基于中文医药文本的实体识别和图谱构建

杨 晔,裴 雷,侯凤贞*

(中国药科大学理学院,医药大数据与人工智能研究院,南京 211198)

摘 要 知识图谱技术促进了新药研发的进展,但国内研究起点晚且领域知识多以文本形式存储,图谱重用率低。因此,本研究基于多源异构的医药文本,设计了以Bert-wwm-ext预训练模型为基础,并融合级联思想的中文命名实体识别模型,从而减少了传统单次分类的复杂度,进一步提高了文本识别的效率。实验结果显示,该模型在自建的训练语料上的F1分数达0.903,精确率达89.2%,召回率达91.5%。同时,将模型应用于公开数据集CCKS2019上,结果显示该模型能够更好地识别中文文本中的医疗实体。最后,利用此模型构建了一个中文医药知识图谱,图谱包含13 530个实体,10 939个属性,以及39 247个相关关系。本研究所提出的中文医药实体识别与图谱构建方法,有望助力研究者加快医药知识新发现,从而缩短新药研发进程。

关键词 中文医药文本;命名实体识别模型;Bert-wwm-ext预训练模型;级联思想;知识图谱

中图分类号 TP391;R28 **文献标志码** A **文章编号** 1000-5048(2023)03-0363-09

doi:10.11665/j.issn.1000-5048.2023030903

引用本文 杨晔,裴雷,侯凤贞.基于中文医药文本的实体识别和图谱构建[J].中国药科大学学报,2023,54(3):363 - 371.

Cite this article as: YANG Ye, PEI Lei, HOU Fengzhen. Entity extraction and graph construction based on Chinese medical text[J]. *J China Pharm Univ*, 2023, 54(3): 363 - 371.

Entity extraction and graph construction based on Chinese medical text

YANG Ye, PEI Lei, HOU Fengzhen*

Institute of Medical Big Data and Artificial Intelligence, School of Science, China Pharmaceutical University, Nanjing 211198, China

Abstract Knowledge graph technology has promoted the progress of new drug research and development, but domestic research starts late and domain knowledge is mostly stored in text, resulting in low rate of knowledge graph reuse. Based on multi-source and heterogeneous medical texts, this paper designed a Chinese named entity recognition model based on Bert-wwm-ext pre-training model and also integrated cascade thought, which reduced the complexity of traditional single classification and further improved the efficiency of text recognition. The experimental results showed that the model achieved the best performance with an F1-score of 0.903, a precision of 89.2%, and a recall rate of 91.5% on the self-built dataset. At the same time, the model was applied to the public dataset CCKS2019, and the results showed that the model had better performance and recognition effect. Using this model, this paper constructed a Chinese medical knowledge graph, involving 13 530 entities, 10 939 attributes and 39 247 relationships of them in total. The Chinese medical entity extraction and graph construction method proposed in this paper is expected to help researchers accelerate the new discovery of medical knowledge, and shorten the process of new drug discovery.

Key words Chinese medical text; named entity recognition model; Bert-wwm-ext pre-training model; cascade thought; knowledge graph

近年来,基于知识图谱的药物新靶点发现[1]、药物不良反应预测[2]以及药物重定位[3]等在药物研发领域取得了有效的成果。在医药领域,知识图谱可以将疾病与药物等相关信息之间的复杂关系以一种图结构的形式呈现,有效解决了知识孤岛的现象[4]。目前,DrugBank、SNOMED-CT以及

收稿日期 2023-03-09 * **通信作者** Tel:025-86185121 E-mail:houfz@cpu.edu.cn

PharmKG等都是医药领域成熟、稳定且规模较大的英文知识图谱，已经被广泛应用于医药研究。而国内对于知识图谱的研究起步较晚，市面上能直接投入应用的中文医药知识图谱还较少。随着我国医药信息化/数字化建设规模的不断扩大，医药领域积累了海量的文本数据。如何从这些中文医药文本中构建知识图谱，对于这些数据的管理与利用有着重要的意义。

知识图谱是一种大规模的语义网络，通常由若干节点和边组成，其中节点表示客观世界中的各种实体、属性，边则表示了两两实体之间的语义关系[5-7]。如图1所示，左边是一个由疾病实体“肺炎链球菌肺炎”、药物实体“头孢曲松”以及两实体之间的关系组成的三元组，记为<“肺炎链球菌肺炎”“治疗药物”“头孢曲松”>。它可以理解为，“肺炎链球菌肺炎”疾病的治疗药物有“头孢曲松”。而由多个这样的实体关系或属性关系形成的三元组，可以组成含有丰富语义的知识图谱，并且能够通过图数据库进行高效地存储与管理[8]。

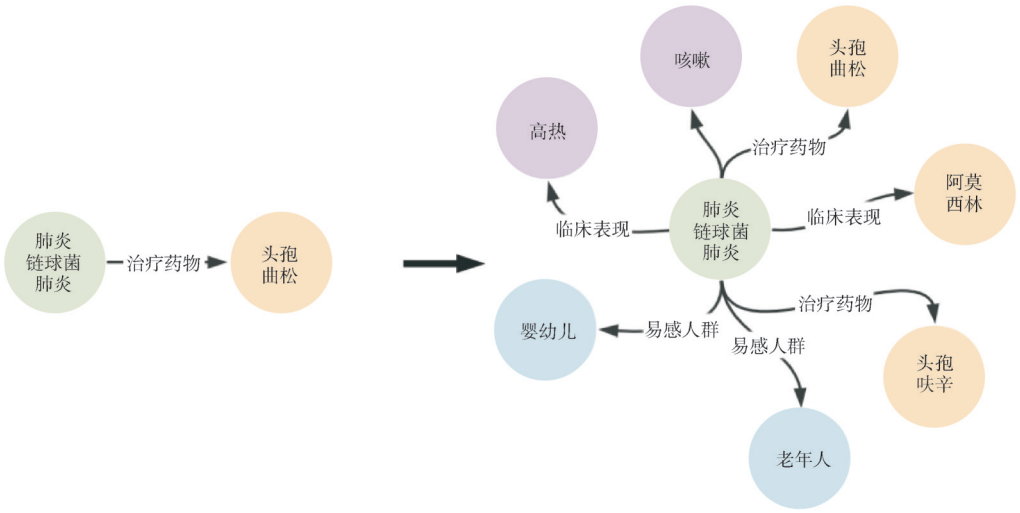

图1 知识图谱组成示例

非结构化文本数据是医疗活动过程中产生的一类重要的信息资源，也是健康医药数据的重要组成部分[9]。该类文本数据包含丰富的医药知识，但很难通过统一的规则处理。信息抽取作为该类文本处理的关键技术，也是构建知识图谱的前提，旨在将复杂文本中有用的信息以结构化统一的形式呈现出来[10]。其中，实体识别是信息抽取最关键的环节，其目的在于提取文本中特定的词汇并将其归为预先定义好的实体类别[11-12]。

传统的命名实体识别主要基于词典及规则的方法实现[12-15]，该方法虽然能取得较高的准确率，但是这需要大量的人力资源和全面的专业知识，并且多变的数据很难依靠有限的规则提取。后来，机器学习逐渐成为了实体识别研究的主流方向，研究者们通过使用机器学习模型，如隐马尔科夫模型[16]、条件随机场(conditional random fields，CRF)[17]以及支持向量机[18]等，结合领域数据的特征进行命名实体的抽取。例如，张朝胜等[19]基于CRF模型结合产品名特有的指标信息特征，构建了英文产品命名实体的自动识别模型。随着信息技术的发展，深度学习作为机器学习研究的一个新领域，在实体识别任务中取得了显著的效果，其性能和效果都超过了传统的算法。常见的深度学习模型有循环神经网络[20]、双向长短期记忆网络(bi-directional long short-term memory，BiLSTM)[21]以及注意力机制[22]等。随着算力的不断提升，许多以深层神经网络为基础的高性能预训练模型应运而生，如典型的Transformer[23]、Bert[24]等。在命名实体识别领域，将预训练好的Bert作为编码层并将该层获得的词向量输入到BiLSTM和CRF中进行特征提取和序列解码的方法已经成为主流并取得了较好的效果[25-26]。例如，许力等[27]将Bert+BiLSTM+CRF组合模型应用于生物医学命名实体识别领域，在BC4CHEMD、NCBI-disease等数据集

上均取得了较好的识别效果。然而，相较于英文，中文的词边界更加难以区分；且在不同的语境下，同一词语的表述也有所不同。医药领域的中文文本更加专业，这为中文医疗实体的识别带来了挑战。

因此，本研究尝试以传统医学教材以及垂直网站等作为数据来源，以自顶向下的方式构建一个中文医药知识图谱（具体流程如图 2 所示）。由于数据中存在大量非结构化的复杂文本，本研究设计、训练、测试并验证了一个适用于中文医疗实体识别的模型，以用于该类复杂文本中实体的自动抽取，从而实现数据结构化处理。本研究能够有效利用现有的医药数据，助力于加速数据驱动的药物发现过程。

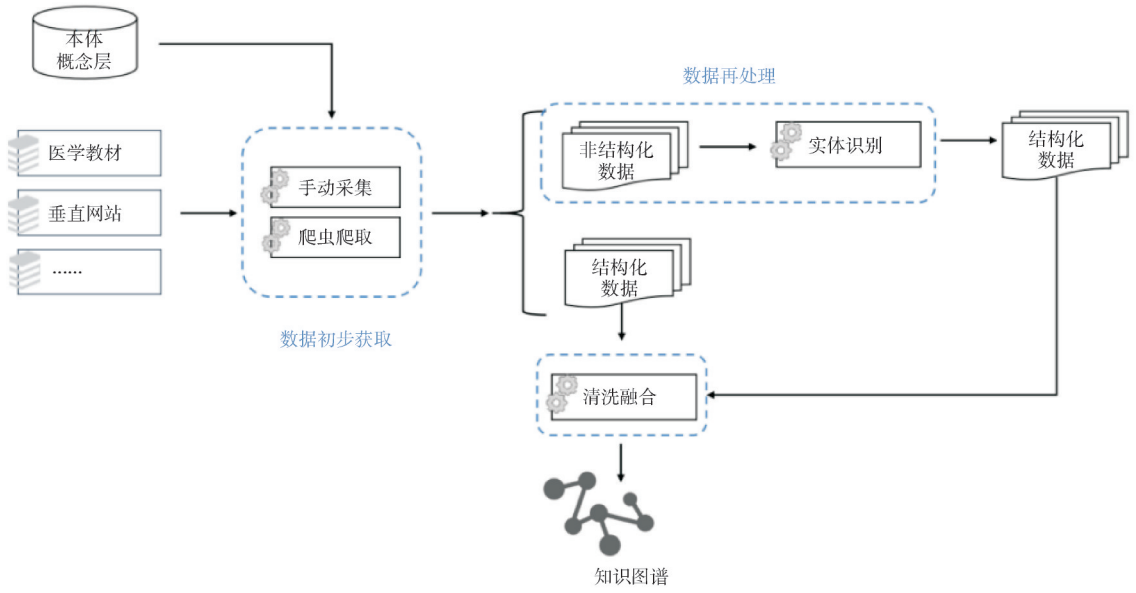

图 2　知识图谱构建流程

1　方　法

1.1　知识图谱本体概念层的设计

本体概念层是知识图谱构建的基础，主要用于对图谱数据层的规范和约束。在图谱构建初期，本研究定义了包含“疾病”“临床表现”“检查”“药物”“人群”和“身体部位”共 6 种实体，药物的“性状”“功能”“用法用量”“规格”“性味与归经”以及“贮藏”等 6 种属性，以及“疾病-检查”“疾病-治疗”等 14 种关系在内的图谱本体概念框架，具体如图 3 所示。

其中，对涉及的相关关系描述和具体示例如表 1 所示。

1.2　数据初步获取

本研究知识图谱数据主要源于传统医学教材（《内科学》和《眼科学》）、三九健康网（http://www.39.net/）、中国医药信息查询平台（https://www.dayi.org.cn）。此外，为了丰富图谱内容，还以 2020 年版《中华人民共和国药典》（https://db.ouryao.com/）为数据来源，加入了具有中国特色的中药数据。

对于传统医学教材，本研究直接使用其电子版文本；而对于其他的数据，则基于本体概念层定义的关系，使用 Python 的 Selenium 库（https://www.selenium.dev/）进行爬取。由于源数据分布均具有半结构化特点，因此，本研究根据本体概念层定义的实体、属性关系，直接建立数据映射。如图 4 示例所示，由于源数据分布均具有半结构化特点，本文根据本体概念层定义的实体、属性关系，对半结构化提取后的数据直接建立映射。将源数据中半结构化分布的关系（如“临床表现”“检查”等）和其对应的具体文本收集存储，以此获得信息记录。

最终，共获得以疾病为中心的信息记录（包括疾病的临床表现、检查等）共 1 992 条，以及以药物为中心的信息记录（包括药物的性状、功能等）2 269 条。

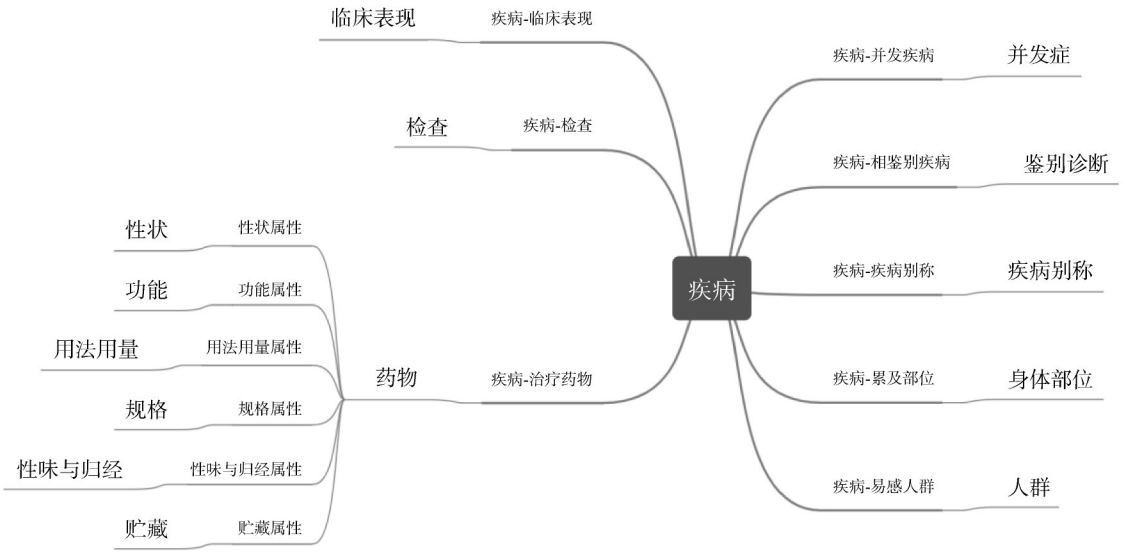

图3 本体概念层框架

表1 本体概念层关系描述和示例

关 系	描 述	示 例
疾病-临床表现	得了某些疾病后,身体出现的一系列的变化,是疾病在患者身上的表现	肺脓肿,临床表现,盗汗
疾病-检查	为了得到更多的由疾病导致的异常表现以及支持诊断而采取的检查项目	睡眠呼吸暂停低通气综合征,检查,心电图
疾病-治疗药物	指预防治疗疾病、调节人的生理功能所对应采取的并规定有适应证或用法用量的物质	肺炎链球菌肺炎,治疗药物,头孢曲松
疾病-并发疾病	是指一种疾病在发展过程中引起另一种疾病的发生	支气管哮喘,并发症,气胸
疾病-相鉴别疾病	根据患者的主诉,与其他疾病鉴别,并排除其他疾病可能的诊断	支气管扩张症,鉴别诊断,慢性支气管炎
疾病-易感人群	指易受疾病感染的群体	肺炎衣原体肺炎,易感人群,学龄儿童
疾病-疾病别称	疾病的另一种名称	急性上呼吸道感染,疾病别称,上感
疾病-累及部位	受疾病影响,出现病变或不适的身体组织结构	消化性溃疡,累及部位,胃
性状属性	药物体的物理特征或形态	丁香,性状属性,本品略呈研棒状且长1～2 cm
功能属性	医药学理论所注明药物的功能	九香虫,功能属性,理气止痛、温中助阳
用法用量属性	药物的使用方法以及一定时间内服用的数量	三七,用法用量属性,研粉吞服:一次1～3 g
规格属性	药物每个单位所含主要成分的量	灵泽片,规格属性,每片重0.58 g
性味归经属性	药物的性质和气味和药物作用的所属定位	安息香,性味归经属性,辛、苦、平;归心、脾经
贮藏属性	药物的储存条件	清咽丸,贮藏属性,密封置阴凉干燥处

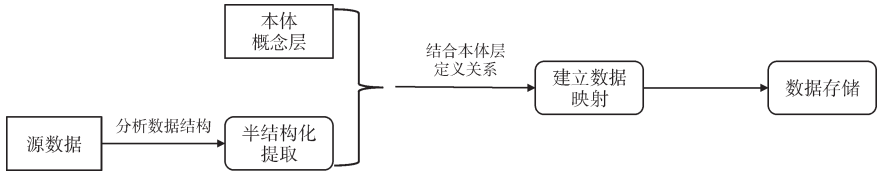

图4 数据初步获取流程

1.3 实体识别实验

1.3.1 *训练语料库构建* 基于上述初步获取的以疾病为中心的信息记录,本研究将其中所有非结构化文本随机打乱,并抽取部分数据作为训练语料。同时,利用标注工具Brat(http://brat.nlplab.org/index.html)在具有医学专业背景人员的协助下进行标注,最后将标注结果以json格式文件储存。如图5所示,文本标注后的结果中包括了“原始文本”和

所有的“实体”信息；对于每一个“实体”，都包括了它的类别、起始位置和结尾位置3项信息。

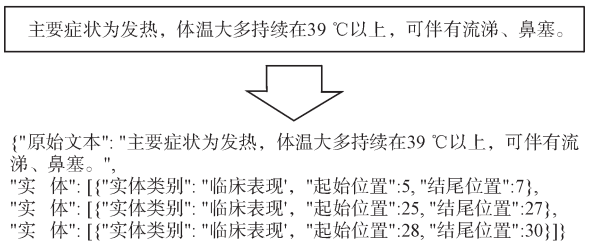

图5　语料的标注和存储示例

1.3.2　模型构建　通用实体识别模型多以Bert预训练语言模型为基层模型获取语义表征向量，但该类模型的设计理念更适用于英文语言，比如分词方式和掩码方式等。由于中文和英文本质上的差异（例如，中文文本通常是由连续字符组成，不同于英文的词与词之间会存在分隔符），掩码对象以字为单元极有可能会导致信息泄露[28]。因此，谷歌官方进一步提出了全词掩码（whole word masking，WWM）任务，即将最小的掩码单元由子词转换为全词。如图6所示，WWM会将同属一个词“咳嗽”的每一个子词“咳”“嗽”全部掩盖，而不是只掩盖某一个子词。这一方式可以让模型在训练过程中获得全词的语义信息，也更适用于中文命名实体识别的任务。

Bert-wwm-ext[29]是由哈尔滨工业大学讯飞联合实验室将Bert预训练语言模型与WWM技术相结合，同时采用更大规模的中文数据集进行训练并加大了训练步数后发布的语言模型。因此，本研究将该模型作为特征表示层模型。CRF是一种以无向图形式表达的概率分布模型，训练数据时可以自动学习标签间的依赖关系来，从而保证最终预测结果的有效性[30]。本研究将CRF层作为位置标签解码层，大大减少了预测序列的错误率。

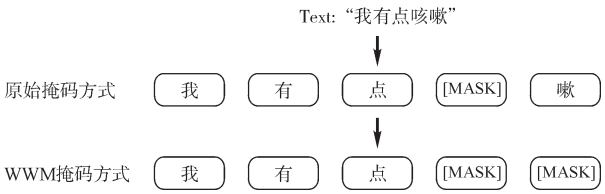

图6　不同掩码方式示例

为了提高识别效果，本研究还融入了级联思想[31]。对于原先的模型在CRF解码时是将所有类别的标签都考虑在内，而级联思想则是将传统单次多任务学习改为两次多任务学习。具体如下：

对于给定的文本序列$\boldsymbol{I}=\{\boldsymbol{i}_1,\boldsymbol{i}_2,\ldots,\boldsymbol{i}_t\}$，其中$\boldsymbol{i}_t$是第$t$个字符的词表征向量。本研究选择使用实体识别任务领域内常用的“B I O S”标注方式作为序列标注标签[32]。其中B表示实体的起始字符，I为实体中间字符，O为非实体字符，S为单个实体字符。如图7所示，对于文本序列“我有点咳嗽”，将非实体标注为“O”，将临床表现类实体的第1个字“咳”标注为“B”，第2个字“嗽”标注为“I”。

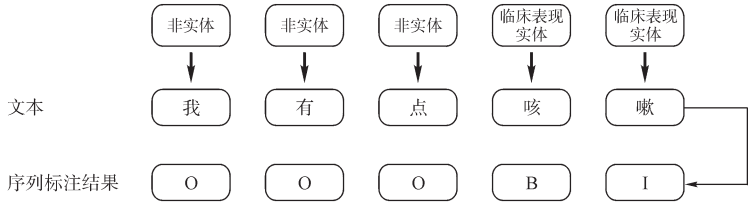

图7　文本序列标注示例

模型通过第1次分类可获得输入序列的“BIOS”分类标签，损失可以通过字符在真实位置标签概率的负对数表示，如式（1）所示：

$$\text{loss}_{\text{p}}=-\sum_{t=1}^{n}\log(P(y_{\text{p}=\text{true}}|\boldsymbol{i}_t)) \tag{1}$$

模型的第2次分类是具体实体类型层面的分类，如“疾病”“临床表现”等，可以通过损失向量点积运算获得，如式（2）所示。其中，mask为掩码信息，可根据不同标签设置权重以此区分出输入序列中的实体范围。

$$\text{loss}_{\text{e}}=-\sum_{t=1}^{n}\log(P(y_{\text{e}=\text{true}}|\boldsymbol{i}_t))\cdot\text{mask} \tag{2}$$

模型最终的损失为第1层和第2层损失相加的结果，如式（3）所示：

$$\text{loss}=\text{loss}_{\text{p}}+\text{loss}_{\text{e}} \tag{3}$$

综上所述，本研究设计的中文医疗命名实体识别模型的整体结构如图8所示。

1.3.3　模型验证　本研究选择常用实体识别模型Bert+BiLSTM+CRF作为对比实验模型，并将Bert替换为Bert-wwm-ext后的组合作为消融实验

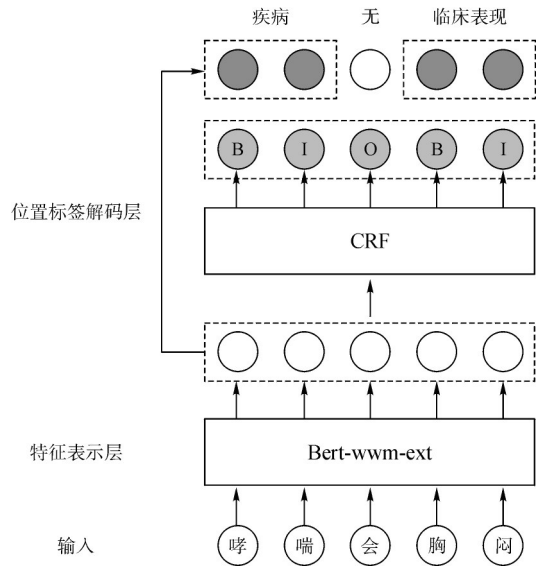

图8 实体识别模型结构

模型。同时为了验证模型的性能,本研究还准备了CCKS2019(http://www.sigkg.cn/ccks2019/)命名实体识别任务数据集。CCKS2019评测竞赛是由中国文学信息学会语言与知识计算专委会举办,旨在为研究人员提供一个测试技术和算法的平台。在该数据集中,官方发布了共1 379条训练语料,其中包括6种医疗实体类别,分别为"疾病和诊断"类共实体2 798个、"检查"类实体共313个、"检验"类实体共511个、"手术"类实体共905个、"药物"类实体共719个以及"解剖部位"类实体共1 933个。

1.3.4 *模型参数设置* 本实验均使用8 GB 1080 Ti GPU和Pytorch(version 1.11.0)进行搭建。模型参数配置中,为抑制过拟合设置Dropout为0.1,学习率设置为4×10^{-5},模型的最大序列长度设置为256。

1.3.5 *模型评价指标* 对于模型性能的评估,本研究使用了F1分数(F1-score)、精确率(precision)以及召回率(recall)作为命名实体识别的评价指标。精确率是指模型所得分类结果中,预测为正样本中真正的正样本的比值,也可叫作查准率;召回率是指模型所得分类结果中,实际为正的样本中被预测为正样本的比值,也可叫作查全率;F1分数则是综合考虑精确率和召回率两个指标后计算得出[33-34],具体计算方法如式(4)~(6)所示:

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{4}$$

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{5}$$

$$\text{F1}=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{6}$$

其中,TP表示被模型预测为正类的正样本;TN表示被模型预测为负类的负样本;FP表示被模型预测为正类的负样本;FN表示被模型预测为负类的正样本。

1.4 *清洗融合*

数据层填充是图谱构建的最关键的一个步骤。在将初步获取的结构化数据,以及基于实体识别模型处理后的结构化数据进行知识融合的过程中,由于数据来源多样,且各来源的知识表达形式也不尽相同。因此,获得的实体数据经常存在多词一义的情况(例如,"视力下降"和"视力减弱"等)。为了解决上述问题,本研究利用Bert-wwm-ext预训练语言模型获取各实体的语义表征向量,并通过余弦距离来衡量各它们之间的相似性,将相似度大于0.9的实体看作同一个实体,并以第一次出现时的表述的为统一标准。对于实体向量$\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$和$\boldsymbol{y}=\{y_1,y_2,\cdots,y_n\}$,其余弦距离可通过公式(7)计算:

$$\text{sim}(x,y)=\frac{\sum_{i-1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}x_i^2}\times\sqrt{\sum_{i=1}^{n}y_i^2}} \tag{7}$$

最后本研究将规范融合的结构化知识数据,以三元组的形式填充进Neo4j(https://neo4j.com/)图数据库中。Neo4j是领域内最为流行的图数据库,一方面,它能够支持海量数据的存储和管理,另一方面它使用的是功能强大的Cypher查询语言,允许在数据库内进行高效的数据检索和更新[35]。

2 结 果

2.1 *实体识别结果分析*

本研究构建的训练语料共1 637条,将其随机以6:2:2的比例依次划分为训练集、验证集和测试集进行实验,具体实体类型与数量分布如表2所示。

在该数据集上,本研究将通用实体识别模型与提出的模型进行了对比试验。从表3结果可以看出,将通用模型Bert+BiLSTM+CRF中的预训练模型换成Bert-wwm-ext后,F1分数、精确率和召回率均有所提高。可以认为,与Bert基线模型相比,

表 2　训练语料信息统计

数据集	文本数	疾病	检查	身体部位	人群	临床表现	药物
训练集	981	587	529	445	458	686	603
验证集	328	223	169	178	148	291	190
测试集	328	192	182	152	171	287	208

Bert-wwm-ext 模型可以获得更加丰富的语义信息和对上下文的理解能力，更加适用于中文文本的挖掘和处理。此外，将传统模型的一次分类任务分解为两次分类后，模型指标 F1 分数、精确率、召回率均比其他模型提高 1% ~ 2%。

表 3　训练语料在不同模型的预测性能结果

模型	F1	精确率/%	召回率/%
Bert + BiLSTM + CRF	0.879	86.7	89.1
Bert-wwm-ext + BiLSTM + CRF	0.887	87.7	89.7
Bert-wwm-ext + 级联 + CRF	0.903	89.2	91.5

为了验证提出模型的性能效果，本研究将公开数据集 CCKS2019 随机按 6∶2∶2 的比例划分成训练集、验证集和测试集，并同样在不同模型上进行对比试验。考虑到原始语料数据中每一条记录的文本长度偏长，在实验前，本研究将其按“。”分隔符号进行了拆分。运行结果如表 4 所示，以 Bert-wwm-ext 为基层的模型较 Bert 能获取更高的识别率。此外，本研究提出的实体识别模型总体性能较通用模型也是有所上升。综合来看，本研究构建的命名实体识别模型对于中文医疗文本的信息提取有着积极的作用。

表 4　CCKS2019 数据集在不同模型的预测性能结果

模型	F1	精确率/%	召回率/%
Bert + BiLSTM + CRF	0.807	79.9	81.5
Bert-wwm-ext + BiLSTM + CRF	0.811	80.1	82.0
Bert-wwm-ext + 级联 + CRF	0.814	80.9	81.9

2.2　知识图谱规模和成果展示

基于中文医疗实体识别模型构建的中文医药知识图谱共包含实体数据 13 530 个，分别为“疾病”4 347 个、“药物”3 561 个、“临床表现”3 852 个、“人群”115 个、“检查”1 462 个和“身体部位”193 个；属性数据共 10 939 个，分别为“性状”2 254 个、“功能”2 173 个、“用法用量”2 204 个、“规格”1 445 个、“性味与归经”611 个和“贮藏”2 252 个；包含关系数据 39 247 个，其中“疾病-临床表现”“疾病-治疗药物”等实体关系 25 965 个，药物的“性状属性”“功能属性”等属性关系 13 282 个。部分可视化中文医药知识图谱示例如图 9 所示。

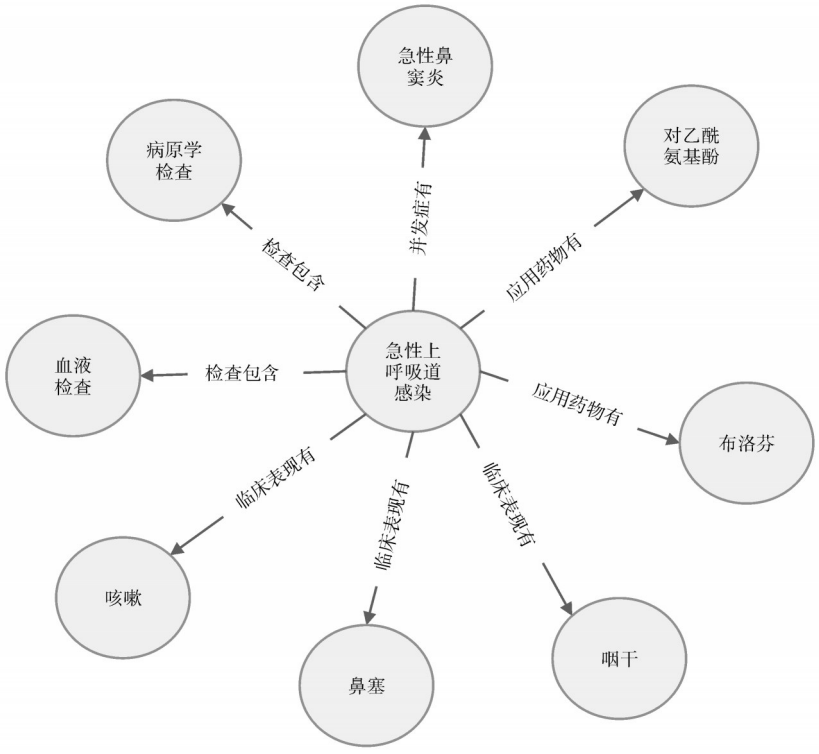

图 9　中文医药知识图谱示例成果展示

3 讨 论

本研究基于传统医学教材和垂直网站等多来源数据,采用自顶向下的方法构建了中文医药知识图谱。同时,为了实现文本中医疗实体的自动识别,本研究以Bert-wwm-ext预训练模型作为基层模型,并将传统的一次多分类任务分解为两次分类任务,构建了适用于中文文本的实体识别模型,并在自建的训练语料集和公开数据集CCKS2019中分别进行了对比试验。结果表明,与通用实体识别模型Bert+BiLSTM+CRF相比,本研究构建的实体识别模型效果更优。

本研究在中文医药文本中实体的高效抽取以及知识图谱的构建上进行了有意义的尝试,有助于研究者们有效利用现有的医药数据,实现基于知识图谱的药物新靶点发现、药物不良反应预测、药物重定位等应用,加速基于知识数据驱动的新药研发过程。然而,本研究用于模型实验的语料数据量还不足够多,在未来的研究工作中仍需引入更多来源的中文医疗文本作为训练数据,以期提高实体识别模型的泛化能力。同时,还会加入实体关系的链接预测和知识表示技术更进一步补全知识图谱。此外,针对药物新靶点发现等应用研究,未来工作还会增加"药物结构"等医药实体类别和相关关系扩大图谱范围,以此为进一步推断疾病和药物之间的关系而作出贡献。

References

[1] Mohamed SK, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings[J]. *Bioinformatics*, 2020, **36**(2): 603-610.

[2] Lukashina N, Kartysheva E, Spjuth O, *et al*. SimVec: predicting polypharmacy side effects for new drugs[J]. *J Cheminform*, 2022, **14**(1): 49.

[3] Li ZX. Relocation of Parkinson′s disease drugs based on knowledge graph[J]. *Inf Technol* (信息技术与信息化), 2022(7): 28-32.

[4] Wu XD, Sheng SJ, Jiang TT, *et al*. Huapu-CP: From knowledge graphs to a data central-platform[J]. *JAS* (自动化学报), 2020 (10): 2045-2059.

[5] Fan YY, Li ZM. Research and application progress of Chinese medical knowledge graph[J]. *J Front Comput Sci Technol* (计算机科学与探索), 2022, **16**(10): 2219-2233.

[6] Qi GL, Gao H, Wu TX. Research progress of knowledge map[J]. *Inf Eng*(情报工程), 2017, **3**(1): 4-25.

[7] Ma XG. Knowledge graph construction and application in geosciences: a review[J]. *Comput Geosci*, 2022, **161**: 105082.

[8] Li ZW, Ding Y, Hua ZY, *et al*. Knowledge graph completion model based on triplet importance integration[J]. *Comput Sci* (计算机科学), 2020, **47**(11): 231-236.

[9] Hu JH, Zhao WQ, Fang A. Research on clinical text processing and knowledge discovery method based on medical big data[J]. *China Digit Med* (中国数字医学), 2020, **15**(7): 11-13, 88.

[10] Guo XY, He TT. A survey of information extraction[J]. *Comput Sci* (计算机科学), 2015, **42**(2): 14-17,38.

[11] de Aquino Silva R, da Silva L, Dutra ML, *et al*. An improved NER methodology to the Portuguese language[J]. *Mobile Netw Appl*, 2021, **26**(1): 319-325.

[12] Liu P, Guo YM, Wang FL, *et al*. Chinese named entity recognition: the state of the art[J]. *Neurocomputing*, 2022, **473**: 37-53.

[13] Wu ST, Liu HF, Li DC, *et al*. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis [J]. *J Am Med Inform Assoc*, 2012, **19**(e1): e149-e156.

[14] Friedman C, Alderson PO, Austin JH, *et al*. A general natural-language text processor for clinical radiology[J]. *J Am Med Inform Assoc*, 1994, **1**(2): 161-174.

[15] Chiticariu L, Krishnamurthy R, Li YY, *et al*. Domain adaptation of rule-based annotators for named-entity recognition tasks[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, Massachusetts. New York: ACM, 2010: 1002－1012.

[16] Eddy SR. Hidden Markov models[J]. *Curr Opin Struct Biol*, 1996, **6**(3): 361-365.

[17] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence dat [C]. ICML. New York: Association for Computing Machinery, 2001:282-289.

[18] Cortes C, Vapnik V. Support-vector networks[J]. *Mach Learn*, 1995, **20**: 273-297.

[19] Zhang CS, Guo JY, Xian YT, *et al*. English product named entity recognition based on conditional random field[J]. *Comput Sci Eng* (计算机工程与科学), 2010, **32** (6): 115-117.

[20] Elman JL. Finding structure in time[J]. *Cogn Sci*, 1990, **14**(2): 179-211.

[21] Cai LQ, Zhou ST, Yan X, *et al*. A stacked BiLSTM neural network based on coattention mechanism for question answering [J]. *Comput Intell Neurosci*, 2019, **2019**: 9543490.

[22] Xu YS, Li L, Gao HH, *et al*. Sentiment classification with adversarial learning and attention mechanism[J]. *Comput Intell*, 2021, **37**(2): 774-798.

[23] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all You need[J]. *arXiv*,2017:1706.03762.

[24] Devlin J, Chang MW, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding[J]. *arXiv*,

2018: 1810.04805

[25] Song YH, Tian SW, Yu L. A method for identifying local drug names in Xinjiang based on BERT-BiLSTM-CRF[J]. *Autom Control Comput Sci*, 2020, **54**(3): 179 – 190.

[26] Chen LM, Liu D, Yang JK, *et al*. Construction and application of COVID-19 infectors activity information knowledge graph[J]. *Comput Biol Med*, 2022, **148**: 105908.

[27] Xu L, Li JH. Biomedical named entity recognition based on BERT and BiLSTM-CRF[J]. *Comput Sci Eng*, 2021(10): 1873-1879.

[28] Hou YT, Abduklimu A, Haridamu A. Research progress of Chinese pre training model[J]. *Comput Sci* (计算机科学), 2022, **49**(7): 148-163.

[29] Cui YM, Che WX, Liu T, *et al*. Pre-training with whole word masking for Chinese BERT[J]. *IEEE/ACM Trans Audio Speech Lang Process*, 2021, **29**: 3504-3514.

[30] Song SL, Zhang N, Huang HT. Named entity recognition based on conditional random fields[J]. *Clust Comput*, 2019, **22**(3): 5195-5206.

[31] Wei ZP, Su JL, 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. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 1476-1488.

[32] Zheng SC, Wang F, Bao HY, *et al*. Joint extraction of entities and relations based on a novel tagging scheme[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 1227-1236.

[33] Luque A, Carrasco A, Martín A, *et al*. The impact of class imbalance in classification performance metrics based on the binary confusion matrix[J]. *Pattern Recognit*, 2019, **91**: 216-231.

[34] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks[J]. *Inf Process Manag*, 2009, **45**(4): 427-437.

[35] Sen S, Mehta A, Ganguli R, *et al*. Recommendation of influenced products using association rule mining: Neo4j as a case study[J]. *SN Comput Sci*, 2021, **2**(2): 1-17.

[**专家介绍**] 侯凤贞,博士,教授,美国哈佛大学访问学者,江苏省“青蓝工程”优秀青年骨干教师。近年来,相继主持和参与多项国家/省级自然科学基金项目,主持多个横向课题。正在开展的研究主要集中在两个方面:一是通过对各种生物医学信号(如心电、脑电、功能磁共振信号)的分析来挖掘生理系统的内在机制,从而为临床应用,如疾病诊断、健康监测等提供参考;二是探索人工智能在大健康领域的应用场景,如药物重定位、睡眠的科学评估、心脏病的精准预测以及老年痴呆症的及早诊断等。以第一作者或通信作者身份在*Sleep*,*Progress in Neuropsychopharmacology & Biological Psychiatry*,*Sleep Medicine*,*Frontiers in Neuroscience*等国际学术期刊上发表研究论文20余篇。