面向语义的电子病历智能文本检索技术研究

惠　婷　申艳妮

（空军军医大学第二附属医院信息科　西安　710038）

〔摘要〕　**目的/意义**　探索面向语义的智能检索方法，以实现电子病历文本数据语义互操作，提升检索效能。**方法/过程**　引入医学专业术语库与层次化扩展机制增强检索，利用预训练语言模型对查询和病历文本进行深度语义编码与相似度计算，实现基于上下文理解的高阶语义检索。**结果/结论**　该方法可显著提升准确率、查全率，为实现高效、精准的电子病历智能检索提供了可行方案。

〔关键词〕　电子病历；智能检索；预训练语言模型；语义分析；语义检索

〔中图分类号〕R-058　　〔文献标识码〕A　　〔DOI〕10.3969/j. issn. 1673-6036. 2026. 02. 013

Study on Intelligent Text Retrieval Technology for Semantic Electronic Medical Records

HUI Ting, SHEN Yanni

Department of Information, Second Affiliated Hospital of Air Force Military Medical University, Xi'an 710038, China

〔**Abstract**〕　**Purpose/Significance** To explore semantic-oriented intelligent retrieval methods, and to achieve semantic interoperability of electronic medical record (EMR) text data and enhance retrieval efficiency. **Method/Process** A medical terminology database and a hierarchical expansion mechanism are used to enhance retrieval. Pre-trained language models are utilized to conduct deep semantic encoding and similarity calculation on queries and medical record texts, achieving advanced semantic retrieval based on context understanding. **Result/Conclusion** The method can significantly improve the accuracy and recall rate, and provide a feasible solution for efficient and precise intelligent retrieval of EMR.

〔**Keywords**〕　electronic medical record (EMR); intelligent retrieval; pre-trained language model; semantic analysis; semantic retrieval

1　引言

电子病历产生于临床治疗过程，是由医疗机构通过信息化系统生成的数字化医疗记录，涵盖临床诊疗记录、检查检验结果、医嘱、手术记录等全流程信息[1]。其中包含丰富的临床细节、诊疗推理过程和患者整体情况等，可为医学知识挖掘、临床辅助诊断及精准医疗提供数据支撑[2]。随着医疗信息化技术的飞速发展，电子病历数据结构不断完善，数据规模呈指数增长。与此同时，个体化医疗的发展使病历信息检索需求日趋复杂，传统关键词匹配检索方式的局限性逐渐突显。一是电子病历中约80%的信息以自由文本形式存在，这些信息（如主诉、诊断等）缺乏固定结构，无法使用传统SQL语句进行查询[3]；二是医学语言词汇具有专业性、逻辑复杂性及语义模糊性，而传统检索方式无法领悟用户检索意图及隐含语义内容[4]，检索结果的关联

〔修回日期〕　2025-12-24

〔作者简介〕　惠婷，工程师，发表论文4篇；通信作者：申艳妮，工程师。

性和准确性较差；三是基于检索扩展算法的检索方式虽然可在传统查询结果中进行二次扩展，提高召回率，但基于权重的结果排序存在漂移风险，降低了检索性能的稳定性[5]。

近年来，以 BERT、GPT 为代表的预训练模型，通过融合人类反馈强化学习与监督微调机制，可有效捕捉文本间复杂语义关联，生成高质量文本[6]，目前已广泛应用于医学问答[7]、临床辅助决策[8]、病历信息提取与结构化[9]等领域，显著提升了领域知识检索的准确性与用户体验。本研究将预训练模型引入医学电子病历文本检索，提升其语义理解与检索能力。

2 总体设计与功能实现

2.1 总体设计

基于医学专业术语库与预训练模型，设计融合术语扩展与语义理解的检索方法，见图 1。

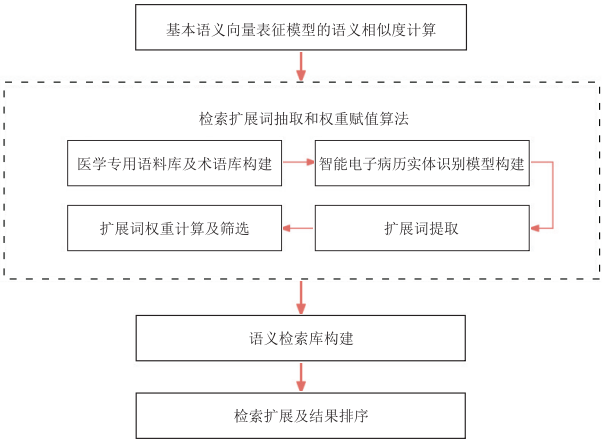

图 1 电子病历智能文本检索技术路线

2.2 基于语义向量表征模型的语义相似度计算

语义相似度计算可精准度量查询意图与病历内容深层含义的一致性[10]，实现高阶语义匹配，提升检索结果的相关性与准确性，为临床决策提供智能支持。从 2022 年 2 月 17 日—11 月 23 日空军军医大学第二附属医院入院电子病历记录中，选取临床症状相似但最终诊断不同、诊断相同但具体表现严重程度差异较大的 750 例病历，并两两配对。提取其中的病历概要、诊疗过程及治疗记录，并进行去隐私化、脱敏处理、去标识化、文本解构、代表性内容采样等标准化处理，生成病历片段对。由 5 名高年资临床专业医生对其进行独立标注，计算 Kappa 系数并评估标注一致性。标注结果以键值对形式保存，按照 7∶2∶1 比例划分训练集、测试集及验证集。针对训练集病历片段对，按 1—5 评分标注语义相似度标签，1 表示几乎不相似，5 表示语义几乎相同，形式为“(句子 x，句子 y，标签)”。选择 CoSENT、BERT、RoBERTa、Sentence-BERT 等模型进行文本向量相似性评估[11-12]，以-1、0、1 分别表示完全负相关、完全不相关及完全正相关。采用网格搜索法确定上述模型的超参数设置。采用 Spearman 相关系数、准确率作为衡量指标，对比结果，见表 1。当 λ 为 30，batch_size 为 64，最大序列长度为 512，语义向量维度为 192 时，CoSENT 模型 Spearman 相关系数及准确率最高。因此采用 CoSENT 模型进行电子病历文本特征间相似性比较。该模型针对中文短文本匹配和语义搜索设计，采用排序损失函数优化训练过程，与传统分类损失函数相比，其通过优化句子对的相对相似性排序评估权重，可构建更具判别性的语义空间，具有较高的训练与推理性能[11]。

表 1 各模型训练后性能对比

模型名称	Spearman 相关系数	准确率（%）
BERT	0.58	68.23
RoBERTa	0.69	71.82
Sentence-BERT	0.75	81.05
CoSENT	0.84	88.64

2.3 检索扩展词抽取和权重赋值算法

2.3.1 医学专用语料库及术语库构建 以脱敏电子病历记录、学术文献、医学结构化数据集、开放术语库等作为数据源。对相关数据进行清洗、去噪、词汇抽取与术语规范化。其中，词汇抽取与术语规范化是核心环节[13-14]，流程如下。第 1 步：基于 Word2Vec 方法结合 BERT 进行命名实体识别[15]，初步抽取疾病、解剖部位、检查、操作等 17.28 万条词汇，合并不同表述的同一概念（如“心肌梗

死”与“心梗”），形成11.07万条词汇。第2步：由5名高年资临床专业医生独立进行人工复核。第3步：针对不同标注者标注的同一批样本，采用一致性检验计算Kappa系数，以评估并提升标注共识。最终建立的医学专用术语库数据统计，见表2。

表2 医学专用术语库数据统计

样本来源	标注词汇数量（万条）	数据来源
电子病历记录	3.12	滨州市政务数据开放平台开放式电子病历、空军军医大学第二附属医院电子病历系统
学术文献	4.18	中国生物医学文献数据库
医学结构化数据集	1.10	医渡云开放数据源Yidu-S4K/N7K
开放术语库	2.67	全国科学技术名词审定委员会制订的“医学卫生健康”规范术语库

2.3.2 智能电子病历实体识别模型构建 基于RoBERTa-wwm-ext预训练模型，识别并抽取用户输入的检索需求中的实体信息。该模型针对完整、复合专业术语的中文命名实体识别任务设计，能够准确地理解词义、把握实体边界，并可通过微调快速适应医疗领域的特殊语境[16]。为提高其在医学专业领域的实体识别效能，对RoBERTa-wwm-ext进行微调，以CCKS2017中文电子病历命名实体识别项目（600例）及中文医学命名实体识别数据集（CMeEE-V2）（23 000例）为数据来源，选取18 000例CMeEE-V2数据作为训练集，2 000例CMeEE-V2数据为验证集，CCKS2017及其余3 000例CMeEE-V2数据作为测试集。将医学文本命名实体划分为疾病、临床表现、药物、医疗设备、医疗程序、身体、医学检验项目、微生物类、科室等类型，采用BIO格式进行标注，而后进行迭代训练。最终测试准确率为85.13%。

2.3.3 扩展词提取 为提高扩展词提取泛化能力，采用开放医学知识图谱OpenKG和国际疾病分类术语集ICD-10作为扩展词来源。通过知识图谱检索工具、ICD-10术语查询功能提取相关术语作为扩展词。提取过程中，进一步通过关系推理、层级查询获取查询词上/下位词，基于下位词迭代查询，细化检索细节；基于上位词遍历同层级术语，提高检索全面性[17]。从医学专用术语库中选择420例医学专用术语作为测试样本，采用上述方法自动提取扩展词，并由3名医学领域专家复核。

2.3.4 扩展词权重计算及筛选 共现频率通过量化词汇在语境中的分布关联间接揭示其语义内容，具有直观性、可解释性，可为扩展词权重计算提供信息补充。因此采用扩展词与查询词共现频率、语义相似度综合评估扩展词权重$\omega(w)$。其中，q_w为扩展词w对应的检索词，q_{all}为查询词集合，$sim(q_w,w)$为q_w与w间语义相似度，$coe(w,q_{all})$为w与查询词集合共现频率。有研究[18]表明，当查询扩展词数量超过20时将显著增加查询漂移概率。由于查询扩展词呈高斯分布[19]，设置扩展词权重置信区间为：$[\mu \pm \sigma]$。当扩展词数量大于20时，剔除置信区间外的扩展词。

$$\omega(w)=\frac{1}{2}\sqrt{sim(q_w,w)+coe(w,q_{all})} \quad (1)$$

2.4 语义索引库构建

语义索引库通过语义向量表征模型将文本数据转换为高维稠密向量[20]，即嵌入向量，便于以向量在向量空间中的几何距离直接表征其对应文本的语义相似性。基于微调后的RoBERTa-wwm-ext模型，将电子病历本文信息逐条转换为多维语义向量。将电子病历诊疗过程记录、治疗记录、病历概要等文本拼接，形成电子病历特征文本，将特征文本输入模型，生成对应语义向量，并通过遍历电子病历数据批量形成语义向量信息。基于Elasticsearch 7.5实现语义向量存储与索引功能，将语义向量信息批量载入索引库，建立索引映射字段，分别设置字段类型及维度为dense_vector及192，结合语义向量表征需求选择相似度排序计算方式。完成索引映射构建后，即可通过最近邻搜索接口检索访问语义向量索引信息。

2.5 检索扩展及结果排序

将用户查询需求转化为语义向量信息，通过索引映射字段匹配索引库记录。匹配结果根据关键词扩展

权重$S_{keyword}$与语义相似度$S_{semantic}$的线性加权分数排序，选择前5条记录作为推荐检索结果。其中，α通过在验证集上进行网格搜索优化确定，用以动态调节词汇精确性与语义泛化性在最终决策中的相对权重。

$$S_{final} = \alpha \times S_{keyword} + (1 - \alpha) \times S_{semantic} \quad (2)$$

3 实验设计与结果分析

3.1 实验设计

以空军军医大学第二附属医院2023年4月10日—5月25日电子病历记录为样本，经去隐私化、脱敏处理，剔除字段信息不完整记录，并按照我国《第二批罕见病目录（2023年）》《中国罕见病定义研究报告2021》筛选出胃肠间质瘤、血友病等477例罕见病电子病历记录，提取其中的诊疗过程记录、治疗记录、病历概要字段内容，并针对每例测试样本设计简单关键词查询、复杂语义查询、罕见病术语查询3类任务。以多维度全文关键词检索（以下简称关键词组）、基于医学术语库的知识图谱检索（以下简称知识图谱组）作为对照组。实验过程遵循双盲原则，利用3组检索工具获取初步结果后，由3名医学领域专家进行独立复核。最终结果依据多数专家意见确定，并以查全率与查准率作为核心评价指标。采用SPSS 23.0作为统计分析工具，针对上述指标采用卡方检验进行统计学分析。

3.2 实验结果及分析

3.2.1 电子病历简单关键词查询测试 3类任务电子病历查询测试结果，见表3。简单关键词查询任务中，3组间查准率差异无统计学意义（$\chi^2 = 1.43$，$P = 0.49$），查全率存在显著差异（$\chi^2 = 8.87$，$P = 0.01$）。两两比较中，知识图谱组、智能文本组的查全率均显著高于关键词组。知识图谱检索利用预定义的概念间关系进行查询扩展和推理，实现字面不同但概念相关的扩展检索。智能文本检索基于深度语义表征，实现语义相似性匹配，克服词汇鸿沟，能够捕捉上下文隐含的语义信息，提高领域适配程度和对专业逻辑的捕捉能力[21]。而多维度全文关键词检索受限于词汇表覆盖度和表述变异性，无法识别未在预定义词典中的术语、缩写、俗称或错误拼写，更无法理解概念间的深层语义关联。

表3 电子病历查询测试结果(%)

组别	简单关键词查询		罕见病术语查询		复杂语义查询	
	查准率	查全率	查准率	查全率	查准率	查全率
关键词组	88.26	66.67	37.94	15.09	66.25	53.46
知识图谱组	84.49	75.26	70.02	64.36	77.99	72.33
智能文本组	86.16	79.66	77.99	71.07	85.75	81.55

3.2.2 电子病历罕见病术语查询测试 罕见病术语查询任务中，3组间查准率、查全率差异均具有统计学意义（$\chi^2 = 86.1$，$P < 0.01$；$\chi^2 = 134.5$，$P < 0.01$）。基于Bonferroni校正进行两两比较发现，智能检索组与知识图谱组均显著优于关键词检索组，且智能检索组显著优于知识图谱组。这表明在术语标准化挑战突出的罕见病领域，依赖字面匹配的传统检索方法已然失效。引入结构化医学知识（如知识图谱）或深度语义模型，通过从“字符串匹配”到“概念匹配”，再到“语义匹配”的范式升级，有望带来检索性能的跨越式提升。

3.2.3 电子病历复杂语义查询测试 在复杂语义查询任务中，知识图谱组的查准率和查全率均显著高于关键词组（$\chi^2 = 16.35$，$P<0.05$；$\chi^2 = 36.38$；$P<0.05$）。其原因可能与医学语言自身的专业化、复杂性及模糊性有关，电子病历中存在复杂术语体系与众多同义词、多义词、缩略语以及分型现象，传统关键词组识别困难。知识图谱包含同义词、上下位词等关系，基于知识图谱的电子病历检索可将患者、症状、诊断、检查、治疗、药品、医生等实体相关联，实现深度关联查询与推理，提高用户提问关键词的兼容性，避免漏检[7]；还能通过实体间的复杂关系，推理发现间接相关信息，进一步提升查全率。例如：当用户搜索“前胸痛”时，系统能自动扩展查询，

找到包含“胸痛”“心前区疼痛”等所有相关术语的病历。智能文本组查准率和查全率分别为85.75%和81.55%，均高于知识图谱组，且差异具有显著性（$\chi^2 = 9.67$，$P<0.05$；$\chi^2 = 11.44$，$P<0.05$）。这表明深度学习模型可避免知识图谱检索受限于内置本体或词汇表等问题。另外，深度学习模型可以端到端地处理整段原始文本，理解复杂的上下文依赖与指代关系。电子病历内也存在复杂的指代消解现象，深度学习模型能够准确关联到所指的具体实体，同时，还能够理解医学诊疗中出现的时序性与动态演变现象，进一步提升了智能文本组的检索准确率。

4 结语

本研究构建面向语义的电子病历智能文本检索技术体系，克服了传统关键词匹配在语义理解与上下文感知方面的局限。通过医学术语库实现标准化与概念映射，解决了领域词汇多样性问题；利用深度学习预训练模型进行语义表示学习，精准捕获非结构化文本中的复杂依赖与指代关系；进而基于向量相似度计算与语义检索库，实现了高效、准确的病历语义检索[6]。本研究仍存在以下局限，一是检索性能依赖标注数据的质量与规模，二是对于病历中长上下文、复杂逻辑的叙事型文本解析能力有待进一步验证。未来技术将向多模态信息融合与知识图谱增强的语义理解方向演进，推动检索系统向主动决策支持与跨机构知识互联升级，为智慧医疗生态构建提供核心支撑，也为医学相关领域信息检索提供技术借鉴。

作者贡献：惠婷负责研究设计及实施、论文撰写；申艳妮负责文献调研、论文审核。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 任蔚文. 基于中文电子病历的疾病诊断方法研究[D]. 北京：北京化工大学，2025.

2 于家畦，康晓东，白程程，等. 一种新的中文电子病历文本检索模型[J]. 计算机科学，2022，49(S1)：32-38.

3 熊回香，周明洁. 电子病历中基于实体识别和共现分析的疾病间语义关系挖掘研究[J]. 情报科学，2025，43(6)：14-27.

4 吴进发. 电子病历搜索引擎中的新词发现和排序技术研究[D]. 成都：电子科技大学，2021.

5 邓兰，徐进. 一种高效安全的密文电子病历多关键字检索方案[J]. 医学信息学杂志，2022，43(1)：49-53.

6 刘洪洲，张鸿. 基于语义调节与两级匹配的图像文本检索方法[J]. 计算机技术与发展，2024，34(12)：100-107.

7 秦乐，勾智楠，王培伍，等. 基于提示引导多跳推理的医学诊断检索增强生成[J]. 计算机应用研究，2025，42(10)：2956-2963.

8 谭平，刘惠娜，韦昌法. 融合大语言模型与知识图谱的抑郁症中西医结合智能问答系统构建研究[J]. 上海中医药杂志，2025，59(7)：1-10.

9 张文超，郜勇，王玉阳，等. 大语言模型技术驱动的全周期病历质量管理设计与实践探索[J]. 中国数字医学，2025，20(8)：9-14.

10 周朝阳，贺艳菊，夏岭梅，等. 融合词性与语义相关性的图书馆智能咨询系统问句相似性计算方法研究[J]. 情报探索，2024(5)：1-8.

11 陈健飞，卜凡亮，王一帆. 基于CoSENT和改进K-means的冒犯性评论文本主题识别[J]. 科学技术与工程，2024，24(31)：13442-13449.

12 张超，杨玉芳. 基于语义智能化识别的非结构化问答信息多模态检索方法[J]. 中国新技术新产品，2025(11)：50-52.

13 杨睿. 基于知识图谱的网络信息资源智能检索系统设计[J]. 电子设计工程，2025，33(10)：103-106，111.

14 贾楠，胡冠宇. 基于大语言模型的科技项目查新方法研究[J]. 图书馆研究与工作，2025(10)：38-43.

15 周潇，高雅倩，樊嘉逸. 基于BERT词嵌入的专利检索策略研究[J]. 情报学报，2023，42(11)：1347-1357.

16 韩振桥，付立军，刘俊明，等. 结合RoBERTa与多策略召回的医学术语标准化[J]. 计算机系统应用，2022，31(10)：245-253.

17 吴欢，车贺宾，王万玲，等. 基于循证医学和电子病历数据的通用医学知识图谱构建[J]. 医学信息学杂志，2025，46(2)：22-28.

18 余传明，蔡林，胡莎莎，等. 基于深度学习的查询扩展研究[J]. 情报学报，2019，38(10)：1066-1077.

19 秦钰淑，杨良怀，朱艳超，等. 融合图像与文本特征的组合检索方法[J]. 电子学报，2025，53(2)：558-567.

20 王诣涵. 基于蒸馏和量化的高效语义检索研究[D]. 北京：北京邮电大学，2024.

21 吉旭瑞，魏德健，张俊忠，等. 中文电子病历信息提取方法研究综述[J]. 计算机工程与科学，2024，46(2)：325-337.