面向在线健康社区UGC的医疗健康知识图谱构建研究*
——以小儿腹泻病为例

孟秋晴[1] 郑铭瑞[1] 田玥璐[1] 刘逸品[1] 王琼弟[2]
（1. 贵州财经大学信息学院，贵阳 550025；2. 南京大学软件学院，南京 210008）

摘要：构建面向在线健康社区用户生成内容（User Generated Content，UGC）数据的医疗健康知识图谱，探究基于用户潜在需求的健康知识抽取，对优化在线健康社区信息组织与检索，支撑在线健康社区知识服务创新具有重要意义。提出基于在线健康社区UGC数据的实体识别组合模型LDA-BERT-BiLSTM-CRF，首先利用LDA主题模型对在线健康社区UGC数据进行主题聚类分析从而提取实体类型，基于细分实体类型利用BERT-BiLSTM-CRF模型进行命名实体识别；然后采用MC-BERT-CasRel模型抽取在线健康社区UGC数据中的重叠三元组，并通过SBERT模型实现实体对齐；最后利用Neo4j图数据库完成知识图谱的存储和可视化。以小儿腹泻病为例，基于所提方法最终构建包含939个实体和3 224个关系的小儿腹泻病知识图谱。与目前主流模型进行对比实验，结果表明，所采用的组合模型LDA-BERT-BiLSTM-CRF与关系抽取模型MC-BERT-CasRel较传统方法知识抽取更准确，实体分类也更具针对性。

关键词：知识图谱构建；在线健康社区；用户生成内容；LDA；知识抽取
中图分类号：G250.73 DOI：10.3772/j.issn.1673-2286.2024.08.002
引文格式：孟秋晴，郑铭瑞，田玥璐，等. 面向在线健康社区UGC的医疗健康知识图谱构建研究：以小儿腹泻病为例[J]. 数字图书馆论坛，2024，20（8）：9-18.

在“互联网+医疗健康”背景下，在线健康社区逐渐成为民众获取医疗健康信息的主要渠道，为用户提供了疾病知识检索、健康问答和在线问诊等多种形式的信息服务，也因此积累了海量医患交互信息。医患交互信息作为在线健康社区中的用户生成内容（User Generated Content，UGC）[1]，受到了相关领域学者的广泛关注。在线健康社区UGC作为网络健康信息资源的重要部分，蕴含了用户所关注的丰富的医疗健康知识。但不同于传统健康信息资源，在线健康社区UGC的海量及碎片化特征加大了对其进行知识组织的难度[2]。在医疗健康领域，基于知识图谱的应用研究蓬勃发展，研究对象从过去的电子病历和医学文献等结构化和半结构化数据，逐步延伸到在线健康社区医患交互信息等非结构化数据。对在线健康社区UGC数据进行知识抽取，构建医疗健康知识图谱，一方面，可以帮助各类在线健康社区对其中的UGC数据进行细粒度整合，尽可能地实现对自有资源的有效组织与利用；另一方面，在线健康社区UGC大多涉及用户最关心的健康话题，能够体现用户潜在健康需求，以在线健康社区UGC为数据源进行知识图谱构建能够拓宽现有医疗健康知识图谱的知识范围，特别是能够

收稿日期：2024-04-14
*本研究得到贵州省科技厅科技计划“‘互联网+医疗’背景下基于用户特征挖掘的医疗资源推荐研究”（编号：黔科合基础-ZK[2021]一般336）、贵州省教育厅青年科技人才成长项目“基于知识图谱的在线医疗社区信息推荐研究”（编号：黔教合KY字[2022]192号）资助。

从用户视角挖掘其关心的健康知识，为后续的健康信息检索及个性化信息服务奠定坚实基础。

在线健康社区医患问答文本中包含了大量医疗健康知识，具有专业性、多样性、实时性等特点。抽取并整合这类非结构化数据中有价值的医疗健康信息，并构建知识图谱，能够为患者提供更便捷的知识查询和获取途径，有利于提高医疗健康信息服务的质量和效率。然而，在线健康社区UGC存在着表述随意、主观性强和个体性差异等问题，准确、有效地抽取UGC中的医疗健康知识成为构建在线健康社区知识图谱的挑战之一。因此，本文选取在线健康社区医患问答文本，提出基于主题聚类的知识抽取框架，通过优化命名实体识别和关系抽取技术，构建面向在线健康社区UGC的医疗健康知识图谱，并以小儿腹泻病为例，验证所提方法的可行性和有效性，以期为医疗健康领域知识图谱构建研究提供方法借鉴。

1 相关研究

医疗健康知识图谱构建主要包括医学知识表示、医学知识抽取和医学知识融合等内容。其中，医学知识抽取是构建知识图谱的核心环节，主要包括命名实体识别和关系抽取。

目前，双向长短期记忆网络（Bidirectional Long Short Term Memory Network，BiLSTM）[3]与条件随机场（Conditional Random Field，CRF）[4]模型成为命名实体识别方法中的主流模型。Zhang等[5]在基于Transformers的双向编码器表示（Bidirectional Encoder Representations from Transformers，BERT）[6]基础上加入BiLSTM和CRF，构建BERT-BiLSTM-CRF模型进行中文临床文本医学命名实体识别，显著提高了医学命名实体识别的准确率和召回率。目前的医学命名实体识别研究大多集中在深度学习方法的应用和调优上，忽视了实体划分对实体识别任务最终效果的影响。多位学者对5类医学领域实体——疾病、症状、药物、检查、治疗进行抽取[7-8]。然而，针对不同数据来源，应基于数据源特征有针对性地细化实体类型及关系类型，否则，会导致抽取出的医疗健康知识不够全面、知识抽取的准确率和召回率不高等问题。黄琼影[9]在对糖尿病社区问答文本进行实体识别时，利用词云（WordCloud）工具对5类基础实体进行了细分，该方法虽提升了实体识别效果，但实体类型划分具有较大主观性，并且无法有效验证最佳实体类型数量。

关系抽取研究方面，目前，联合学习方法抽取效果较好，可进一步细分为基于参数共享和基于序列标注方式的两类联合模型[10]，Zheng等[11]采用了基于序列标注的联合解码实现实体、关系的联合抽取。然而，采用序列模式抽取实体关系会削弱模型捕捉长距离依赖关系的能力，不能有效提取非结构化数据中的重叠三元组。重叠三元组是指在关系抽取任务中多个实体对之间存在相同关系的情况，会导致模型在预测时难以确定实体对之间的具体关系，从而增加预测的难度。2020年，Wei等[12]提出了一种基于级联二元标记的三元组抽取框架CasRel，也称层叠指针网络，用来应对重叠三元组抽取任务，并在公开数据集上进行了实验，取得了不错的效果。周俊等[13]基于RoBERTa-wwm编码改进CasRel，抽取特定领域文本的实体间关系，F1值提升到了91.86%。

综上，目前知识抽取任务中的实体类型划分主要基于主观判断，缺乏客观验证。另外，在线健康社区医患问答文本中存在大量的重叠三元组，例如，在语句“小儿腹泻可以采用思密达和妈咪爱来治疗”中，“思密达”和“妈咪爱”都与“小儿腹泻”存在“药物治疗疾病”关系，当这种同一实体在同一关系中被重复计算的情况出现，会导致传统模型的关系分类器产生混乱，从而增加关系预测的难度，且现有研究中还没有较好的可处理医疗健康文本中重叠三元组的方法。因此，针对现有问题，主要开展以下3个方面的工作。

（1）提出实体识别组合模型LDA-BERT-BiLSTM-CRF，旨在通过挖掘最佳的实体类型，提升实体识别效果。

（2）提出关系抽取模型MC-BERT-CasRel，旨在解决医疗健康文本中重叠三元组的抽取问题，提高医疗健康文本关系抽取的准确性。

（3）将在线健康社区医患问答文本作为UGC数据来源，以小儿腹泻病为例进行知识图谱构建及可视化展示，从而验证所提方法的可行性和有效性。

2 研究设计

2.1 研究框架

面向在线健康社区UGC的医疗健康知识图谱构建框架如图1所示，整体构建流程包括数据获取、LDA-BERT-BiLSTM-CRF组合模型实体识别、关系抽取、

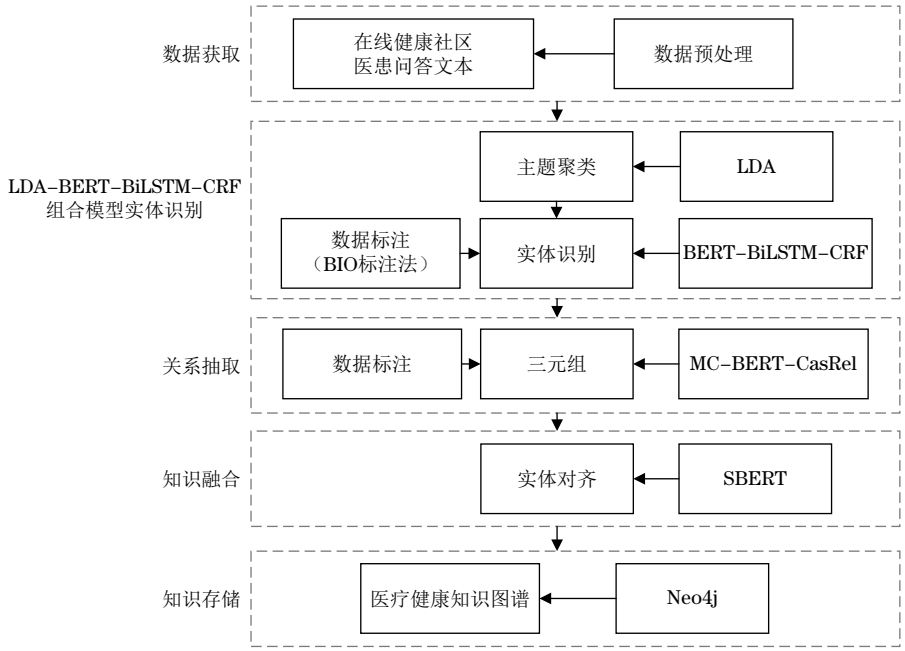

图1　面向在线健康社区UGC的医疗健康知识图谱构建框架

知识融合、知识存储5个子模块。首先采集在线健康社区医患问答文本并进行数据预处理；然后利用LDA-BERT-BiLSTM-CRF组合模型进行实体识别，利用MC-BERT-CasRel模型抽取重叠三元组，并通过SBERT模型计算词汇相似度，对数据进行实体对齐；最后通过Neo4j图数据库对知识图谱进行存储和可视化展示。

2.2　实体识别组合模型

LDA-BERT-BiLSTM-CRF组合模型架构如图2所示，由LDA主题聚类层、BERT层、BiLSTM层和CRF层组成。通过LDA主题聚类层对在线健康社区医患问答文本进行主题聚类，得到数个主题以及各个主题对应的特征词，并根据特征词归纳出各主题对应的实体类型，从而对文本进行实体标注。BERT层将经过标注的文本序列a_1、…、a_m逐词映射为向量表示，并导入BiLSTM层，利用两个相反方向的LSTM进一步捕捉序列中的前后依赖关系。CRF层利用全局特征对序列进行联合建模，更好地捕捉标签序列之间的依赖关系，确保生成的标签序列满足一定的约束条件，如BIO规则。

2.3　MC-BERT-CasRel

CasRel为级联二元标记框架，利用两级联步骤提取三元组[12]，主要包括编码层和解码层。为了在医患问答文本关系抽取任务中进一步提高CasRel编码层的语言特征表示能力，在编码层采用生物医学领域预训练语言模型MC-BERT[14]。与传统BERT掩蔽随机词汇的方法不同，MC-BERT通过掩蔽医学实体，将生物医学知识注入中国生物医学表征倾向，并在生物医学特定领域的大型语料库中进行了预训练。针对目前医疗健康文本中存在大量重叠三元组的问题，选用MC-BERT-CasRel作为关系抽取模型，试图提高已有模型抽取重叠三元组的能力。

3　小儿腹泻病知识图谱构建

3.1　实验数据

随着三孩政策的实施，公众对婴幼儿健康问题的关注度日益增加。腹泻病是婴幼儿的常见病，有关数据显示，我国5岁以下儿童腹泻病发病率为201%，平均每年每个儿童发病3.5次，死亡率为0.51%[15]，因此，小儿腹泻病的预防与诊治引起广泛关注。选取“寻医问药网”有问必答版块下的小儿腹泻病医患问答文本作为研究对象，运用Python 3.10程序爬取网页信息，采集信息包括病情描述和医生回复。共采集到4 761条数据，对其进行数据清洗，剔除医生回复为空值的8条数据和

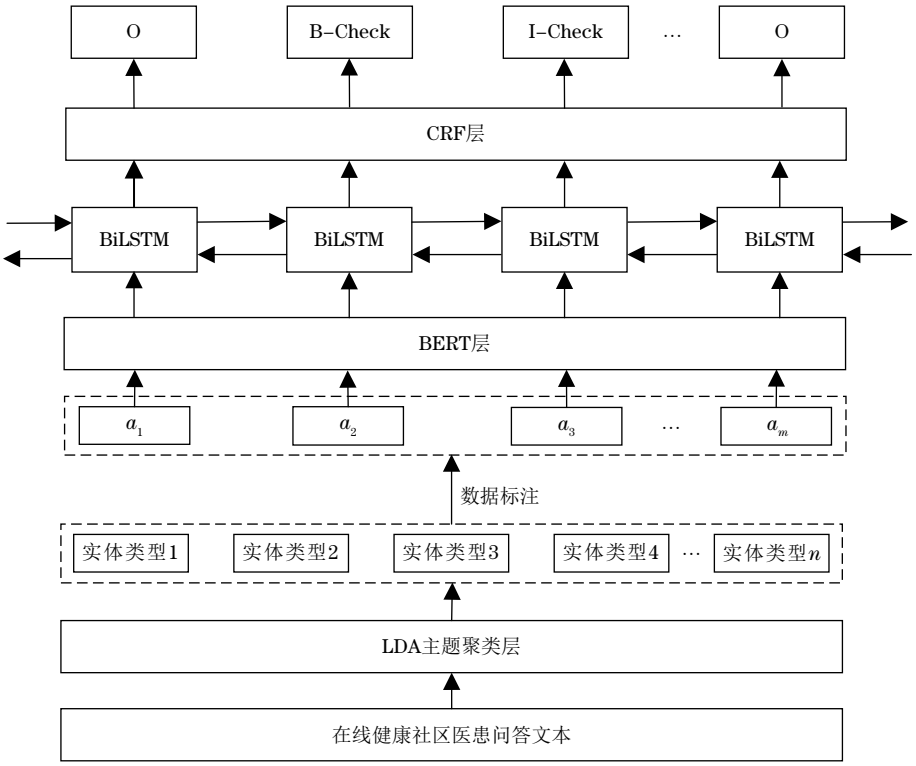

图2　LDA-BERT-BiLSTM-CRF组合模型架构

100条无关数据，共得到4 653条有效数据。由于每条数据的病情描述与医生回复高度匹配，将病情描述与医生回复进行拼接，并转化为15 411个语句，供后续知识抽取实验所用，实验数据如表1所示。

表1　实验数据示例

序号	示例
1	小儿腹泻呈水样状、喷射状、绿色怎么治
2	对于小儿腹泻呈水样状、喷射状、绿色的情况，首先要考虑的是感染性腹泻
3	在治疗方面，孩子应多喝水，保持身体水分和电解质平衡，并适当补充维生素
4	作为医生，我建议您先确定小儿是否饮食正常、是否脱水
5	如果饮食正常、没有脱水症状，可以继续观察，多让小儿喝水，饮食清淡

3.2　实体识别

根据提出的LDA-BERT-BiLSTM-CRF组合模型，首先通过LDA主题模型对小儿腹泻病医患问答文本进行主题聚类，从而得到最佳主题数量下主题对应特征词的概率分布；然后对各主题下概率排名靠前的特征词进行场景描述[16]，归纳出每个主题对应的医疗健康实体类型，从而对文本进行实体标注；最终将标注好的文本序列输入BERT-BiLSTM-CRF模型进行训练和预测，完成对小儿腹泻病医患问答文本的实体识别。

3.2.1　实体类型提取

（1）数据预处理。对小儿腹泻病医患问答文本进行数据预处理，包括去重、分词、去停用词等。在“哈工大停用词表”的基础上，结合词频统计方法，将“你好”“您好”“谢谢”等出现频次高但无实际意义的词删除，最终构建的停用词表包含1 913个词。为避免在分词过程中误判专有医学词，构建自定义词表，在词表中添加“妈咪爱”“蒙脱石散”“双歧杆菌”等词，最终构建的自定义词表包含167个词。

（2）LDA主题建模。在数据预处理的基础上，使用CountVectorizer参数对小儿腹泻病医患问答文本进行特征提取，将原始文本数据转化为向量表示，以捕获词汇的出现频率、权重等信息，供LDA主题模型进行训练和推断，并采用Scikit-learn中的LatentDirichletAllocation库构建主题模型。

在使用LDA主题模型对文本进行分析时，通常需要设置主题数量k。Blei等[17]提出的困惑度（Perplexity）

作为衡量语言模型预测性能的重要指标，已被广泛应用于判断最优主题数量，因此，采用困惑度来确定小儿腹泻病医患问答文本的最优主题数量。当k设定为10时，模型的困惑度较低，值为196.05，且一致性（Coherence）较高，值为0.473 3，主题分类效果较好，因此，最终将k设定为10。

（3）实体类型提取。实体类型提取过程与结果如表2所示。通过LDA主题分析得到小儿腹泻病医患问答文本的10个主题，以及每个主题对应的概率排名前15的特征词。结合小儿腹泻病患者在实际问诊中的语料特征，对每个主题下的高概率特征词进行场景描述[16]，总结归纳出最符合当前主题下高概率特征词的医疗健康实体类型。例如：Topic 1中，感染性、细菌、病毒、轮状病毒等特征词都代表病因，疾病、症状等特征词贴合某种病因诱发疾病或症状的场景，故将Topic 1定义为“病因”，场景描述为“病因诱发疾病”“病因导致症状”；Topic 2中，食物、生冷、油腻、刺激性等特征词都代表食物，症状、拉肚子、肠炎等特征词体现了由食物引起的某类症状或疾病，以及因疾病或症状不宜食用某类食物的情景，故将Topic 2定义为“食物”，场景描述为“食物诱发症状”“食物诱发疾病”“症状不适宜食物”“疾病不适宜食物”。根据主题聚类结果，在疾病、症状、药物、治疗和检查5类实体的基础上增加了食物、病因、部位、人群和预防措施5类实体，并增加对10类医疗健康实体的场景描述。

表2　实体类型提取过程与结果

主题	特征词	实体类型	场景描述
Topic 0	小儿、家长、腹部、妈妈、拉肚子、婴儿、症状、儿童、乳酶、助消化、肠炎、复方、患儿、口服液、肚脐	人群	人群出现疾病、人群出现症状
Topic 1	感染性、原因、疾病、方法、细菌、病毒、症状、轮状病毒、药物、婴儿、病因、小儿、肠道、儿童、患儿	病因	病因诱发疾病、病因导致症状
Topic 2	食物、饮食、生冷、油腻、消化、刺激性、症状、拉肚子、妈妈、母乳喂养、腹部、营养、调理、蔬菜、肠炎	食物	食物诱发症状、食物诱发疾病、症状不适宜食物、疾病不适宜食物
Topic 3	肠炎、疾病、炎症、化验、常规、白细胞、颗粒、抗生素、拉肚子、细菌性、小儿、复查、输液、症状、蛋白	疾病	疾病并发相关疾病、疾病诱发相关症状、药物治疗疾病、疾病适宜食物
Topic 4	拉肚子、症状、腹痛、疼痛、母乳、精神、脱水、辅食、母乳喂养、婴儿、食物、吃奶、妈妈、体重、牛奶	症状	症状适宜食物
Topic 5	指导、医生、调理、意见、母乳喂养、病情、用药、水分、补充、药物、小儿、症状、外用、疼痛、皮肤	预防措施	预防措施防止疾病、预防措施防止症状
Topic 6	补液、输液、医院、精神、水分、拉肚子、电解质、症状、轮状病毒、肠炎、急性、药物、病毒性、脱水、病毒	治疗	治疗改善症状、治疗改善疾病、治疗采用药物
Topic 7	肠道、调理、腹部、调节、菌群、胃肠道、症状、拉肚子、功能、炎症、药物、婴儿、菌群、发育、粘膜	部位	身体部位出现症状、身体部位出现疾病
Topic 8	药物、颗粒、小儿、脾胃、婴儿、中药、中成药、间隔、症状、体温、拉肚子、疫苗、活菌、庆大霉素、复方	药物	药物改善症状、药物针对病因
Topic 9	医院、常规、检查、化验、对症、原因、血常规、拉肚子、检查一下、药物、肠炎、症状、炎症、腹部、病因	检查	检查证实疾病、检查发现病因

3.2.2　命名实体识别

（1）数据标注。为了验证实体类型提取的效果，利用Doccano文本标注工具，以5类基础医疗健康实体（疾病、症状、药物、治疗、检查）和提出的10类医疗健康实体（疾病、症状、药物、治疗、检查、食物、病因、部位、人群、预防措施）分别标注1 000个相同的小儿腹泻病医患问答语句，并抽取其中的800个语句作为训练集、200个语句作为测试集。

采用BIO标注策略，按照B-X、I-X和O进行标注，其中：B代表一个实体的开始位置，I代表一个实体的内部位置，X代表具体的实体类型；B-X即当前字符是实体类型的起始部分，I-X即当前字符是实体类型的中间或结束部分，O即当前字符不属于任何实体。例如，语句{腹泻是许多病毒感染的常见症状。}，BIO标注应为{‘B-Symptom’，‘I-Symptom’，‘O’，‘O’，‘O’，‘B-Reason’，‘I-Reason’，‘I-Reason’，‘I-Reason’，‘O’，‘O’，‘O’，‘O’，‘O’，‘O’}，其中“腹泻”为实体“症状”，“病毒感染”为实体“病因”。

（2）实验对比。采用BERT-BiLSTM-CRF模型

分别对5类基础医疗健康实体标注数据和提取的10类医疗健康实体标注数据进行训练，采用准确率、召回率和F1值对实验结果进行评价。模型参数设置如表3所示。

两组实验结果如表4所示，可以看出，基于LDA主题聚类提取实体类型的实验组的准确率、召回率和F1值均高于对照组，其中，准确率提升到0.842 3，召回率提高了0.093 1，F1值提高了0.063 1。结果表明，提取实体类型的实体识别效果比未提取实体类型的实体识别效果更好。

表3　BERT-BiLSTM-CRF模型参数

模型参数	设置
模型隐藏节点	128
词向量维度	150
优化算法	Adam
Dropout	0.1
学习率	10^{-5}
Epochs	10
Batch Size	8

表4　提取实体类型实验结果对比

实体类型	BERT-BiLSTM-CRF（未提取实体类型）			LDA-BERT-BiLSTM-CRF（已提取实体类型）		
	准确率	召回率	F1值	准确率	召回率	F1值
治疗	0.644 7	0.569 8	0.604 9	0.800 0	0.790 1	0.795 0
药物	0.963 4	0.877 8	0.918 6	0.980 0	0.924 5	0.951 5
症状	0.815 5	0.736 8	0.774 2	0.858 7	0.792 6	0.824 3
疾病	0.812 0	0.742 2	0.775 5	0.830 0	0.709 4	0.765 0
检查	0.900 0	0.360 0	0.514 3	0.941 2	0.551 7	0.695 7
病因				0.811 1	0.858 8	0.834 3
预防措施				0.836 4	0.807 0	0.821 4
部位				0.640 0	0.533 3	0.581 8
人群				0.795 5	0.875 0	0.833 3
食物				0.870 5	0.846 2	0.858 2
平均值	0.814 7	0.718 1	0.763 4	0.842 3	0.811 2	0.826 5

因此，将小儿腹泻病医患问答文本的实体类型确定为10类（疾病、症状、药物、治疗、检查、食物、病因、部位、人群、预防措施）。

（3）实体识别。在15 411个语句中选取3 500个语句作为实验数据进行标注，并抽取其中的2 800个语句作为训练集、700个语句作为测试集，导入BERT-BiLSTM-CRF模型进行训练。该模型测试下的各项评价指标结果如表5所示。

从表5可以看出，准确率、召回率和F1值的平均值都超过0.800 0，表明模型实体识别的整体效果较好，其中，准确率达到了0.914 1，F1值达到了0.873 6。少数实体的个别评价指标表现一般，例如疾病和症状。这是因为在数据标注任务中，将“小儿腹泻”和“婴幼儿腹泻”等词视为疾病实体，而将“腹泻”视为一种具体的临床症状，并标注为症状实体，模型容易因为人工标注的问题而产生混淆，从而导致评价指标表现一般。并且由于涉及身体部位的语料不多，训练样本量较少，部位实体识别的各项评价指标表现不太理想。除此之外，其他实体识别的准确率、召回率和F1值表现优秀，均在0.900 0左右。因此，使用改良的模型对剩余11 911个语句进行实体抽取，共抽取出8 199个医疗健康实体。

表5　实体识别实验结果

实体类型	准确率	召回率	F1值
治疗	0.940 0	0.901 0	0.920 1
药物	0.956 3	0.891 3	0.922 6
症状	0.906 6	0.784 1	0.840 9
疾病	0.849 5	0.750 0	0.796 6
检查	0.901 4	0.901 4	0.901 4
病因	0.965 5	0.900 6	0.931 9
预防措施	0.886 2	0.825 8	0.854 9
部位	0.824 7	0.740 7	0.780 5
人群	0.954 6	0.843 6	0.853 4
食物	0.956 3	0.912 0	0.933 6
平均值	0.914 1	0.845 9	0.873 6

3.3　关系抽取

3.3.1　实体关系规则设计

综合表2所示的10个实体类型对应的场景描述，以及小儿腹泻病医患问答文本的语料特点，设计了25条小儿腹泻病医患问答文本实体关系规则，如表6所示。

表6　小儿腹泻病实体关系规则

主体	关系	对象	符号表示
人群（Crowd）	出现（Appear）	症状（Symptom）	CAS
人群（Crowd）	出现（Appear）	疾病（Disease）	CAD
病因（Reason）	诱发（Induce）	疾病（Disease）	RID
病因（Reason）	导致（Cause）	症状（Symptom）	RCS
食物（Food）	诱发（Induce）	疾病（Disease）	FID
食物（Food）	诱发（Induce）	症状（Symptom）	FIS
症状（Symptom）	适宜（Suitable）	食物（Food）	SSF
症状（Symptom）	不适宜（Not Suitable）	食物（Food）	SNF
疾病（Disease）	适宜（Suitable）	食物（Food）	DSF
疾病（Disease）	不适宜（Not Suitable）	食物（Food）	DNF
疾病（Disease）	并发相关（Concurrence）	疾病（Disease）	DCD
疾病（Disease）	诱发相关（Induce）	症状（Symptom）	DIS
预防措施（Measure）	防止（Prevent）	症状（Symptom）	MPS
预防措施（Measure）	防止（Prevent）	疾病（Disease）	MPD
预防措施（Measure）	防止（Prevent）	病因（Reason）	MPR
治疗（Treatment）	改善（Improve）	症状（Symptom）	TIS
治疗（Treatment）	改善（Improve）	疾病（Disease）	TID
治疗（Treatment）	采用（Adopt）	药物（Medicine）	TAM
部位（Body）	出现（Appear）	症状（Symptom）	BAS
部位（Body）	出现（Appear）	疾病（Disease）	BAD
药物（Medicine）	治疗（Treatment）	疾病（Disease）	MTD
药物（Medicine）	改善（Improve）	症状（Symptom）	MIS
药物（Medicine）	针对（Aim）	病因（Reason）	MAR
检查（Check）	发现（Discover）	病因（Reason）	CDR
检查（Check）	证实（Discover）	疾病（Disease）	CDD

3.3.2　模型参数设置及数据标注

采用MC-BERT-CasRel模型抽取小儿腹泻病医患问答文本的实体间关系，模型参数设置如表7所示。

从15 411个小儿腹泻病医患问答语句中选取5 600个语句进行实体间关系标注，抽取其中4 480个语句作为训练集、1 120个作为测试集。语句标注示例如表8所示。

表7　MC-BERT-CasRel模型参数

模型参数	设置
模型隐藏节点	128
词向量维度	256
优化算法	Adam
Dropout	0.1
学习率	10^{-5}
Epochs	10
Batch Size	8

3.3.3　实体间关系抽取

利用MC-BERT-CasRel模型对小儿腹泻病医患问答文本进行关系抽取，同时，与传统关系抽取模型BiLSTM-CRF进行对比实验，各项评价指标的对比结果如表9所示。

可以看出，在关系抽取实验中，MC-BERT-CasRel的准确率、召回率和F1值均高于传统关系抽取模型BiLSTM-CRF。结果表明，相比传统关系抽取模型，MC-BERT-CasRel能够更好地处理在线健康社区医患问答文本重叠三元组的问题。MC-BERT-CasRel模型下各细分关系抽取的准确率、召回率和F1值如表10所示。利用训练好的MC-BERT-CasRel模型对未经过关系抽取的数据进行三元组提取，将提取出的4 357个结果存入[‘Subject’，‘Relation’，‘Object’]关系列表。

3.4　实体对齐

在线医患问答文本中存在实体共指的现象，例如“受凉”“着凉”“受寒”都指代相同意义。这种现象可能是命名规则不同、名称简写和变体等原因造成的，会导致实验中存在大量的冗余数据，从而降低知识图谱的构建质量。为了解决这个问题，采用语义相似度模型SBERT[18]，通过微调预训练模型BERT将词汇映射到向

表8　关系抽取标注示例

序号	实验数据	头实体	尾实体	关系
1	轮状病毒引起的小儿腹泻	轮状病毒	小儿腹泻	RID
2	益生菌可以调节肠道菌群，促进肠道蠕动	调节肠道菌群	益生菌	TAM
3	消化问题可能会导致腹泻和泡沫大便	消化问题	腹泻	RCS
4	思密达适用于某些细菌感染引起的腹泻	思密达	细菌感染	MAR
5	小儿腹泻主要表现为大便稀烂	小儿腹泻	大便稀烂	DIS

表9　关系抽取实验结果对比

关系抽取模型	准确率	召回率	F1值
BiLSTM-CRF	0.670 0	0.680 0	0.660 0
MC-BERT-CasRel	0.884 6	0.898 2	0.891 4

量空间中，计算词汇嵌入向量之间的余弦相似度，捕捉词汇之间的语义关系和相似性，从而判断多个词汇表达是否对应同一个实体。首先，利用小儿腹泻病医患问答数据微调BERT模型，并且经过Pooling操作生成每个词汇固定长度的嵌入向量表示；然后，选择重要词汇，使用余弦相似度公式计算该嵌入向量与其他嵌入向量之间的相似度，通过计算得到的相似度值，可以评估两个词汇之间的语义关系和相似性，值越接近1表示两个词汇在语义上越相似；最后，通过SBERT模型计算与该词汇语义较接近的5个词汇，部分相似词汇如表11所示。

由表11可知，在训练好的SBERT模型中查询“受凉”词向量，获取和“受凉”相似度较高的5个词汇（着凉、受寒、受风、受风寒、吹风），并且将这5个词汇归类为实体“受凉”。同理，查询其他词汇并获取与该词相似度较高的词汇，进行统一归类，能够减少冗余数据，提高知识图谱的构建质量。最终对所抽取的8 199个实体和4 357个三元组进行实体对齐操作，得到939个实体与3 224个三元组。

3.5　基于Neo4j的知识存储与可视化

Neo4j是一个高性能的图数据库，其中节点表示实体，边表示实体间的语义关系，采用查询语言Cypher执行复杂的图形查询操作。将经过实体对齐的实体与实体间关系存储到Neo4j图数据库中，构建可视化的小儿腹泻病知识图谱。知识图谱中存储了10类小儿腹泻病

表10　细分关系抽取结果

关系	准确率	召回率	F1值
CAS	0.947 6	0.956 3	0.951 9
CAD	0.833 3	0.892 9	0.862 1
RID	0.857 1	0.885 2	0.871 0
RCS	0.915 9	0.945 3	0.930 4
FID	0.750 0	0.750 0	0.750 0
FIS	0.971 4	0.944 4	0.957 7
SSF	0.923 8	0.923 8	0.923 8
SNF	0.920 6	0.950 8	0.935 5
DSF	0.884 6	0.718 8	0.793 1
DNF	0.823 5	0.777 8	0.800 0
DCD	0.894 7	0.809 5	0.850 0
DIS	0.919 6	0.933 7	0.926 6
MPS	1.000 0	0.739 1	0.850 0
MPD	0.916 7	0.814 8	0.862 7
MPR	0.812 5	0.787 9	0.800 0
TIS	0.783 1	0.875 7	0.826 8
TID	0.875 0	0.777 8	0.823 5
TAM	0.807 5	0.909 6	0.855 5
BAS	0.664 2	0.54 76	0.684 7
BAD	0.675 2	0.44 32	0.527 0
MTD	0.921 1	0.648 1	0.760 9
MIS	0.884 8	0.901 2	0.893 0
MAR	0.835 1	0.843 8	0.839 4
CDR	0.804 9	0.916 7	0.857 1
CDD	1.000 0	0.705 9	0.827 6

表11　相似词汇示例

查询词汇	相似词汇	相似度	查询词汇	相似词汇	相似度	查询词汇	相似词汇	相似度
受凉	着凉	0.983 250	吃坏东西	吃错东西	0.929 164	肠道功能不良	肠道消化不良	0.949 425
	受寒	0.982 390		吃不合适	0.920 348		肠道功能不调	0.934 053
	受风	0.945 855		吃东西不对	0.918 356		肠功能不良	0.898 585
	受风寒	0.934 283		吃坏了东西	0.914 429		肠道动力不足	0.864 096
	吹风	0.824 599		吃东西不合适	0.909 426		肠消化不良	0.859 821

相关实体节点与25类实体间关系，共包含939个节点与3 224个关系。

对小儿腹泻病知识图谱进行查询，可以获取与小儿腹泻病有关联的实体，例如症状、病因、药物等实体，用户根据幼儿自身表现的症状确定病因，从而对症下药；还能够查询到宜食和忌食食物，以及预防措施等实体，从而做好对小儿腹泻病的预防和护理。利用MATCH查询语句对DNF（疾病不适宜食物）进行查询，查询结果如图3所示。由查询结果可见，小儿腹泻病不适宜的食物有冻品、变质食物、乳糖制品等。由于本研究基于在线健康社区UGC数据，该领域数据不仅包括具体食物的名称，还包括描述食物性质的词汇，例如温软、温热、辛辣等，在数据标注过程中将该类词汇归类为食物，并在小儿腹泻病知识图谱中以食物节点呈现，有助于用户在查询信息时获取到可食用或不可食用食物的特征。小儿腹泻病知识图谱还可以根据已存在的逻辑关系发现新的实体间关系，实现对隐性知识的挖掘[19]，例如通过查询小儿腹泻病及其相关联症状的宜食和忌食食物，能够定制符合病患发病机理的常规食谱，从用户角度推动了个性化医疗的发展[20]。

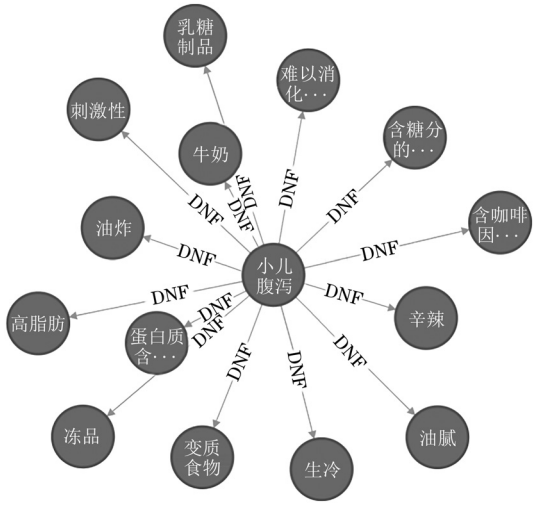

图3　小儿腹泻病患者不适宜食物

4　结语

本文针对在线健康社区中的UGC数据进行医疗健康知识图谱构建，提出了实体识别组合模型LDA-BERT-BiLSTM-CRF，旨在精准提取UGC中的医疗健康实体。首先，该组合模型有效提升了基于BERT-BiLSTM-CRF的5类基础医疗健康实体识别的准确率、召回率和F1值。其次，采用MC-BERT-CasRel模型解决在线健康社区医患问答文本中重叠三元组的抽取问题，实验结果表明，相比传统的关系抽取模型，MC-BERT-CasRel能够更好地处理重叠三元组问题。最后，通过SBERT模型对知识抽取的结果进行实体对齐，利用Neo4j图数据库构建可视化的小儿腹泻病知识图谱，为医疗健康领域知识图谱构建提供参考借鉴。通过本文提出的知识图谱构建方法，从小儿腹泻病医患问答文本中抽取健康知识，能够根据数据特征有效识别出疾病、症状、药物、治疗、检查、食物、病因、部位、人群和预防措施10个实体类型以及各实体类型之间的关系，从而获取社区用户所关注的医疗健康知识，提高医疗健康信息获取和理解的效率和质量，进而支撑后续医疗健康知识服务应用。

由于本研究仅针对小儿腹泻病文本进行实验，且所获取的实验数据量有限，包含部位等实体信息的语料略显欠缺，知识抽取结果可能不够全面。未来可在大规模数据集上进行模型训练，提升模型的抽取效果，同时可在本研究构建的知识图谱基础上利用其知识推理与检索优势，进一步开展知识推荐和智能问答等应用研究。此外，所构建的小儿腹泻病知识图谱仅基于在线健康社区UGC中的医患交互数据，未来可拓宽UGC数据来源范围，并可进一步结合医学文献、电子病历以及生物信息学数据库等，构建多源异构医疗健康知识图谱。

参考文献

[1] 陈旖旎，周晓英，岳丽欣，等. 移动UGC社区用户健康信息采纳行为意愿的影响因素[J]. 图书情报知识，2022，39（5）：82-95.

[2] 毕崇武，王冰艳，杨瑞仙，等. 基于群体认知图式的健康UGC知识标注研究[J]. 情报理论与实践，2023，46（10）：182-191.

[3] SCHUSTER M，PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing，1997，45（11）：2673-2681.

[4] LAFFERTY J，MCCALLUM A，PEREIRA F C N. Conditional random fields：probabilistic models for segmenting and labeling sequence data[C]//Proc. 18th International Conf. on Machine Learning，2001.

[5] ZHANG M Y，WANG J，ZHANG X J. Using a pre-trained language model for medical named entity extraction in Chinese clinic text[C]//2020 IEEE 10th International Conference on Electronics Information and Emergency Communication

（ICEIEC），2020：312-317.

[6] DEVLI J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics，2018：4171-4186.

[7] 苏娅，刘杰，黄亚楼. 在线医疗文本中的实体识别研究[J]. 北京大学学报（自然科学版），2016，52（1）：1-9.

[8] 张帆，王敏. 基于深度学习的医疗命名实体识别[J]. 计算技术与自动化，2017，36（1）：123-127.

[9] 黄琼影. 在线医疗社区问答文本的知识图谱构建研究[D]. 广州：华南理工大学，2020.

[10] 董美，常志军. 一种面向中医领域科技文献的实体关系抽取方法[J]. 图书情报工作，2022，66（18）：105-113.

[11] ZHENG S C，WANG F，BAO H Y，et al. Joint extraction of entities and relations based on a novel tagging scheme[J]. ArXiv e-Prints，2017：arXiv：1706.05075.

[12] WEI Z P，SU J L，WANG Y，et al. A novel cascade binary tagging framework for relational triple extraction[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics，2020：1476-1488.

[13] 周俊，郑彭元，袁立存，等. 基于改进CASREL的水稻施肥知识图谱信息抽取研究[J]. 农业机械学报，2022，53（11）：314-322.

[14] ZHANG N Y，JIA Q H，YIN K P，et al. Conceptualized representation learning for Chinese biomedical text mining[EB/OL]. [2024-02-05]. https://www.semanticscholar.org/reader/2b01b3334ce950c76c9c3c2c9146a7f0ce79cc50.

[15] 小儿腹泻病[EB/OL]. [2023-11-03]. https://baike.baidu.com/item/小儿腹泻病/12677256?fr=ge_ala.

[16] 李倩，王帅. LDA模型下我国公共图书馆微信平台阅读推广内容主题研究[J]. 图书情报工作，2022，66（8）：72-83.

[17] BLEI D M，NG A Y，JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research，2003（3）：993-1022.

[18] REIMERS N，GUREVYCH I. Sentence-BERT：sentence embeddings using Siamese BERT-Networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing，2019：3982-3992.

[19] 曾桢，赵浩宇. 基于文献的中国近代史知识图谱构建与实证研究[J]. 数字图书馆论坛，2022（4）：35-42.

[20] 廖开际，黄琼影，席运江. 在线医疗社区问答文本的知识图谱构建研究[J]. 情报科学，2021，39（3）：51-59，75.

作者简介

孟秋晴，女，博士，副教授，研究方向：网络信息组织、信息服务。
郑铭瑞，男，硕士研究生，通信作者，研究方向：数据挖掘与数据分析，E-mail：442823913@qq.com。
田玥璐，女，硕士研究生，研究方向：数据挖掘与信息服务。
刘逸品，女，硕士研究生，研究方向：数据挖掘与信息服务。
王琼弟，男，硕士研究生，研究方向：数据挖掘与数据分析。

Construction of Medical Health Knowledge Map for UGC in Online Health Community: Taking Child Diarrheal Disease as an Example

MENG QiuQing[1] ZHENG MingRui[1] TIAN YueLu[1] LIU YiPin[1] WANG QiongDi[2]
(1. School of Information, Guizhou University of Finance and Economics, Guiyang 550025, P. R. China;
2. Software Institute, Nanjing University, Nanjing 210008, P. R. China)

Abstract: It is of great significance to construct the medical health knowledge map oriented to the user generated content (UGC) data of online health community and explore the health knowledge extraction based on the potential needs of users to optimize the information organization and retrieval of online health community and support the knowledge service innovation of online health community. This paper proposes a combined entity recognition model LDA-BERT-BiLSTM-CRF based on UGC data of online health communities. We use the LDA topic model to perform thematic cluster analysis on UGC data of online health communities to extract entity types. Based on subdivision entity type, BERT-BiLSTM-CRF model is used to identify named entity. Then, MC-BERT-CasRel model is used to extract overlapping triples from UGC data in online health communities. Entity alignment is realized by SBERT model. Finally, the storage and visualization of knowledge map are realized by using Neo4j graph database. Taking child diarrheal disease as an example, a knowledge map of child diarrheal disease containing 939 entities and 3 224 relationships is constructed based on this method. Compared with the current mainstream models, the results show that the combined model LDA-BERT-BiLSTM-CRF and the relationship extraction model MC-BERT-CasRel are more accurate than the traditional knowledge extraction methods, and the entity classification is more targeted.

Keywords: Knowledge Map Construction; Online Health Community; UGC; LDA; Knowledge Extraction

（责任编辑：王玮）