; Large language model; Pathology records; Bladder cancer

本文为面向真实世界的知识挖掘和与知识图谱补全研究系列文章的第四篇，如前所述[1-3]，电子病历数据作为真实世界数据的典型代表，含有大量非结构化信息，如临床叙述，需要将其中的有效信息抽取出来方能支撑后续的真实世界研究。然而，医学信息的抽取因医学语言的复杂性、可变性及医学领域的低容错率而充满挑战。传统人工抽取由具备专业医学知识的人员进行，成本高昂，且面临效率低、管理难的挑战。通过设计和创建数据标注平台可以一定程度提升人工抽取的效率和质量，便于多人抽取和专人审核。

近年来，基于 Transformer 架构的预训练语言模型在自然语言处理领域取得了突破性进展[4]。预训练语言模型是一种深度学习模型，它在大量未标记的文本数据上进行训练，以学习语言的通用表示。例如，BERT 模型及基于 BERT 优化的 RoBERTa 模型，证明了通过大规模预训练然后微调到特定任务上可以显著提高多种自然语言处理任务的性能[5-6]。随着参数规模的增大，出现了以 GPT-3 为代表的大语言模型（large language models，LLMs）。LLMs 已被应用于电子病历这种复杂语言的信息提取、文本摘要、识别疾病表型等任务[7]，但目前仍面临产生“幻觉”、特定领域知识欠佳、训练成本高昂等挑战[8]。为此，检索增强生成（retrieval augmented generation，RAG）技术被开发用于最大限度地减少幻觉，通过从可靠数据源中识别和整合相关信息来辅助 LLMs 生成更加准确和可靠的答案[9]。

病理结果通常作为肿瘤确诊的金标准，其以长文本形式记录疾病分级、分期、分型等重要信息。

本研究以膀胱癌病理报告为例，搭建真实世界数据标注平台，探索 RAG 结合 GPT-3.5 模型及基于 BERT、RoBERTa 模型的预训练 – 微调方法辅助挖掘病理报告中癌症分型、分期的效果，为这些方法在电子病历信息抽取的实践应用提供参考。

1 真实世界数据标注平台搭建

1.1 数据来源

以武汉大学中南医院 2015—2022 年出院诊断为膀胱癌的患者的病理记录为本研究的数据源。研究已通过武汉大学中南医院伦理委员会审批（批号：科伦 [2022002K]），所有数据均进行了去隐私化处理。

1.2 多学科团队组建

本研究组建了包括临床、计算机、病理、医学信息学、循证医学等领域的多学科专家团队，通过多学科团队利用计算机技术对电子病历进行文本挖掘。

1.3 确定提取内容及标注指南

通过与泌尿外科临床专家、病理学专家就膀胱癌病理报告关注重点、病理报告内容进行解读和培训，最终确定提取病理报告中的癌症分型、TN 分期、分级、组织学亚型、血管淋巴管浸润情况、手术标本是否包含肌层、周围组织浸润情况、基底部浸润情况、淋巴结清扫情况等信息。

此外，考虑到早期部分病理记录中未明确 TN 分期，本研究综合临床和病理专家意见制订了未明确记录 TN 分期时依据报告内容（如肿瘤的浸润深度、淋巴结浸润情况等）人工提取分期的标注指南，并基于预提取结果及反馈修订确定最终的标注指南。同时在正式标注前对标注员及审核员均进行了规范化培训。

1.4 基于标注平台进行文本标注及审核

本研究计算机团队搭建了标注平台，支持多位标注员同时进行标注和审核员审核。本研究基于开放共享的中文医学知识图谱模型 OMAHA Schema 中已有的域和属性进行文本标注 [10]，若未能找到足够契合的属性，将基于提取目的进行补充。

标注平台的数据提取界面如图 1 所示。对每条病理记录进行标注时，先依据句号或句意由标注员分割、填入“原文内容”，再基于 Schema 中的域及属性标注该“原文内容”中需要提取的信息填入“值”，并且使用“组”来标记同一原文内容内多个属性的归属。由于部分早期病理记录未直接报告 TN 分期等变量信息，需依据标注指南推断此类信息，因此本研究也搭建了推理界面，完成疾病分型、分期的提取。

每条数据标注完成后进入待审核区，由医学专家进行内容审核。审核未通过的数据修改后再次进入审核，直至数据通过审核。

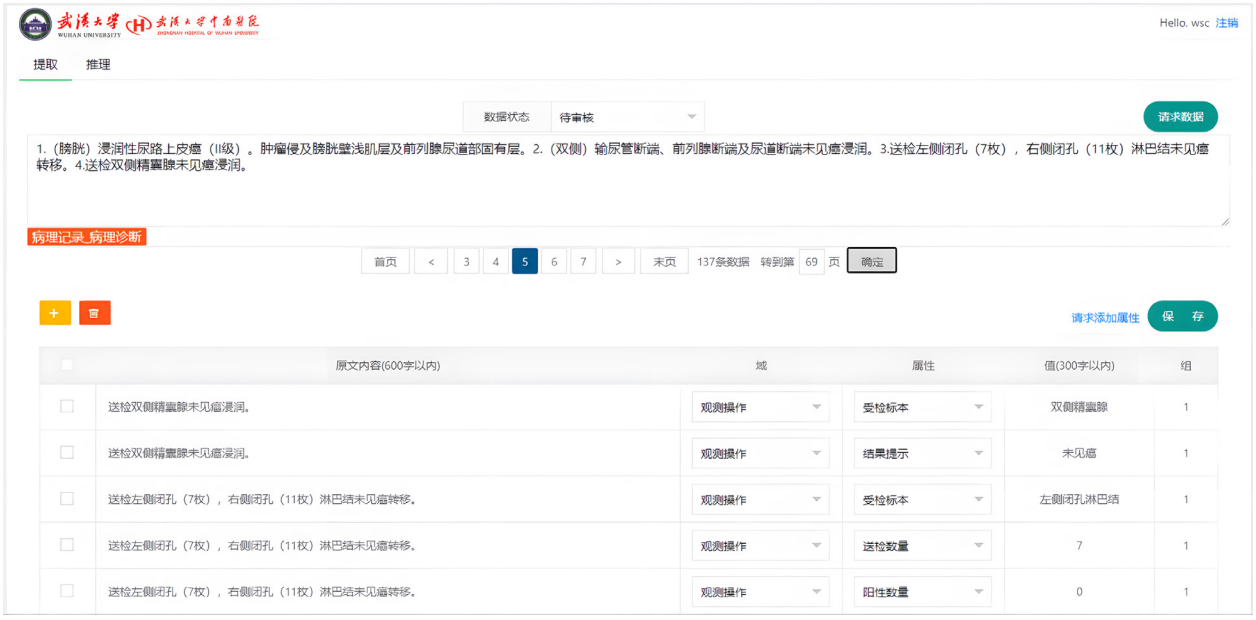

图1 真实世界数据标注平台提取界面示例

Figure 1. Example of extraction interface of the real-world data annotation platform

2 自动化抽取方法及效果评价

2.1 RAG结合GPT-3.5模型方法

利用上一步基于标注平台已经完成标注及审核的膀胱癌病理记录作为语料库，为预训练语言模型提供模型参数以外的医学领域专业知识，从而辅助模型做出更准确的提取。以疾病的分期、分型提取为例，当借助 LLM 辅助提取病历文本中的诊断结果（疾病分期、分型）时，基于 RAG 技术先查询人类专家已标注、审核的病历文档库，返回相关文档作为参考，输入预训练语言模型，辅助其生成该新病历文本中的诊断结果，完成信息抽取。抽取的结构化信息可以作为真实世界研究的数据基础，也可以反馈于临床医生，帮助其快速获取疾病诊断信息，服务于诊疗过程。同时，新的诊疗记录又将补充进入待标注的病历文档库，实现系统闭环。应用流程如图 2 所示。

本研究采用 Pyserini 的中文 BM25 模型作为信息检索模型，选择 Open AI 的公开接口“gpt-3.5-turbo”作为本方法的 LLM。综合考虑效率和准确度等因素，选择检索结果输出前五及前十的相关病历，分别作为 LLM 的参考，探索其效果，模型分别表示为 MedRAG@5 及 MedRAG@10。为了提高准确度，对于膀胱癌分型、T 分期和 N 分期三个提取任务，分别做三次平行的生成。每次生成任务都能参考对应的相关文档。

2.2 基于BERT、RoBERTa预训练语言模型的预训练-微调方法

本研究选择了两个主流的中文预训练语言模型作为预训练 - 微调方法的基础模型，再连接前馈神经网络（feed-forward neural network，FFNN）构成的分类器，形成的网络结构如图 3 所示。该网络结构同时用于膀胱癌分型、T 分期和 N 分期的抽取。模型共享同一个预训练编码器，但使用不同的 FFNN 做分类器。[CLS] 是默认初始词向量，其经过预训练模型编码器的输出可以认为是病历文档的词向量表示。[CLS] 输出经过 FFNN 分类器输出结果。

2.2.1 基础模型

以下两个主流的中文预训练语言模型作为预训练 - 微调方法的基础模型：

① Chinese-BERT-wwm-Finetuned：采用科大讯飞团队的基于预训练模型 BERT[5] 的改进版本和中文训练数据特化版本 [11]。使用 HuggingFace 网站上开源的预训练模型参数。

② Chinese-RoBERTa-wwm-Finetuned：采用科大讯飞团队的基于预训练模型 RoBERTa[6] 的改进版本和中文训练数据特化版本 [11]。使用 HuggingFace 网站上开源的预训练模型参数。

2.2.2 微调的样本设计

在上述预训练模型的基础上，设计测试小样本微调和全训练集微调的效果。由于设计的基于 RAG 技术的抽取模型并不需要针对下游任务的额

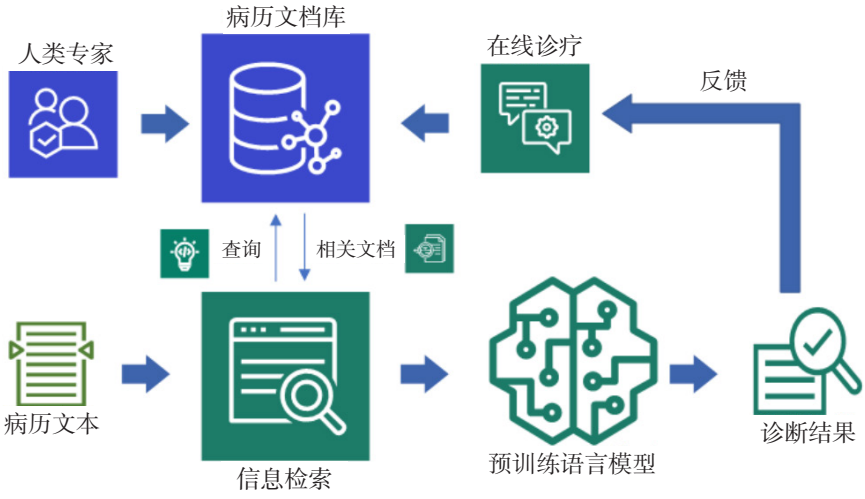

图2 检索增强生成技术用于电子病历文本挖掘的应用流程

Figure 2. Application process of retrieval augmented generation technology for electronic medical record text mining

外模型训练工作，因此预训练 – 微调方法做零样本学习才能形成公平的对比。但是预训练 – 微调方法的 FFNN 使用的是随机初始化的参数，无法使用完全的零样本学习，因此设计使用小样本学习。小样本的设定是在数据集中采集足够的病历，直到每个分型、分期的类别都至少有一个样本。此外，以全训练集数据作为微调样本，测试预训练 – 微调模型效果。

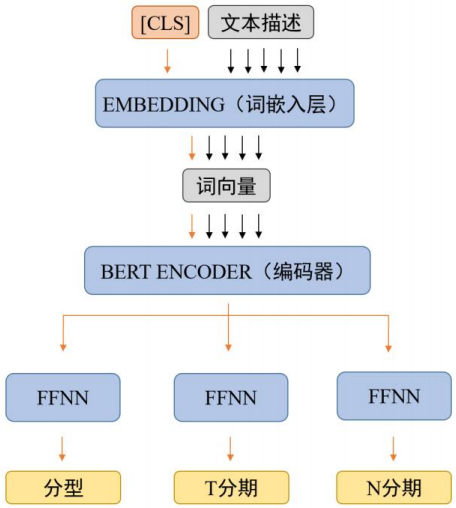

图3 传统的预训练–微调训练范式在分型分期数据上的模型结构示意图

Figure 3. Schematic diagram of the model structure on typing and staging data using the traditional pre–training and fine–tuning training paradigm

2.3 数据集划分及模型评价

基于已标注、审核的 484 条膀胱癌病理记录进行实验，每条记录涵盖有提取的膀胱癌分型、T 分期和 N 分期结果。将数据集划分成训练集、验证集和测试集，其中，训练集、验证集、测试集分别包含 384、48、52 条数据。以准确率和 F1 值作为各模型效果的评价指标，比较基于 RAG 技术取前 5 或前 10 个样本作为参考的抽取模型（MedRAG@5、MedRAG@10）的结果，以及 BERT 模型和 RoBERTa 模型在小样本微调的结果。

3 结果

由于预训练 – 微调模型使用验证集结果选择最好的模型，因此验证集相对测试集的表现更好。从测试集结果来看，对于分型提取，基于 RAG 技术的抽取模型的效果与小样本微调的传统预训练 – 微调模型效果相当，对于 T 分期提取，前者优于后者，对于 N 分期，前者劣于后者。同时 MedRAG@10 模型效果有时劣于 MedRAG@5，见表 1。

预训练 – 微调模型在全训练集微调的结果见表 2。在分型及分期的提取上，其准确率、F1 值较小样本微调的结果均有较大幅度的提升，在多数指标上 RoBERTa 模型表现优于 BERT 模型，但是对于 N 分期，RoBERTa 模型由于更大的参数量出现了过拟合。相对来说，T 分期的提取准确率低于分型和 N 分期。

选择表现性能更佳的模型为前述真实世界研究数据标注平台生成基于模型的推荐抽取结果，示例如图 4。通过模型辅助自动抽取，可以大幅提升信息的抽取效率，节省了人工标注的成本，只需通过“审核”完成最终提取。

表1 验证集、测试集膀胱癌分型及分期抽取结果（%）

Table 1. Bladder cancer typing and staging extraction results on validation set and test set (%)

数据集	方法	分型		T分期		N分期		平均F1值
		准确率	F1值	准确率	F1值	准确率	F1值	
验证集	MedRAG@5	27.08	19.10	37.50	32.66	29.17	17.71	23.16
	MedRAG@10	27.08	18.88	33.33	28.15	31.25	17.56	21.53
	BERT#	45.83	36.88	8.33	8.50	70.83	60.27	35.22
	RoBERTa#	39.58	27.51	12.50	12.76	68.75	57.10	32.46
测试集	MedRAG@5	32.69	23.50	40.38	33.78	28.85	15.18	24.15
	MedRAG@10	32.69	23.35	44.23	37.04	26.92	14.32	24.90
	BERT#	32.69	24.22	15.38	12.46	65.38	53.65	30.11
	RoBERTa#	36.54	25.44	17.31	15.90	65.38	53.90	31.75

注：#使用小样本微调；MedRAG@5. 基于RAG技术，取前5个样本作为参考的GPT-3.5抽取模型；MedRAG@10. 基于RAG技术，取前10个样本作为参考的GPT-3.5抽取模型；BERT. 来自变换器的双向编码器表示（bidirectional encoder representations from transformers）；RoBERTa. 一种经过鲁棒优化的BERT预训练方法（a robustly optimized BERT pretraining approach）。

表2 基于全训练集微调的预训练–微调方法结果（%）

Table 2. Results of pre–training and fine–tuning method based on fine–tuning of full training set (%)

数据集	方法	分型		T分期		N分期		平均F1值
		准确率	F1值	准确率	F1值	准确率	F1值	
验证集	BERT	87.50	86.31	77.08	75.90	87.50	85.20	82.47
	RoBERTa	91.67	90.74	83.33	80.86	87.50	85.36	85.65
测试集	BERT	67.31	66.08	50.00	47.13	80.76	78.81	64.01
	RoBERTa	73.08	71.06	53.85	50.18	75.00	73.65	64.96

注：BERT. 来自变换器的双向编码器表示（bidirectional encoder representations from transformers）；RoBERTa. 一种经过鲁棒优化的BERT预训练方法（a robustly optimized BERT pretraining approach）。

图4 标注平台中疾病分型分期自动抽取作为推荐结果的示例

Figure 4. Example of automatic extraction of disease typing and staging as recommended results in the annotation platform

注：分期“值”为“T分期，N分期”结果。

4 讨论

本研究从真实世界电子病历数据中的非结构化数据处理问题出发，探索了真实世界数据标注平台的搭建及工作流程，同时基于已标注及审核的膀胱癌病理数据，探索了 RAG 结合大模型及当前主流的预训练 – 微调方法的自动抽取疾病分型、分期的效果。

为了实现海量医学数据尤其是电子病历数据中非结构化数据到结构化数据的转换，已有较多研究团队开始开展医学知识标注体系的设计及系统构建[12]，以及面向医疗文本的标注平台构建[13–14]。本研究为了更方便高效地标注、审核膀胱癌电子病历数据，建立了真实世界数据标注平台，通过其可视化界面及任务领取、标注、审核的流程设计及后台管理，优化了人工提取的质量和效率。

RAG 技术是当前自然语言处理领域的热门研究方向之一[9, 15]，RAG 模型可以被看作是赋予了生成式模型自主搜寻资料的能力。其优点在于可以有效利用知识库中的信息来提高生成的准确性、流畅性和丰富性，同时避免了直接使用大规模知识库的计算和存储开销。本研究利用 RAG 技术检索相关病理记录文本及其标注结果，并将这些信息与待提取的新病理记录文本相结合，通过添加提示后，一并输入 LLM 进行处理。LLM 会从前面已标注的样本中得到指导，判断待提取的病理记录中膀胱癌的分型分期结果。

自动化抽取的结果显示，基于 RAG 技术的抽取方法对于 T 分期的抽取优于基于小样本微调的预训练 – 微调方法，但在 N 分期及分型抽取上劣于或微劣于小样本微调的预训练 – 微调方法。在小样本场景，两种方法对比，RAG 技术的优势在于不需要针对下游数据的训练过程，大大减轻了迁移到其他医学领域的难度，但尚待在跨机构、跨病种数据中验证，此外，RAG 能够展示出

推理过程中提供参考依据的相关文档，这为理解模型的决策过程提供了一定的可解释性。对于基于 RAG 技术的抽取模型，参考 10 个病理记录的模型效果有时反而劣于参考 5 个病理记录的模型，可能是因为用于检索的、已标注的数据库仍相对较小，在碰到罕见分期或分型结果时，检索模型会纳入更多的相关度较差的样本，进而干扰 LLM 的判断。

总体上，基于 RAG 技术的抽取方法及基于小样本微调的预训练 – 微调这两种方法抽取非结构化病理报告中分型、分期信息的平均 F1 值低于 0.4，效果欠佳。而基于全训练集微调的预训练 – 微调方法在验证集、测试集的平均 F1 值分别在 0.8、0.6 以上，抽取效果较为理想，提示对于疾病特异性较强的非结构化病理记录，通用领域的预训练语言模型参数中存储的知识无法覆盖此专业领域，需要以一定量的专业领域数据进行微调后，方能提升模型抽取的准确率。在微调过程中，模型会学习特定任务的信息，同时保留之前学到的通用语言特征。

本研究中 RAG 结合大模型方法抽取效果欠佳，可能有以下几方面的原因：一是 RAG 对检索系统的依赖性较强，模型的表现在很大程度上依赖底层检索系统的效果，本研究使用的 BM25 检索算法的检索能力可能有待提升，后续工作可以考虑使用更先进的中文检索算法；二是本研究所使用的大模型 gpt-3.5-turbo 在中文医疗领域能力有待提升，后续工作可以考虑使用在中文医疗领域表现更好的大模型；三是用于检索的数据库有待扩展，后续工作可以考虑增加已标注的病历数据或指南等权威文件、文献等更新的知识作为外部知识库，提升抽取的准确率；四是，分型、分期信息在部分病理报告中未明确报告，因此对于这些缺失信息的抽取还涉及推理过程，例如，N 分期的判断需要先明确淋巴结转移部位及对应数量，再结合 N 分期判断标准综合推断。因此对比直接抽取肿瘤大小等原文中已存在的信息，分型、分期的抽取更加困难，这可能也导致了本研究的准确率不够理想。

前期研究中基于正则表达式方法提取膀胱癌病理报告中 T 分期、N 分期的 F1 值分别为 0.90、0.88[3]，抽取的准确率优于本研究中基于预训练语言模型的方法。可能是因为病理报告中包含许多专业术语、缩写以及复杂的句子结构，模型未在预训练语料库中充分接触到这些专业词汇或表达，导致其在微调阶段难以准确识别。例如，“送检右盆腔（3/4 枚）淋巴结可见癌转移”指得是送检了右盆腔 4 枚淋巴结，其中阳性淋巴结有 3 枚，这一表述规则性较强，而对于这类格式固定、规则明确的文本信息，正则表达式能够通过精确的规则直接定位和抽取信息，存在优势。但对于复杂或不规则的语言模式，编写正则表达式规则就比较费时费力，扩展到其他领域时泛化能力也较差。考虑到本研究是基于预训练语言模型的初步尝试，标注数据有限，未来有待更大的数据集、更先进的算法、更适配的模型、更多元的外部知识库以深入探索预训练语言模型在电子病历信息抽取中的效果。

综上所述，尽管人工智能技术在医疗领域展现出巨大的应用潜力，但在处理真实世界电子病历数据中的非结构化信息方面仍存在一定的挑战。本研究通过真实世界标注平台的搭建进行膀胱癌病理数据的标注与审核，并基于已标注的数据初步探索了 RAG 结合大模型方法及基于 BERT 模型、RoBERTa 模型的预训练 – 微调方法自动化抽取疾病分型、分期的效果，并将自动化抽取结果作为标注平台的推荐结果加速电子病历数据的非结构化信息抽取，以加速知识挖掘与数据赋能。未来研究应继续优化模型和训练策略，以提高信息抽取的准确性和效率，为医疗领域的人工智能应用提供更加可靠的技术支持。在此基础上，有望推动医疗信息化进程，助力临床决策支持系统的发展，为患者提供更优质的医疗服务。

参考文献

1 李绪辉，阎思宇，陈沐坤，等 . 面向真实世界的知识挖掘与知识图谱补全研究（一）：真实世界数据与知识图谱概述 [J]. 医学新知，2023, 33(2): 130-135. [Li XH, Yan SY, Chen MK, et al. Research on real-world knowledge mining and knowledge graph completion (I): overview of real-world data and knowledge map[J]. Yixue Xinzhi Zazhi, 2023, 33(2): 130-135.] DOI: 10.12173/j.issn.1004-5511.202301018.

2 阎思宇，李绪辉，陈沐坤，等 . 面向真实世界的知识挖掘与知识图谱补全研究（二）：非结构化电子病历信息抽取方法及进展 [J]. 医学新知，2023, 33(5): 358-365.

[Yan SY, Li XH, Chen MK, et al. Research on realworld knowledge mining and knowledge graph completion (II): methods and progress of information extraction from unstructured electronicmedical records[J]. Yixue Xinzhi Zazhi, 2023, 33(5): 358-365.] DOI: 10.12173/j.issn.1004-5511.202301016.

3 马文昊, 石涵予, 黄桥, 等. 面向真实世界的知识挖掘与知识图谱补全研究(三): 基于正则表达式对膀胱癌真实世界数据的结构化信息抽取[J]. 医学新知, 2024, 34(3): 312-321. [Ma WH, Shi HY, Huang Q, et al. Research on real-world knowledge mining and knowledge graph completion (III): structured information extraction from real world data of bladder cancer based on regular expression[J]. Yixue Xinzhi Zazhi, 2024, 34(3): 312-321.] DOI: 10.12173/j.issn.1004-5511.202308006.

4 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv: 1706.03762, 2017. http:arxiv.org/abs/1706.03762.

5 Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2019. http:arxiv.org/abs/1810.04805.

6 Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized bert pretraining approach[J]. arXiv: 1907.11692, 2019. http:arxiv.org/abs/1907.11692.

7 Li L, Zhou J, Gao Z, et al. A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs)[J]. arXiv: 2405.03066, 2024. http://arxiv.org/abs/2405.03066.

8 Li D, Kadav A, Gao A, et al. Automated clinical data extraction with knowledge conditioned LLMs[J]. arXiv: 2406.18027, 2024. http://arxiv.org/abs/2406.18027.

9 Pelletier AR, Ramirez J, Adam I, et al. Explainable biomedical hypothesis generation via retrieval augmented generation enabled large language models[J]. arXiv: 2407.12888, 2024. http://arxiv.org/abs/2407.12888.

10 董文波, 孙仕亮, 殷敏智. 医学知识推理研究现状与发展[J]. 计算机科学与探索, 2022, 16(6): 1193-1213. [Dong WB, Sun SL, Yin MZ. Research and development of medical knowledge graph reasoning[J]. Journal of Frontiers of Computer Science & Technology, 2022, 16(6): 1193-1213.] DOI: 10.3778/j.issn.1673-9418. 2111031.

11 Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for Chinese BERT[J]. arXiv: 1906.08101, 2019. abs/1906.08101. http://arxiv.org/abs/1906.08101.

12 马鹤桐, 王序文, 沈柳, 等. 医学知识标注体系设计与系统构建[J]. 中国卫生标准管理, 2023, 14(21): 1-4. [Ma HT, Wang XW, Shen L, et al. Medical knowledge labeling system design and annotation system construction[J]. China Health Standard Management, 2023, 14(21): 1-4.] DOI: 10.3969/j.issn.1674-9316.2023.21.001.

13 张坤丽, 赵旭, 关同峰, 等. 面向医疗文本的实体及关系标注平台的构建及应用[J]. 中文信息学报, 2020, 34(6): 36-44. [Zhang KL, Zhao X, Guan TF, et al. A platform for entity and entity relationship labeling in medical texts[J]. Journal of Chinese Information Processing, 2020, 34(6): 36-44.] DOI: 10.3969/j.issn.1003-0077.2020.06.006.

14 张辉, 连万民, 刘翔, 等. 麻醉与围术期医学科数据标注平台的设计与实现[J]. 中国数字医学, 2021, 16(1): 96-100. [Zhang H, Lian WM, Liu X, et al. Design and implementation of the data annotation platform of department of anesthesia and perioperative medicine[J]. China Digital Medicine, 2021, 16(1): 96-100.] DOI: 10.3969/j.issn.1673-7571.2021.01.021.

15 Alkhalaf M, Yu P, Yin M, et al. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records[J]. J Biomed Inform, 2024, 156: 104662. DOI: 10.1016/j.jbi.2024.104662.

收稿日期：2024 年 09 月 02 日 修回日期：2024 年 10 月 11 日
本文编辑：李绪辉 曹 越

引用本文：阎思宇, 谭杰骏, 朱海锋, 等. 面向真实世界的知识挖掘与知识图谱补全研究(四): 真实世界数据标注平台搭建及基于预训练语言模型的自动化抽取方法探索[J]. 医学新知, 2024, 34(11): 1276-1283. DOI: 10.12173/j.issn.1004-5511.202409004.
Yan SY, Tan JJ, Zhu HF, et al. Research on real-world knowledge mining and knowledge graph completion (IV): construction of a real-world data annotation platform and exploration of automatic extraction method based on pre-trained language models[J]. Yixue Xinzhi Zazhi, 2024, 34(11): 1276-1283. DOI: 10.12173/j.issn.1004-5511.202409004.