计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(10):37−46 [doi: 10.15888/j.cnki.csa.009635]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于知识图谱和预训练语言模型的儿童疫苗接种风险预测①

吴英飞[1], 刘　蓉[1], 李明燕[2,3], 季　钗[2,3], 崔朝健[1]

[1](杭州师范大学 信息科学与技术学院, 杭州 311121)
[2](浙江大学医学院附属儿童医院 儿童保健科, 杭州 310003)
[3](国家儿童健康与疾病临床医学研究中心, 杭州 310003)
通信作者: 季　钗, E-mail: 6198011@zju.edu.cn

摘　要: 基层医疗机构的医生缺少患病儿童疫苗接种风险的判断能力, 通过学习高水平医院医生的经验来研发儿童疫苗接种风险预测模型, 从而帮助基层医疗机构医生快速筛查高风险患儿, 是一种可行的方案. 本文提出了一种智能化的基于知识图谱的疫苗接种建议推荐方法.首先, 提出了一种基于预训练语言模型的医学命名实体识别方法ELECTRA-BiGRU-CRF, 用于门诊电子病历命名实体抽取. 其次, 设计疫苗接种本体, 定义关系及属性, 基于 Neo4j 构建了中文儿童疫苗接种知识图谱. 最后, 基于构建的中文疫苗接种知识图谱, 提出了一种基于预训练语言模型进行显著性类别指导的疫苗接种建议分类推荐方法. 实验结果表明, 本文研究方法可以为医生提供辅助诊断, 对于患病儿童能否接种疫苗提供决策支持.

关键词: 中文电子病历; 预训练语言模型; 知识图谱; 命名实体识别; 疫苗接种建议

引用格式: 吴英飞,刘蓉,李明燕,季钗,崔朝健.基于知识图谱和预训练语言模型的儿童疫苗接种风险预测.计算机系统应用,2024,33(10):37–46. http://www.c-s-a.org.cn/1003-3254/9635.html

Risk Prediction of Child Vaccination Based on Knowledge Graph and Pre-trained Language Model

WU Ying-Fei[1], LIU Rong[1], LI Ming-Yan[2,3], JI Chai[2,3], CUI Zhao-Jian[1]

[1](School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China)
[2](Child Healthcare Department, Children's Hospital of Zhejiang University School of Medicine, Hangzhou 310003, China)
[3](National Clinical Research Center for Child Health, Hangzhou 310003, China)

Abstract: Primary healthcare providers lack the ability to assess the risk of vaccination for children with certain illnesses. It is a viable solution to developing a risk prediction model for pediatric vaccination, by leveraging the experience of healthcare professionals in tertiary hospitals, to assist primary healthcare providers in swiftly identifying high-risk pediatric patients. This study proposes an intelligent method for vaccine recommendations based on a knowledge graph. Firstly, a method for medical named entity recognition called ELECTRA-BiGRU-CRF, based on pre-trained language models, is proposed for named entity extraction from outpatient electronic medical records. Secondly, a vaccination ontology is designed, with relationships and attributes defined, to construct a Chinese childhood vaccination knowledge graph based on Neo4j. Finally, a method for vaccine recommendations guided by significant categories using pre-trained language models is proposed based on the constructed knowledge graph. Experimental results indicate that the proposed methods can provide diagnostic assistance to physicians and offer support for deciding whether vaccines can be administered to children with certain illnesses.

① 基金项目: 浙江省自然科学基金 (TGY24H260008)
收稿时间: 2024-02-07; 修改时间: 2024-03-05, 2024-05-06; 采用时间: 2024-05-09; csa 在线出版时间: 2024-08-21
CNKI 网络首发时间: 2024-08-22

Key words: Chinese electronic medical record; pre-trained language model (PLM); knowledge graph; named entity recognition (NER); vaccination recommendation

患病儿童接种疫苗后可能产生异常反应, 严重者甚至危及生命[1]. 2015 年起我国范围内免疫接种后不良事件 (adverse events following immunization, AEFI) 报告呈上升趋势[2]. 受一些疫苗事件和报道的影响, 部分公众对疫苗产生怀疑态度, 甚至产生疫苗犹豫[3], 不利于疫苗接种工作的良性循环.

由于相关知识的不完全了解, 社区卫生服务工作者通常对于患病儿童的疫苗接种犹豫不决[1], 尤其对于经验较少的新手医生[4]而言. 目前, 专科医院开设了儿童保健科免疫接种咨询门诊, 专家医生凭借其临床经验为患病儿童诊断并给出疫苗接种处理建议. 通过学习高水平医院医生的经验来研发儿童疫苗接种预测模型, 从而帮助基层医疗机构医生快速筛查高风险患儿, 是一种可行的方法.

因此, 迫切需要一种智能化方法, 通过利用已有的医疗知识来辅助医生对疾病儿童健康状况进行评估, 从而给出合理的接种建议决策. 这将有助于提高医疗服务的质量和效率, 更好地满足病人的需求.

医学知识图谱是推进智慧医疗的关键技术, 也是医疗信息智能化管理的基础[5], 被广泛应用在医学辅助诊断、个性化推荐、药物间关系发现等方面. 医学知识图谱与人工智能技术相结合, 可以提高医生的诊断效率, 在智慧医疗建设中具有关键的作用. 从电子病历中抽取知识对构建知识图谱至关重要, 但是中文电子病历文本形式多样, 其中存在的实体信息具有语义模糊、上下文不确定等特点. 高质量的知识抽取工作一定程度上影响着知识图谱下游任务的开展.

在这一背景下, 本研究提出了一种基于预训练模型的儿童免疫接种咨询门诊电子病历的命名实体识别 (named entity recognition, NER) 方法 ELECTRA-BiGRU-CRF. 在基于该方法构建出的中文疫苗接种知识图谱上, 本文提出一种基于预训练语言模型的显著性类别指导的疫苗接种推荐方法. 通过对比基线模型, 验证了本文所提出方法的有效性. 本文研究方法为免疫干预工作中的医生提供了宝贵的资源, 能够有效地辅助医生对儿童健康进行评估并给出合理的接种建议.

1 相关工作

1.1 疫苗接种风险分析研究

随着免疫规划的不断发展, 对 AEFI 事件的关注逐渐增加, 尤其是在儿童预防接种工作中[6].

在预测儿童疫苗接种风险的做法上, 研究者们大多采用统计学方法. 例如, 胡丹丹等人[7]分析神经系统疾病儿童免疫接种现状、相关的神经系统反应以及禁忌证等, 提出了不同神经系统疾病儿童免疫接种策略. Wang 等人[8]对疫苗接种与系统性红斑狼疮和风湿关节炎发病风险的关系进行了系统评价和分析.

统计学方法大多对接种疫苗后的儿童疫苗进行监测, 统计分析不良反应报告发生率, 以提出预测和规避疫苗接种不良反应的方法, 且取得了一定的进展. 但这种方法需要进行随访跟踪和系统分析, 耗费大量的人力物力, 同时也难以避免医疗资源紧张和浪费等问题.

1.2 知识图谱应用研究

医学知识图谱的应用是目前垂直领域知识图谱的研究热点之一[9], 其旨在有组织地充分利用庞大的数据资源, 为医疗信息化提供支持.

越来越多的研究人员致力于医学领域知识图谱的开发与应用. 比如, Chandak 等人[10]整合 20 个高质量的资源, 构建了疾病丰富和具有功能性的多模态知识图谱 PrimeKG. Joshi 等人[11]设计了一种深度神经网络, 用于预测药物在生产阶段或临床试验阶段引起的不良反应. Tao 等人[12]提出了一种分类模型用于从个人健康知识库中进行知识挖掘来帮助发现潜在的疾病威胁.

医疗领域知识图谱的研究呈现出蓬勃发展的趋势, 而中文医疗知识图谱的发展面临着语言特殊、知识来源繁杂、开源程度不足等问题, 构建仍然存在着许多困难[13,14]. 构建医学知识图谱的关键在于图谱三元组的抽取, 而在三元组抽取工作中, 医学命名实体的抽取直接决定了构建的知识图谱质量.

1.3 预训练语言模型

大规模预训练语言模型 (pre-trained language model, PLM)[15]是近年来发展起来的重要技术, 在自然语言处理领域取得了显著进展. PLM 通过大规模的无监督学习, 让模型在大量文本数据上进行预先训练, 从

而学到丰富的语言表示. BERT[16]是 PLM 中的代表作之一, 通过预先训练双向 Transformer 编码器[17], 使得模型能够同时考虑上下文的信息, 从而更好地捕捉词汇之间的关系. BERT 在问答、文本分类和命名实体识别等任务上取得了领先水平.

预训练语言模型的发展过程中涌现了一系列变体和改进方法, 通过引入不同的训练策略、更大的数据集或改进的架构, 进一步提升了在各种任务上的性能, 医学领域的预训练语言模型也应运而生. 目前 PLM 也面临着一些挑战, 模型巨大的参数规模和训练过程需要大量的计算资源, 在部分场景下限制了模型的应用范围. 其次医学文本术语的特殊性和敏感性也带来了应用的挑战.

1.4 实体识别方法研究

NER 任务旨在从医学文本中识别各种实体. 目前 NER 方法主要分为 3 种. 基于字典和规则的方法侧重通过匹配的方法去处理文本, 但对医学词典库的依赖性高且维护成本巨大. 随着监督学习的发展, 基于统计和机器学习的方法成为主流. 然而传统的机器学习算法需要复杂的特征工程, 且模型泛化能力受限. 相较之下, 深度学习方法的显著优势在于其端到端的训练过程, 无需手动设计特征.

近年来融合深度学习方法和机器学习模型的 NER 方法不断涌现, 用以提高效率和结果精度[18]. Wang 等人[19]提出了两种扩展 BiLSTM 的不同架构和特征表示方案来处理中文电子病历实体识别任务. 自 Transformer 被提出以来, 各种预训练语言模型层出不穷. 有研究表明, 使用预训练语言模型作为词嵌入层可以提高实体识别模型的序列标注性能[20]. 将深度神经网络用于文本特征学习, 无需人工手动定义模板, 且具有较强的泛化能力, 已成为目前 NER 任务的主流方法.

2 关键技术研究

2.1 基于 ELECTRA-BiGRU-CRF 的实体识别方法

在构建儿童疫苗接种知识图谱的过程中, 从电子病历中提取有价值的医疗信息至关重要. 本文提出了一种基于预训练语言模型 ELECTRA[21]的医学信息实体识别模型 ELECTRA-BiGRU-CRF. 该模型主要分为 3 层, 分别为 ELECTRA 层、BiGRU 层以及 CRF 层. 模型结构图如图 1 所示. 模型实体识别的过程包括以下步骤.

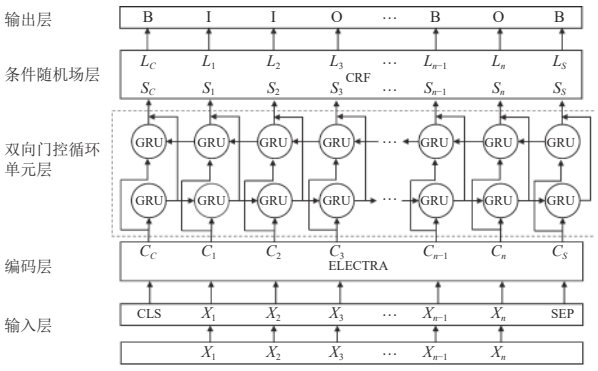

图 1 ELECTRA-BiGRU-CRF 模型框架图

1) 对原始语料进行预处理, 将文本数据$T=(X_1,X_2,\cdots,X_n)$编码为序列$X=(\text{CLS},X_1,X_2,\cdots,X_n,\text{SEP})$, 其中$X_i$表示第 i 个单词或字符.

2) 将序列$X=(\text{CLS},X_1,X_2,\cdots,X_n,\text{SEP})$输入到 ELECTRA 模型的编码器中. 生成器和判别器利用对抗学习的思想来理解每个单词在句中的上下文关系, 生成词的嵌入值. 这一步骤获得了文本的序列特征和语义表示, 得到了文本的向量字符表示$C=(C_C,C_1,C_2,\cdots,C_n,C_S)$, 其中$C_i$表示第 i 个单词的上下文特征.

3) 将$C=(C_C,C_1,C_2,\cdots,C_n,C_S)$送入 BiGRU 层, 捕捉序列中的长距离依赖关系从而获取序列最大分值, 形成输出标签$S=(S_C,S_1,S_2,\cdots,S_n,S_S)$, 其中$S_i$表示序列中每个位置 i 处的标签, 这些标签用于表示文本中每个单词或者字符是否属于某个实体类型.

4) 最后利用条件随机场模型处理标签偏差, 学习标签之间的约束关系进而确定标签为某一实体类型的概率$L=(L_C,L_1,L_2,\cdots,L_n,L_S)$, 获得最终的预测结果.

在模型中所使用的预训练语言模型 ELECTRA 主要由生成器 (Generator) 和判别器 (Discriminator) 两部分组成. 其中, 生成器是一个与 BERT 相同的掩码语言模型. 在训练过程中, 生成器的训练方式是将输入序列$T=(X_1,X_2,\cdots,X_n)$进行掩码, 得到包含被掩盖 token 的预测序列之后进行恢复, 得到X^{corrupt}. 具体来说, 输入序列$T=(X_1,X_2,\cdots,X_n)$在经过编码之后可以得到上下文向量表示序列. 而对于给定的位置 t (仅限定于位置X_t上的字符是被掩码时), 生成器通过 *Softmax* 层输出X_t的概率来恢复或者生成原始的 token, 以此来学习输入序列中被掩盖 token 的上文语义. 生成器在学习过程中使用负对数似然函数作为损失函数.

对于判别器来说, 其将经过生成器恢复后的输出序列当作输入序列, 对于每一个 token, 判别器需要进

行预测每个位置的 token 是否与原文一致. 具体来说, 对于恢复序列X^{corrupt}, 判别器通过二元交叉熵损失函数, 将其与原始输入序列X进行对比判断, 区分输入序列中的每个 token 是原始的还是被替换的, 学习不同位置上的 token 在序列中的含义.

ELECTRA 的最终损失函数是由生成器和判别器的损失一同组成. 在预训练阶段, 生成器和判别器共享嵌入层, 并同时训练, 旨在让模型学会理解上下文和语义, 以及判断替换的 token 是否合理. 在微调阶段, 生成器和判别器二者继续合作. 通过这样的过程, 模型能够更好地理解和处理非结构化文本的 NER 任务.

2.2 基于 Neo4j 的疫苗接种知识图谱构建

2.2.1 本体设计

本体设计是构建知识图谱的基础环节之一. 为获得良好得可视化效果, 本体设计需具有良好的可解释性. 本研究根据疫苗接种领域的需求, 参考相关标准和数据集, 确定了构建本体的目标和范围, 然后在 Protégé 上完成了对疫苗接种本体的构建工作.

目前疫苗接种领域尚未出现公开的特定本体模式, 但存在较多的涉及疫苗管理的数据集可供参考. 本文参照 2017 年实施的中华人民共和国卫生行业标准 WS 375.19-2016《疾病控制基本数据集第 19 部分: 疫苗管理》对实体核心类、类之间的关系和属性进行定义. 对于实体类别的定义, 采取单一化的值类型, 例如诊断的内容属于疾病类型, 由医学专业名词术语构成, 命名具有统一性、规范性. 然而实体属性大多由各类型的实体及非专业名词术语构成, 表现为口语化描述性语句构成的长段文本, 难以进行枚举. 例如家族史的描述可能包括“奶奶白细胞偏低”或者“父亲有家族遗传哮喘”.

在本体构建时, 本文使用了当下流行的本体开发工具之一 Protégé. 该工具提供了关于概念类、关系和属性等实例的构建方式, 便于用户构建和管理本体. 本研究关注的实体主要为 7 个核心大类: 患者、手术、药物、查体、诊断、暂缓疫苗、接种建议. 根据确定的实体, 定义了 6 种语义关系: 所做手术、所做诊断、所做查体、服用药物、接种建议、暂缓疫苗. 此外, 需要定义患者类的属性. 对于定义好的本体结构, 通过 Ontograf 进行可视化, 并在图 2 中进行展示.

2.2.2 Neo4j 存储与图谱三元组表示

Neo4j 数据库是一种高性能的、可拓展图形数据库, 为知识图谱的存储、管理和可视化提供了技术基础[22]. Neo4j 使用节点和关系的形式来组织和存储数据, 其中节点表示实体, 关系表示实体之间的连接和关联. 用户可以在节点和关系上定义属性和标签, 以更好地描述实体和关系的特征. Neo4j 还提供了拥有图查询语言 Cypher, 它是一种面向图分析的声明式查询语言, 专门用于在检索、操作和管理图数据库中的数据, 执行复杂的图查询和遍历操作.

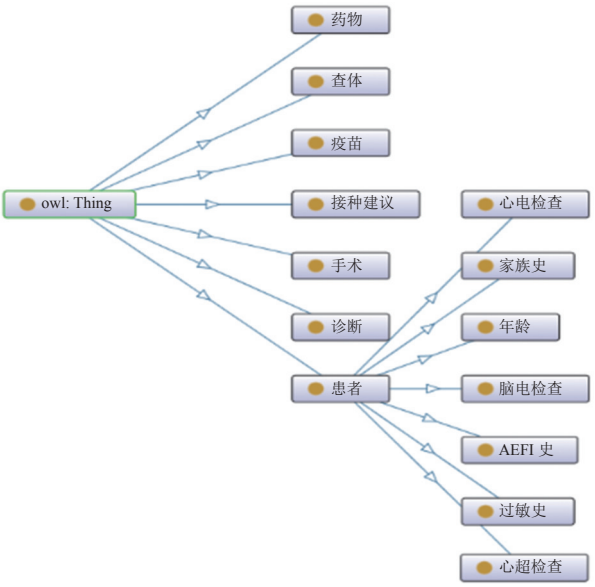

图 2 疫苗接种本体

广义的知识图谱可表示为$G=\{E,R,V,A,T_{RE},T_{VA}\}$, 其中$E$代表知识图谱中的实体集合, R代表图谱中的关系, V代表属性值, A代表节点属性. $T_{RE}=(E,R,E)$表示为实体和实体间的关系三元组, $T_{VA}=(E,V,A)$则表示实体节点和属性值的三元组. 本研究所构建的儿童疫苗接种知识图谱由上述两种三元组构成. 为了方便表示, 将知识图谱表示为$G=<H_k,R_k,T_k>$, $k\in[1,n]$, $n=|G|$, 在这个定义中, G代表整个图谱, H_k代表图谱中第 k 个三元组的头实体, T_k代表图谱中第 k 个三元组的尾实体, R_k代表H_k和T_k之间的关系. 通过对每个三元组头实体和尾实体的组合, 可以建立一个完整的知识图谱.

在实际的儿童疫苗接种工作中, 医生需要综合考虑疾病儿童的各项检查、过敏史、手术情况等健康状况, 以确定儿童能否接种疫苗. 因此在疫苗接种推荐过程中, 将患者节点考虑为中心节点, 患者的年龄和过敏史等实体作为节点的属性.

对于第 2.2.1 节抽取的 10 类实体, 对其进行后处

理, 规定如下: 对于药物实体, 使用药物字典表将药物名称统一规范. 若诊断实体上下文出现亲属称谓词表中的称谓, 则将该诊断实体映射至图谱三元组中的家族史节点. 将过敏症状与过敏源合并为图谱中的过敏史节点. 对于辅助检查实体, 划分为脑电检查、心电检查、心超检查, 分别映射到患者节点的属性中. 在表 1 中给出这了三元组实体关系的定义, 其中患者拥有姓名编号、脑电检查、心电检查、心超检查、过敏史、家族史、年龄和 AEFI 史属性.

表 1 实体关系定义

序号	实体A	实体B	实体关系
1	患者	诊断	所做诊断
2	患者	查体	所做查体
3	患者	手术	所做手术
4	患者	药物	所用药物
5	患者	暂缓疫苗	暂缓疫苗
6	患者	接种建议	接种建议

2.3 基于显著性类别指导的接种建议推荐工作

本文构建了以患者为中心节点的异构知识图谱, 患者间可能存在部分共享实体, 但实体之间并不必然存在联系. 在实际诊疗过程中, 接种建议的推荐并非依赖于某一特定实体进行决策, 而是需要医生综合考虑儿童当前的身体状况. 这也代表着在查询过程中, 需要考虑各个实体对于接种建议的影响. 为了能够综合考虑各类实体, 需要在图谱中查找相似病历的接种建议, 即在图谱中查找最相似的患者子图, 以达到辅助诊断的目的.

通过利用第 2.1 节的自动化抽取方法, 可以得到医学实体数据, 然后采用三元组的形式进行存储, 进而将患者信息转为患者子图. 尽管这种方法能简明扼要地表示一份电子病历的关键信息, 但也不可避免地导致部分信息的丢失. 原电子病历文本可能包含一些潜在有用信息, 例如主诉文本存在对于病情口语化、细节化的大量描述, 这些信息对医生的诊断决策具有一定的参考价值. 然而, 如果直接将这些信息输入模型进行学习, 会引入过多的噪声, 因此不能将电子病历文本直接送入模型进行处理. 为此, 本文在参考 Wang 等人[23]提出的显著性类别指导方法上进行了改进, 来实现接种建议的推荐. 显著性即代表了一个句子对文档中心思想的贡献程度, 其分配意味着显著性在文档中所有句子的分布情况. 本文同时考虑患者子图信息及该患者原本未经过处理的电子病历文本信息, 利用预训练语言模型, 训练了基于显著性类别指导的接种建议分类推荐模型. 该模型的训练过程如图 3 所示.

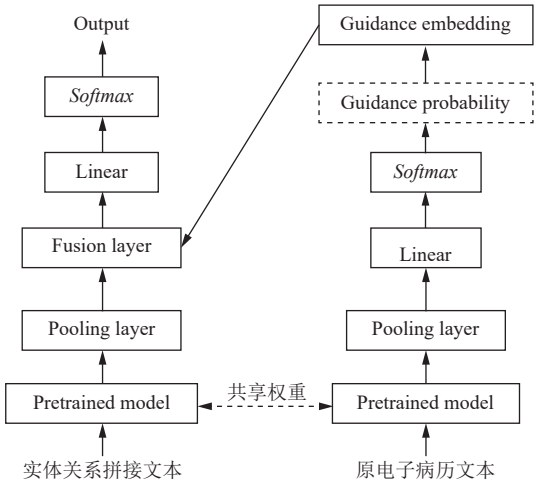

图 3 疫苗接种显著性类别指导方法结构图

具体实验步骤如下.

1) 在知识图谱中使用 Cypher 语言的 match 语句遍历每个患者和其拥有的所有关系及对应实体. 将患者所有的属性值、连接的实体和关系进行拼接, 形成关键信息文本库, 即本文的输入文本库, 输入文本使用式 (1) 中的$Text_{\text{ent}}$来表示.

$$E_{\text{ent}} = Pooling(Pretrained(Text_{\text{ent}})) \tag{1}$$

2) 利用预训练语言模型学习输入文本, 得到式 (1) 中的E_{ent}, 同时将原始电子病历文本$Text_{\text{com}}$输入到预训练语言模型中, 经过池化层处理后得到文本的编码, 即式 (2) 中的E_{com}.

$$E_{\text{com}} = Pooling(Pretrained(Text_{\text{com}})) \tag{2}$$

3) 根据式 (3), 通过线性层和 *Softmax* 层学习到电子病历文本属于某个类别的概率$Guid_{\text{probability}}$.

$$Guid_{\text{probability}} = Softmax(W^{\text{T}} \cdot E_{\text{com}}) \tag{3}$$

4) 初始化表中每个类别的编码向量, 并将得到的概率与类别编码向量相乘后, 得到类别指导向量 E_{guid}, 该向量包含了电子病历文本属于不同文本类别的一些关键性的信息. 此外, 式 (4) 中的$P(class = l \mid x)$代表电子病历文本属于哪一类的接种建议的概率, 即$Guid_{\text{probability}}$.

$$E_{\text{guid}} = \sum_{l=1}^{l} E(class = l) P(class = l \mid x)) \tag{4}$$

5) 将步骤 4) 的结果与子图查询得到关键信息的文本编码结果通过一个融合层 (fusion layer) 进行融合, 参照 Gong 等人[24]提出的融合方法, 具体见式 (5). 式 (6) 则为最终的输出公式.

$$E_{\text{fusion}} = Concat(E_{\text{ent}}, E_{\text{guid}}, E_{\text{ent}} \odot E_{\text{guid}}, E_{\text{ent}} - E_{\text{guid}}) \quad (5)$$

$$Output = Softmax(V^{\text{T}} \cdot E_{\text{fusion}}) \quad (6)$$

3 实验分析

3.1 数据集

本研究使用的数据来源于自东部某三甲医院儿童免疫接种咨询门诊电子病历数据集系统. 采集的免疫接种咨询门诊电子病历数据共 4 000 份. 去除 174 份接种建议空缺的病历数据, 剩余 3 826 份电子病历, 每一份病历中包含若干描述性的语句. 门诊电子病历内容包含门诊日期、性别、主诉、既往史、现病史、家族史、诊断、处理建议、查体以及各种辅助检查报告等内容. 收集并初筛之后的就诊数据需要进行进一步清洗, 在去除不合法字符、重复字符以及冗余字段数据后, 其中一条病历的内容为: 主诉: 免疫接种咨询. 患儿, 21 月龄, 因“癫痫”前来免疫接种咨询. 2019.09.30–2019.10.10 因“1. 癫痫; 2. 先天性甲状腺功能减退; 3. 高胆红素血症; 4. 鹅口疮”住院治疗. 2019 年 11 月患儿因抽搐于我院就诊, 诊断为癫痫, 给妥泰等药物口服治疗, 2020 年 4 月 26 日和 5 月 2、17、18 日, 之后未再出现抽搐. 目前德巴金 (一次 3 mL , 一天两次) 及妥泰 (一次一片, 一天两次) 口服治疗中, 优甲乐 (晨起顿服 1/2 片, 一天一次) 口服中. 基因检测为 PACS2 基因一个变异, 早发性婴儿癫痫性脑病 66 型. 查体: 神清, 精神可, 心肺听诊无殊, 腹软, 无压痛, 肝脾肋下未及肿大, 肌张力可. 诊断: 癫痫, 过敏症状: 湿疹, 过敏源为鸡蛋. 数据集采用 BIO 标注方式进行标注, 即将每个元素标注为“B-X”“I-X”“O”, 其中 B-X 代表 X 实体的起始位置, I-X 表示 X 实体起始位置之后的所有词, O 表示该元素不属于任何标注的实体类型.

3.2 实验评价标准和环境配置

对于模型的评价, 本研究使用准确率 (*Accuracy*)、精确率 (*Precision*)、召回率 (*Recall*) 和 *F*1 分数 (*F*1-score) 作为评价标准. 在式 (7) 中, *Accuracy* 是指分类正确的样本占总样本个数的比例. 如式 (8) 所示, *Precision* 指在分类正确的正样本个数占判定为正样本的样本个数的比例. *Recall* 则是指分类正确的正样本个数占真正的正样本个数的比例, 其定义如式 (9) 所示. 在式 (10) 中, *F*1-score 通过对 *Precision* 和 *Recall* 加权计算得出, 用以综合反应模型的性能.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1\text{-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$

其中, *TP* 为模型正确地将正类样本预测为正类别的数量, *TN* 为模型正确地将负类样本预测为负类别的数量, *FN* 为模型错误地将正类样本预测为负类别的数量, *FP* 为模型错误地将负类样本预测为正类样本的数量. 此外, 研究的实验环境配置参数如表 2 所示.

表 2 实验环境配置

相关配置	配置参数
操作系统	Ubuntu 18.04.4 LTS
GPU	NVIDIA Corporation GV100 [TITAN] (rev a1)
显存大小	12 GB
Python	3.8.0
PyTorch	1.12.1
Cuda	11.2

3.3 实体识别实验

实体识别实验选取 1 000 位儿童的电子病历, 在医生的指导下, 对数据进行标注. 最终用于模型训练的语料库为 3 972 份, 其中 80% 作为训练集, 剩余的 20% 中的 50% 作为测试集, 50% 作为验证集. 基于 PyTorch 实现了相关模型, 在文本输入阶段, 句子最大截断长度设置为 256, 每次训练的样本数为 16. 词向量表示阶段, 采用的预训练模型为在中文语料库上进行预训练的 base 版本的 ELECTRA 模型, 默认使用 12 个注意力头的 Transformer, 取训练语言模型 12 层, 每层包含 768 个隐藏层. 模型训练中, 将学习率设置为 5E–5, 丢失率设置为 0.1, 训练轮次设置为 20.

实验采用 BERT-CRF、BERT-BiLSTM-CRF、RoBERTa-BiGRU-CRF 与本文所提出的模型进行对比, 其对比结果如表 3 所示. 由表 3 可知, 在模型整体层面, ELECTRA-BiGRU-CRF 模型相较于对比模型 BERT-CRF 的 *Precison*、*Recall* 和 *F*1-score 分别提升了 9.25%、8.95% 和 9.13%. 造成这一结果的原因是 ELECTRA

和 BERT 的区别, 即在掩码处理方面, BERT 从每个样本随机选取 15% 的 token 进行掩盖, 其中 80% 被真实 token 替换, 而在预测阶段, 并非对所有被掩盖的 token 及逆行预测, 而只是针对其中一小部分进行预测. 这么做会造成预训练与微调时的不匹配. 相比之下, ELECTRA 在预测任务中, 判别器从所有输入的 token 中学习, 避免了 BERT 中的不匹配问题.

表 3 模型对比结果表 (%)

模型	*Precision*	*Recall*	*F*1-score
BERT-CRF	80.99	86.78	83.78
BERT-BiLSTM-CRF	84.26	88.56	86.36
RoBERTa-BiGRU-CRF	86.81	89.52	88.14
ELECTRA-CRF	87.55	95.47	91.34
ELECTRA-BiGRU-CRF	90.24	95.73	92.91

相较于 BERT-BiLSTM-CRF 的 *Precision*、*Recall* 和 *F*1-score 分别提升了 5.98%、7.17% 和 6.55%. 相较于 RoBERTa-BiGRU-CRF 的 *Precision*、*Recall* 和 *F*1-score 分别提升了 3.43%、6.21% 和 4.77%. 由此可见, 本研究所提出模型的学习能力优于对比模型, 具有较优的识别性能. 在 10 类实体识别层面, 对模型结果进行分析. 如表 4 所示, 以相对来说表现最好的 RoBERTa-BiGRU-CRF 为例, 与本文提出的 ELECTRA-BiGRU-CRF 进行对比, 分析结果发现: 在模型整体层面, ELECTRA-BiGRU-CRF 较 RoBERTa-BiGRU-CRF 的提升幅度并不显著, 但在每一类的实体识别中, RoBERTa-BiGRU-CRF 的性能表现得不如 ELECTRA-BiGRU-CRF 稳定, 尤其在年龄、手术、AEFI 史的识别方面表现较差. 总的来说, 在儿童免疫接种咨询门诊电子病历实体识别任务中, ELECTRA-BiGRU-CRF 模型更有优势.

表 4 RoBERTa-BiGRU-CRF 与 ELECTRA-BiGRU-CRF 模型识别实体效果对比 (%)

类目	RoBERTa-BiGRU-CRF			ELECTRA-BiGRU-CRF		
	Precision	*Recall*	*F*1-score	*Precision*	*Recall*	*F*1-score
年龄	62.96	65.38	64.15	63.88	100.00	77.96
诊断	92.69	94.82	93.75	99.11	95.93	97.49
过敏症状	95.65	95.65	95.65	97.46	100.00	98.71
过敏原	95.83	100.00	97.87	100.00	97.00	98.48
查体	99.00	98.28	99.14	99.86	99.86	99.86
辅助检查	73.20	77.37	75.23	98.86	99.06	98.96
AEFI史	11.53	18.75	14.26	57.03	97.21	71.88
手术	63.63	80.00	70.88	94.72	90.98	92.81
药物	88.71	85.93	87.30	95.28	99.90	97.54
疫苗	96.55	87.50	91.80	98.38	92.94	95.58

如表 5 所示, 通过实验发现, 在 ELECTRA-CRF 模型中引入 BiGRU 模块, 能够增强序列建模和上下文理解能力, 模型的性能会更为稳定. 进一步分析发现, 在未引入 BiGRU 模块的情况下, 部分实体识别性能不稳定, 例如“过敏原”实体的识别表现出色, 但“过敏症状”识别的准确率却不理想.

表 5 ELECTRA-CRF 与 ELECTRA-BiGRU-CRF 模型识别实体效果对比 (%)

类目	ELECTRA-CRF			ELECTRA-BiGRU-CRF		
	Precision	*Recall*	*F*1-score	*Precision*	*Recall*	*F*1-score
年龄	95.00	100.00	97.44	63.88	100.00	77.96
诊断	97.11	98.33	97.72	99.11	95.93	97.49
过敏症状	26.27	100.00	42.11	97.46	100.00	98.71
过敏原	100.00	100.00	100.00	100.00	97.00	98.48
查体	99.41	99.70	99.55	99.86	99.86	99.86
辅助检查	98.46	98.78	98.62	98.86	99.06	98.96
AEFI史	72.28	91.40	80.58	57.03	97.21	71.88
手术	98.27	99.61	98.93	94.72	90.98	92.81
药物	94.85	96.34	95.58	95.28	99.90	97.54
疫苗	92.11	95.19	93.62	98.38	92.94	95.58

总体而言, ELECTRA-BiGRU-CRF 在学习电子病历文本数据的特征信息方面是有效的. 具体而言, 对于“诊断”“过敏症状”“过敏原”“查体”“辅助检查”“药物”“疫苗”这 8 类实体识别的 *F*1-score 均在 95% 以上. 然

而“年龄”和“AEFI 史”的识别效果欠佳. 分析可能的原因有两点.

1) 在儿童免疫接种咨询门诊中, 儿童的年龄记录精度通常高于成年人, 精确到月份, 例如 5 月龄、1 岁 2 月龄. 这导致了在病历中, 儿童的年龄会与病历描述中记录的既往病史和部分检查检验的年月产生重叠和歧义, 这会导致模型对实体的边界识别困难, 无法准确区分在当前语境下实体的类型.

2) 对于 AEFI 史, 由于疫苗接种反应记录大多为医生手动填写, 因此存在相关描述语言风格存在不统一的情况, 且该实体中关于症状实体口语化描述多, 会一定程度上造成“AEFI 史”实体识别效果欠佳的问题. 此外, 实验的数据集中, “AEFI 史”的样本量相对其他实体较少, 这也会一定程度上对模型学习造成影响.

综上所述, ELECTRA-BiGRU-CRF 相较于其他对比模型而言, 无论是在模型整体表面层面, 还是单个实体识别的层面, 都展现出了自身的优势.

3.4 疫苗接种知识图谱存储与可视化

本文选用 Neo4j 图数据库对儿童疫苗接种知识图谱进行存储. 在图数据库存储中, 图的顶点表示知识图谱中的实体, 关系使用顶点与顶点之间的边进行表示.

Neo4j 图数据库支持多种方式进行数据导入以满足不同场景和需求. 在本研究中, 对电子病历数据进行处理后, 形成实体和关系的 JSON 文件, 通过 py2neo 工具包对实体识别后的数据进行批量导入, 使用 Cypher 语言进行节点的增、删、改、查等操作.

本研究基于 3 826 份电子病历, 构建了中文儿童疫苗接种知识图谱. 该图谱共有 6 569 个顶点, 实体间共有 26 336 种关系. 在图 4 中对知识图谱的部分内容进行了可视化展示, 其中不同的颜色代表不同的节点类型, 比如紫色代表患者节点, 红色代表接种建议, 蓝色代表诊断.

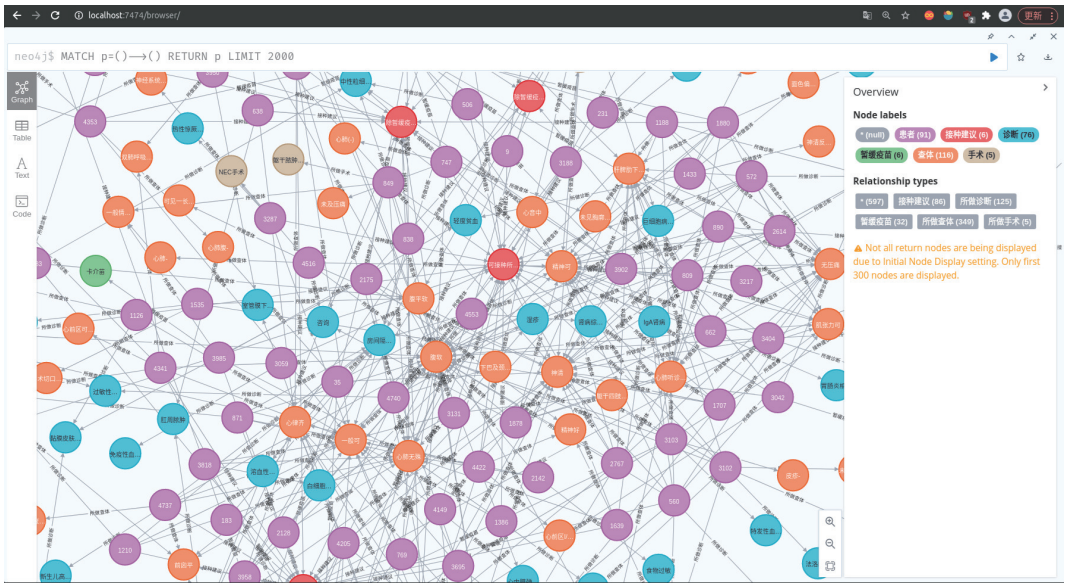

图 4 基于 Neo4j 的图谱可视化

图 5 主要是对患者节点属性进行可视化展示, 其中患者节点包含了 8 种属性, 使用患者姓名编号作为患者的唯一标识.

3.5 基于显著性类别指导的接种建议推荐实验

实验选取 3 826 份电子病中的 80% 作为训练集, 剩余的 20% 中的 50% 作为测试集, 50% 作为验证集. 基于显著性类别指导实验中, 基于 PyTorch 实现了相关模型, 微调时模型输入的句子最大截断长度设置为 512, 每次训练的样本数为 8, 学习率设置为 3E–5, 丢失率设置为 0.1, 优化器为 AdamW. 如表 6 所示, 本研究所采用的电子病历数据集一共拥有 6 类接种建议, 经医生的指导将 1、2 两类接种建议合并, 3、4 类接种建议合并, 共计 4 大类接种建议.

在实验中, 采用了在中文医学语料库上进行预训练的语言模型, 包括 BERT[16]、DKPLM[25]、ERNIE[26] 进行了分类预测实验. 实验结果如表 7 和表 8 所示, 在 BERT 模型上, 采用显著性类别指导方法, 精确率提高了 2.58%. 在 DKPLM 模型上, 精确率提高了 2.3%. 在

BERT 和 DKPLM 上, *F*1-score 均有所提升. 而在 ERNIE 模型上, 采用显著性类别指导方法精确率也有所提升, 表现相对来说较为稳定. 总体而言, 基于显著性类别指导的方法在模型表现结果稳定的情况下, 精确率都得到了提升. 这表明通过引入原始电子病历文本信息, 并使用接种建议显著类别指导的方法帮助模型理解文本信息, 能够让模型分类准确率均得到提升, 从而提高接种建议推荐的准确性.

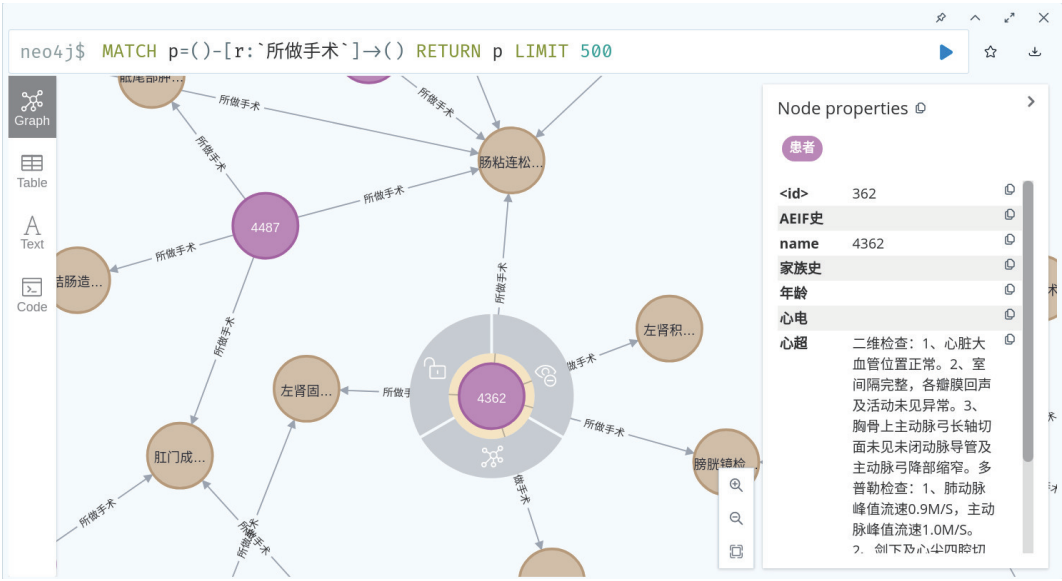

图 5 患者节点属性可视化

表 6 接种建议类型

编号	接种建议	数目
1	可接种所有疫苗, 风险同正常儿一致	1 629
2	可接种所有疫苗, 每次接种一剂次	693
3	除暂缓疫苗外, 其余疫苗可正常接种	406
4	除暂缓疫苗外, 其余疫苗可正常接种, 每次接种一剂次	161
5	暂缓所有疫苗接种	214
6	其他	723

表 7 基于预训练语言模型的推荐结果 (%)

模型	*Accuracy*	*Precision*	*Recall*	*F*1-score
BERT	75.99	72.71	77.65	74.60
DKPLM	77.66	74.95	77.49	76.01
ERNIE	77.03	75.07	75.34	75.18

表 8 基于显著性类别指导的推荐结果 (%)

模型	*Accuracy*	*Precision*	*Recall*	*F*1-score
BERT_guid	77.45	75.29	74.57	74.74
DKPLM_guid	79.33	77.25	77.62	76.74
ERNIE_guid	77.45	75.80	74.57	75.07

4 结论与展望

本文基于儿童疫苗接种咨询门诊电子病历数据构建知识图谱, 实现了儿童疫苗接种建议推荐工作. 为实现知识图谱的构建, 本文提出了 ELECTRA-BiGRU-CRF 实体抽取方法, 该方法相比于其他基线模型, 在疫苗接种领域具有更好的识别性能, 为后续基于知识图谱的疫苗接种推荐工作奠定基础. 在疫苗接种推荐工作中, 本研究采用接种建议显著性类别指导的思想, 让语言模型在学习图谱存储的患者信息的同时, 考虑原电子病历文本的类别信息, 以提升接种建议推荐的学习效果. 该方法在 3 种不同的模型上进行实验, 实验结果表明分类推荐准确率均得到了提升. 本文提出的方法能够更有效地完成知识图谱构建以及接种建议分类推荐工作.

医学知识图谱的构建及其下游任务的应用一直以来都是复杂且严谨的工作. 在后续工作中, 将引入更多罕见病患儿的病历数据, 研究数据完整性和规范性方案, 以平衡数据类别分布, 并细化数据类别, 进而提高接种建议分类推荐模型在准确率上的表现. 此外, 本研究未来将探索图神经网络在异构知识图谱查询方法中的应用.

参考文献

1 王富云, 刘桂红. 特殊健康儿童预防接种评估及不良反应处理. 广州医药, 2023, 54(10): 46–51.
2 陈福星, 潘雪娇, 梁辉, 等. 浙江省 2015–2019 年儿童型三价灭活流感病毒裂解疫苗预防接种不良反应报告发生率.

中国疫苗和免疫, 2022, 28(4): 436–439. [doi: 10.19914/j.CJVI.2022084]
3 宋祎凡, 曾淇民, 余文周, 等. 我国儿童监护人免疫规划疫苗犹豫现状及影响因素分析. 中国预防医学杂志, 2023, 24(12): 1303–1308. [doi: 10.16506/j.1009-6639.2023.12.006]
4 Stassen P, Westerman D. Novice doctors in the emergency department: A scoping review. Cureus, 2022, 14(6): e26245. [doi: 10.7759/cureus.26245]
5 王彩云, 郑增亮, 蔡晓琼, 等. 知识图谱在医学领域的应用综述. 生物医学工程学杂志, 2023, 40(5): 1040–1044.
6 王丹蕾, 邱五七, 都率. 疫苗接种风险沟通中的信任问题与建议. 卫生软科学, 2023, 37(4): 82–86. [doi: 10.3969/j.issn.1003-2800.2023.04.020]
7 胡丹丹. 神经系统疾病儿童免疫接种策略. 临床儿科杂志, 2022, 40(3): 184–188. [doi: 10.12372/jcp.2022.21e0028]
8 Wang B, Shao XQ, Wang D, *et al*. Vaccinations and risk of systemic lupus erythematosus and rheumatoid arthritis: A systematic review and meta-analysis. Autoimmunity Reviews, 2017, 16(7): 756–765. [doi: 10.1016/j.autrev.2017.05.012]
9 Zhang XM, Meng MM, Sun XL, *et al*. FactQA: Question answering over domain knowledge graph based on two-level query expansion. Data Technologies and Applications, 2020, 54(1): 34–63. [doi: 10.1108/DTA-02-2019-0029]
10 Chandak P, Huang KX, Zitnik M. Building a knowledge graph to enable precision medicine. Bioinformatics, 2023, 10(1): 67. [doi: 10.1038/s41597-023-01960-3]
11 Joshi P, Masilamani V, Mukherjee A. A knowledge graph embedding based approach to predict the adverse drug reactions using a deep neural network. Journal of Biomedical Informatics, 2022, 132: 104122. [doi: 10.1016/j.jbi.2022.104122]
12 Tao XH, Pham T, Zhang J, *et al*. Mining health knowledge graph for health risk prediction. World Wide Web, 2020, 23(4): 2341–2362. [doi: 10.1007/s11280-020-00810-1]
13 An B. Construction and application of Chinese breast cancer knowledge graph based on multi-source heterogeneous data. Mathematical Biosciences and Engineering, 2023, 20(4): 6776–6799. [doi: 10.3934/mbe.2023292]
14 Liu FF, Liu MT, Li MT, *et al*. Automatic knowledge extraction from Chinese electronic medical records and rheumatoid arthritis knowledge graph construction. Quantitative Imaging in Medicine and Surgery, 2023, 13(6): 3873–3890. [doi: 10.21037/qims-22-1158]
15 Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 2020, 30(4): 681–694. [doi: 10.1007/s11023-020-09548-1]
16 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.
17 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
18 Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016. 473–482. [doi: 10.18653/v1/N16-1056]
19 Wang Q, Zhou YM, Ruan T, *et al*. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. Journal of Biomedical Informatics, 2019, 92: 103133. [doi: 10.1016/j.jbi.2019.103133]
20 Dai ZJ, Wang XT, Ni P, *et al*. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. Proceedings of the 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. Suzhou: IEEE, 2019. 1–5. [doi: 10.1109/CISP-BMEI48845.2019.8965823]
21 Clark K, Luong MT, Le QV, *et al*. ELECTRA: Pre-training text encoders as discriminators rather than generators. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.
22 Quan XP, Cai WJ, Xi CH, *et al*. AIMedGraph: A comprehensive multi-relational knowledge graph for precision medicine. Database, 2023, 2023: baad006. [doi: 10.1093/database/baad006]
23 Wang F, Song KQ, Zhang HM, *et al*. Salience allocation as guidance for abstractive summarization. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi: Association for Computational Linguistics, 2022. 6094–6106.
24 Gong CG, Yu JF, Xia R. Unified feature and instance based domain adaptation for aspect-based sentiment analysis. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 7035–7045. [doi: 10.18653/v1/2020.emnlp-main.572]
25 Zhang TL, Wang CY, Hu N, *et al*. DKPLM: Decomposable knowledge-enhanced pre-trained language model for natural language understanding. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI, 2022. 11703–11711. [doi: 10.1609/aaai.v36i10.21425]
26 Wang Q, Dai ST, Xu BF, *et al*. Building chinese biomedical language models via multi-level text discrimination. arXiv: 2110.07244, 2022.

(校对责编: 孙君艳)