面向儿科疾病的实体及实体关系标注语料库构建

昝红英[1,2] 刘涛[1,2] 陈俊富[1,2] 李俊卓[1,2] 牛常勇[1] 赵悦淑[2,3]

张坤丽[1,2] 穗志方[2,4]

（1. 郑州大学 信息工程学院,河南 郑州 450001；2.鹏城实验室,广东 深圳 518052；3.郑州大学第三附属医院，河南 郑州 450001； 4.北京大学 计算语言学教育部重点实验室,北京 100871）

摘要：针对当前医学语料库涵盖实体分类以及实体关系难以满足精准医学发展需求的问题，本文从儿科疾病入手，参考现有的医学命名实体和实体关系标注体系，在医学领域专家的指导下，制定了适合儿科学的命名实体和实体关系的标注体系及详细标注规范；利用自行开发的标注工具，在采用机器学习进行预标注实体及实体关系后；以标注规范为指导，进行多轮人工标注，完成了 298 余万字的儿科医学文本中的实体及关系进行标注，形成了面向儿科疾病的实体及实体关系标注语料库。所构建的语料库包含 504 种儿科常见疾病，共标注命名实体 23,603 个，实体关系 36,513 个，多轮标注一致性分别为 0.85 和 0.82。抽取已构建实体及关系标注语料库中的多元组，形成了儿科医学知识图谱，并开发了基于知识图谱的儿科医学知识问答系统。

关键词：儿科疾病；语料库建设；命名实体；实体关系;知识图谱

Corpus Construction for Named-Entity and Entity Relations for Paediatric Diseases

ZAN Hongying[1,2], LIU Tao[1,2], CHEN Junfu[1,2], NIU Changyong[1], ZHAO Yueshu[2,3]
ZHANG Kunli[1,2] ,SUI Zhifang [2,4]

(1.School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China；
2.The Peng Cheng Laboratory,Shenzhen, Guangdong 518052, China；
3.The Third Affiliated Hospital of Zhengzhou University, Zhengzhou,Henan 450001, China；
4.Key Laboratory of Computational Linguistics, Ministry of Education,Peking University, Beijing 100871, China)

Abstract: In view of the current medical corpus covering entity type and entity-relationship is challenging to meet the demand of precision medical development problems, this article obtains from the pediatric disease, reference of existing medical named entity and entity-relationship annotation scheme, under the guidance of experts in the field of medicine, which is suitable for pediatrics set of named entities and entity-relationship annotation scheme and detailed annotation standard; Using the self-developed annotation tool, after using machine learning to pre-tagging entity and entity-relationship; Under the guidance of annotation scheme, multiple rounds of manual tagging were carried out to complete the labeling of entities and relationships in pediatric medical texts of more than 298 million words, forming a corpus of entity and entity-relationship labeling oriented to pediatric diseases. The constructed corpus contains 504 common pediatric diseases, 23,603 named entities, 36,513 entity relationships, and the consistency of multiple rounds of tagging is 0.85 and 0.82, respectively. Multiple tuples in constructed entity and relational annotated corpus were extracted to form paediatric medical knowledge graph, and a paediatric medical knowledge answering system based on knowledge graph was developed.

Keywords: Corpus construction; Named entities; Pediatrics; Knowledge graph

0 前言

命名实体识别和实体关系抽取是信息抽取中的重要任务，同时也是篇章理解的核心技术，对信息检索、问答系统、信息

基金项目：国家社科基金重大项目（18ZDA315），河南省高等学校重点科研项目(20A520038)，河南省科技攻关项目（192102210260），河南省科技攻关计划国际合作项目（172102410065）

过滤、机器翻译等有重要意义。 正因为这两个任务的重要性，很多研究者在不同的应用领域都展开了深入的研究。医学文献的主要特征是未登录词数量庞大、文本充斥着大量命名实体、新的命名实体不断涌现、很多命名实体形式多样、命名实体以及实体间关系包含丰富的医学专业知识。因此医学领域是命名实体识别和实体关系抽取研究的重要领域。

中国是世界上儿童人数最多的国家。2015 年，0 至 14 岁儿童人口约为 2.27 亿[1]。伴随着全面二胎政策的开放，高龄产妇和新生儿会逐渐增多，儿童健康形式也愈发严峻，亟需建立儿科相关的医学知识图谱改善现有诊疗模式，以便快速、准确的获得临床决策支持。

经典教材作为学习相关知识的重要途径，专业性和严谨性被大众普遍认可。因此，本研究以经典儿科教材为依据，探索语料标注规范，构建儿科领域的命名实体及实体关系语料库，为医学命名实体识别和关系抽取、医学信息挖掘研究和知识图谱构建提供可靠的数据支撑。

目前，国内外在医学命名实体及实体关系的语料库的构建上都有了一定的进展。2006 年 Meystre 等[2]构建了涉及 80 种常见的医疗术语的命名实体标注语料，并且对每个医疗问题标注了其修饰词信息。共包含 160 份文档，文档类型包括病程记录、出院小结等信息，标注过程每份文档由两名医生分别标注，第 3 位医生负责裁定不一致的标注。2008 年美国梅奥诊所[3]研构建了 160 份医疗文档规模的命名实体语料，包括住院记录、门诊记录以及出院小结 3 种类型的医疗文档，对其中的疾病实体进行了标注，并且首次对实体和实体关系的修饰信息进行了细致的分类。2009 年 Roberts A 等[4]构建了 2 万份癌症患者病历的标注语料，用于开发和评估从患者病历中自动提取临床重要信息的系统，并且详细的介绍语料库的建设和标注方法。2014 年 Aurélie Névéol 等[5]采用机器标注和人工校对的方式构建了涉及 15 类实体的命名实体标注语料，共包含 2,500 篇医学文章，103,056 字，26,409 个实体概念。2017 年 Leonardo Campillos 等人[6]构建了 500 份文档规模的法语命名实体及实体关系语料库，文档类型包含出院小结、程序报告(如放射学报告)、医生来信和处方，对 11 类实体和 37 类关系进行了标注。

在国内，2013 年 Lei 等人[7]收集了北京协和医院 800 份电子病历并由 2 名医生标注完成了命名实体语料库，文档类型包括入院记录、出院小结，命名实体的分类借鉴 2010 年 I2B2 的实体分类，把治疗细分为药物和过程。为了展开中医病历命名实体研究，Wang 等人[8]于 2014 年构建了医学症状名的语料，包含 11,613 条主诉，标注工作由在职医生完成。2016 年杨锦锋[9]等人结合中文电子病历中命名实体的特点，制定了命名实体和实体关系的详细标注规范，通过手工标注构建了标注体系完整、规模较大的中文电子病历标注语料库，语料库包含病历文本 992 份。经过调研，目前中文医学命名实体和实体关系语料库主要存在以下几点问题:(1)大部分标注语料来源于互联网和电子病历，缺乏结构性和权威性。(2)命名实体分类以及实体关系类型不够丰富，无法完整的表示文本的语义信息。

《儿科学》[10]和《临床儿科学》[11]是国内儿科临床医学专业经典的教材和参考书，其中《儿科学》更注重理论学习，《临床儿科学》更多涉及医生的临床经验，本文以上述两本教材为依据对基础语料进行标注。相较于其他语料库，本语料全面贴合儿科特点。目前，该语料库已完成字数 298 万的标注工作，命名实体数目 23,603，实体关系数目达到 36,513。本语料库为后续进行医学知识图谱的构建提供了支持。

本文组织结构如下：第 1 章介绍语料库标注体系规范；第 2 章介绍语料库构建方法；第 3 章对已构建语料库进行了数据统计以及标注一致性分析。第 4 章给出本文结论。

1 命名实体和实体关系标注体系

参考 I2B2 2010 评测数据[12]以及中文电子病历语料库构建[9]的标注体系，本文将标注体系分为儿科保健体系和疾病体系，儿科保健主要涉及儿童健康、营养，目前该工作正在进行中。疾病体系以儿科疾病为中心，涵盖 11 类医学实体，45 种子关系的命名实体和实体关系标注体系，同时将传统的三元组扩展为六元组，即每一元都可以添加属性描述，信息描述更全面。

疾病体系将命名实体分为 11 大类，分别为疾病、部位、症状、药物、检查、其他治疗、手术、流行病学、预后、其他和社会学，并使用不同的参考标准界定每一类实体涵盖的范围。

命名实体的标注有三个基本原则：

（1）不重叠标注，即同一字符串不能标注为两钟不同的实体类型；

（2）不嵌套标注，即一个实体不能在另一个实体的内部；

（3）实体尽可能不包含标点符号（、，。：；）以及连接词（或、和、以及）；

围绕以疾病为中心，拓展疾病与其他实体间的关系包括：疾病-部位、疾病-症状、疾病-检查、疾病-疾病、疾病-其他治疗、疾病-手术、疾病-药物、疾病-流行病学、疾病-预后、疾病-其他、疾病-社会学等 11 类型关系。每种关系下面细分为子关系，具体的 45 种子类型如**错误!未找到引用源。**所示。

实体关系的标注原则：

（1）优先标注同句内的关系，若同句内不存在实体关系，允许跨句标注；

（2）关系属性和实体属性不确定的，优先标注为实体属性；

标注样例：急性肾炎临床表现轻重悬殊，咽炎为诱因者病前 6~12 天多有发热。

<急性肾炎[咽炎为诱因者]-临床症状[病前 6~12 天]-发热>

其中，”咽炎为诱因者”为实体急性肾炎的属性，”病前 6~12 天”为关系属性。

表 1 实体关系

实体 1	关系类型	实体 2	实体 1	关系类型	实体 2
疾病	疾病分型	疾病	疾病	发病率	流行病学
	鉴别诊断			死亡率	
	并发症			潜伏期	
	并发症（药物）			多发地区	
	并发症（术后）			多发群体	
	相关（导致）			多发季节	
	相关（转化）			传播途径	
	相关（症状）		疾病	病因	社会学
疾病	发病部位	部位		病理生理	
	累及部位			发病机制	
疾病	临床症状	症状		高危因素	
	临床体征			遗传因素	
	治疗后症状			病史	
疾病	辅助检查	检查	疾病	预后状况	预后
	影像学检查			预后生存率	
	病理学检查			预后生存时间	
	内窥镜检查		疾病	阶段	其他
	实验室检查			预防	
	筛查			就诊科室	
疾病	药物治疗	药物		标注住院时间	
疾病	手术治疗	手术		出院标准	
疾病	辅助治疗	其他治疗		注意事项	
任意实体	同义词	与实体 1 相同			

2 语料库构建过程

语料构建的核心工作是制定标注规范并依据规范对语料标注．目前主流的 3 种语料标注模式[13]有：(1)领域专家标注，该标注模式适于专业领域的语料标注，能够确保标注的质量，但标注成本高，周期长；(2)众包标注，该标注模式能够以较低的成本标注较大规模的语料，但仅限于简单的标注任务，并且标注过程也需要精心的设计，来保证标注质量；(3)团体标注，这种标注模式构建语料的过程类似于信息检索评价集的构建，能够在不依赖于专家的情况下，构建出高质量的语料，但对标注团体有很高的要求。

依据上一节提出的命名实体和实体关系的疾病标注体系制定了命名实体和实体关系标注规范，采用团体标注+领域专家的模式，对 504 种常见儿科疾病的文本进行标注，整个标注过程历时四个月，共有 2 名医学专家，9 名硕士生和 6 名本科生参与标注工作。为了提高标注效率，开发了配套的标注工具，如**错误!未找到引用源。**所示。

2.1 标注数据准备

经过和医学专家的讨论，结合医院的真实门诊记录，我们从基础语料中选取了 504 种儿科最为常见的疾病，所选疾病涉及消化系统疾病、免疫系统疾病、呼吸系统疾病、心脑血管疾病等 8 类疾病，各类疾病数量分布如**错误!未找到引用源。**所示。每个疾病对应一个单独的文本，这些文本都是半结构化文本数据，每个文本都含有多个小标题说明该部分文本描述的内容，比如“流行病学”、“诊断”、“治疗”、“病因”等．这种半结构化信息可以很好地指导命名实体的识别以及实体类型的判断，并且对实体关系的标注也能起到提示作用。

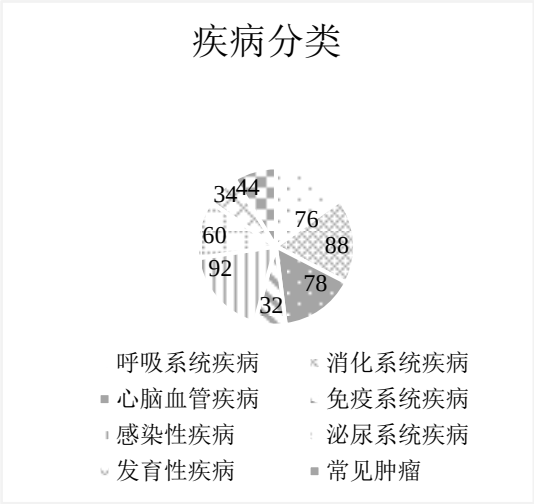

图 2　疾病分类

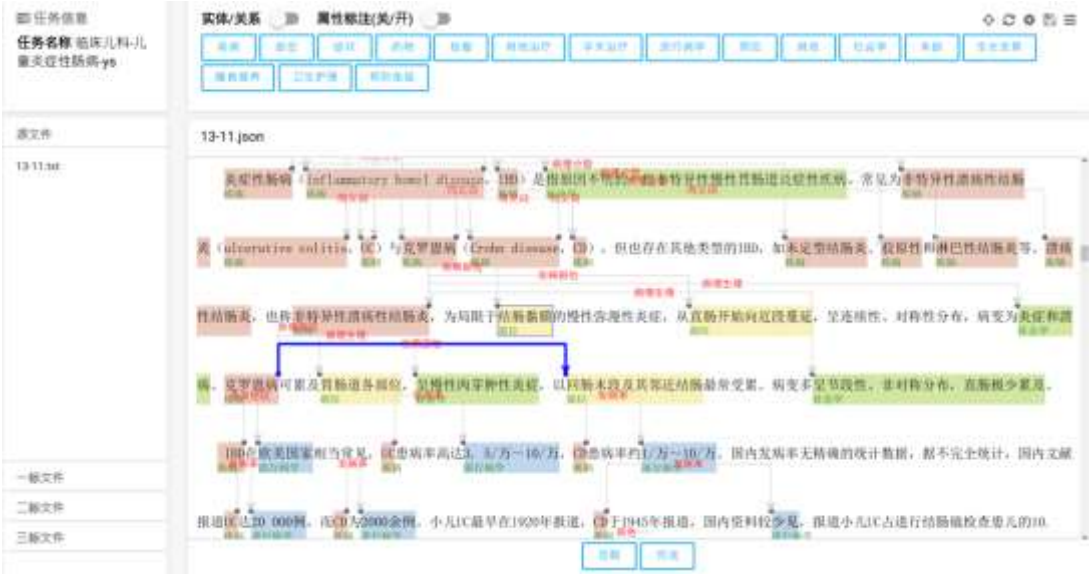

图 1 标注工具

2.2 标注过程

医学命名实体和实体关系标注规范的制定需要专业的医学知识，我们在分析了

标注语料的特点之后，参考中文电子病历标注规范，同医学专家讨论，制定出初步规范，采用多轮迭代模式进行标注规范的更新以及标注工作，整体流程如图 2 所示。

在第一阶段，详细的分析了《儿科学》和《临床儿科学》的文本以及儿科学的特点，在医学专家的指导下，确立了命名实体和实体关系的分类体系，制定了标注规范的第一版，开发了标注工具，并开始构建实体资源库。实体资源库主要包含疾病实体资源库、症状实体资源库、药物实体资源库。疾病实体资源库的构建来源于 MeSH(Medical Subject Headings)[14]主题词表、ICD-10(International Classification of Diseases)[15] 、ATC(Anatomical Therapeutic Chemical)[16]。症状实体资源库和药物实体资源库主要来源于互联网资源，收集多个不同医学百科类网站的资源，如医学百科，中国疾病知识总库，并对这些资源进行融合，确保实体资源库质量。

在第二个阶段，首先进行预标注，预标注目的在于减少重复的劳动，节省人力。依据在第一阶段收集的实体资源库采用最大双向匹配来对标注语料进行预标注操作，当然，预标注的质量难以得到保证，还需标注人员在正式标注过程中进行检查修改。

正式标注过程采用多轮迭代模式保证标注过程中的准确性和一致性。每个文本由两个标注者独立标注，简称为 A，B。A 标注完成后，B 进行二次标注，A、B 标注不一致和不确定的地方记录下来，经过与医学专家的讨论找出解决方案，再由 A 返回语料中进行修改，形成最终的三标版本。在此过程中，不断的对实体资源库进行更新，同时也会根据标注人员的反馈修订标注规范，使其更加贴合语料。

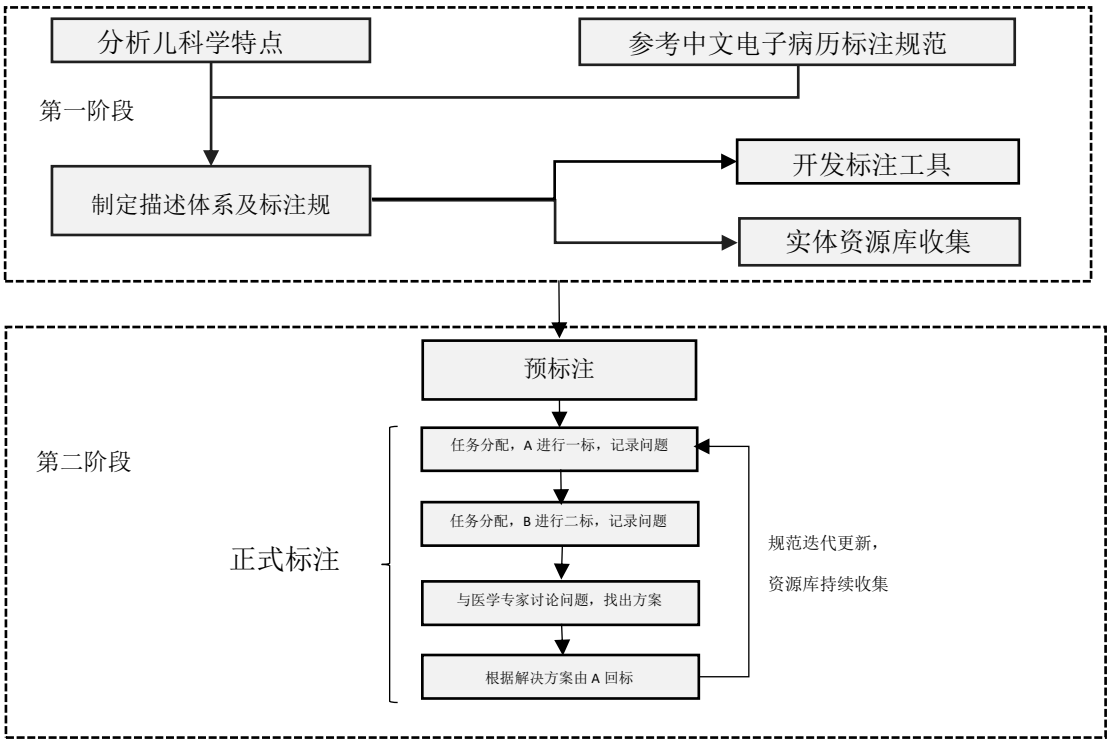

图 2 标注流程

2.3 标注过程常见问题及经验

标注语料库是自然语言处理高层应用的基础，然而标注语料库的构建是困难的，特别是充斥着大量专业术语和专业知识的领域。因此，在构建本语料库的过程中遇到了不少问题，也总结了一些经验。

1） 医学领域语料库的构建涉及大量的专

业词汇，因此医学术语及名词资源的收集有助于界定实体边界和确定实体类型。

2）制定规范之前，需要充分考察语料涉及的范围，制定过程中需和医学专家紧密沟通，但也不能完全以医学专家主导，因为我们把握的标注粒度不同，比如医学专家对我们的实体类型病理学检查和内窥镜检查不认同，原因是内窥镜检查目的是进行病理学检查，可把两种类型合并，而在我们的标注中，这些是要区分开来的。

3）在正式标注之前需要对标注人员进行试标注，目的是训练标注人员以及逐步的完善规范。在正式标注过程中为了保证标注的质量，不确定的先不标注并记录下来，每周组织讨论会对这些问题进行讨论，积极听取医学专家的意见。

3 构建结果

3.1 标注语料质量分析

标注一致性(Inter-Annotator Agreement，IAA)是指两个独立标注人员标注结果达成一直的程度[16]，基于 IAA 指标评估手工标注数据的可靠性已经有了广泛的研究。在实体及实体关系手工标注语料库研究中，通常使用 F 值计算 IAA[18]。

具体做法是将最终标注结果 B 作为标准答案，计算首次标注结果 A 的精确度（P）和召回率（R），进而计算 F 值，计算公式如式（1）-（3）所示：

$$P = \frac{A\text{和B一致数目}}{\text{B的总数}} \quad (1)$$

$$R = \frac{A\text{和B的一致数目}}{\text{A的总数}} \quad (2)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (3)$$

确定实体一致性时，只有当实体文本、实体类型标签和起止位置均相同时认为实体标注是一致的；确定关系一致性时，只有关系类型以及关系中两个实体的类型和起止位置均相同时，认为关系标注是一致的。经统计，我们构建的语料库命名实体识别和实体关系一致率分别达到了 0.85 和 0.82，说明我们构建的语料库是可靠的[19]。对标注结果的命名实体数量和实体关系数量统计见**错误!未找到引用源。**。

为了方便以后的命名实体识别和实体关系抽取研究，我们把标注结果按照如下的方式存储。

标注样例：急性肾炎临床表现轻重悬殊，**咽炎为诱因**者**病前 6~12 天**多有**发热**。

命名实体标注结果的存储格式如下:

EN=急性肾炎 SO=2621 EO=2625 TP=疾病 EA=咽炎为诱因

EN(Entity Name)代表命名实体的名称，SO(Start Offset)代表实体在文档中的起始位置，EO(End Offset)代表实体在文档中的结束位置，TP(Type)代表实体的类型，EA(Entity Attribute)代表实体的属性或者限制条件。

实体关系标注结果的存储格式如下:

E1={EN=急性肾炎 SO=2621 EO=2625 TP=疾病 A=咽炎为诱因者} E2={EN=发热 SO=2656 EO=2658 TP=症状} RT=临床症状 RA=病前 6~12 天

RT(Relation Type)表示关系的类型，RA(Relation Attribute)表示关系的属性或者限制条件。

表 2 实体及实体关系类型数量统计

实体类型	数量	关系大类	数量
疾病	6,710	疾病-疾病	7,960
药物	1,550	疾病-药物	2,950
其他治疗	1,161	疾病-其他治疗	1,570
检查	1,599	疾病-检查	2,859
部位	670	疾病-部位	1,229
症状	6,662	疾病-症状	12,218
流行病学	1,004	疾病-流行病学	1,480
手术	402	疾病-手术	573
预后	366	疾病-预后	515

其他	419	疾病-其他	509
社会学	3,060	疾病-社会学	4,650
总计	23,603	总计	36,513

3.2 儿科知识图谱构建

由于语料库的数据源有多个，在不同数据源有对同一实体的不同表达，即使在同一个数据源里也可能存在这种情况，我们采用实体对齐的方法，对来自不同的数据源的实体在同一框架规范下进行异构数据整合，进而构建儿科医学知识图谱，其可视化展示如图 3 儿科知识图谱可视化

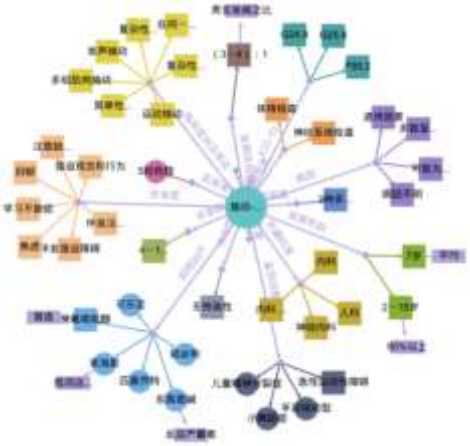

图 3 儿科知识图谱可视化

3.3 儿科知识问答

基于儿科知识图谱，开发了儿科知识问答系统，界面如图 5 所示。通过语义理解、规则匹配识别用户意图，从数据库中匹配正确答案返回给用户，同时能够给出相关的推荐。不同于传统的搜索引擎，基于儿科医学知识图谱的问答系统依赖于知识图谱中存储的知识。例如“唐氏综合征有什么症状？”，系统将在数据库中检索唐氏综合征的症状并将结果返回。

图 4 儿科知识问答系统

4 结语

本文主要以儿科学专业教材作为基础语料标注依据，介绍了语料库的标注体系和标注过程。语料的标注需要在标注规范的基础上进行，本文在中文电子病历医学命名实体和实体关系标注规范的基础上增加了实体类型，丰富了实体关系，由传统三元组扩展为六元组来进行标注，信息描述更为全面。目前，该语料库已完成字数 298 余万的标注工作，命名实体数目 23,603，实体关系数目达到 36,513。基于该语料库构建了儿科医学知识图谱，在此基础上开发了儿科医学知识问答系统。

人工标注的语料库被认为是金标准语料库，但构建规模庞大的语料库单单依靠人力是不切实际的，因此后续工作会尝试使用主动学习以及远程监督的方法来进行半自动化标注，提高标注效率。另一方面，也会继续完善基于儿科知识图谱的问答系统，为医生、患者和普通用户提供高质量的医疗咨询服务。

参考文献

[1] Liu Y, Yang LL, Xu SY, et al. Pediatrics in China: challenges and prospects[J]. World Journal of Pediatrics, February 2018, Volume 14, Issue 1, pp 1–3.

[2] Meystre S, Hang PJ. Natural language processing to extract medical problems from

electronic clinical documents: Performance evaluation.Journal of Biomedical Informatics, 2006, 39(6): 589-599.
[3] Campillos L , Louise Deléger, Grouin C , et al. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT)[J].Language Resources and Evaluation, 2018, 52(2):571-601.
[4] Roberts A, Gaizauskas R, Hepple M, et al. Building a semantically annotated corpus of clinical texts[J]. Journal of Biomedical Informatics, 2009, 42(5):950-966.
[5] Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization.Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing - BioTxtM2014. 2014:24-30
[6] Savova GK, Masanz JJ, Ogren PV, Zheng J, Solm S, Kipper-Schuler KC, Chute CG. Mayo clinical text analysis and knowledge extraction system(cTAKES): Architecture，component evaluation and applications. Journal of the American Medical Informatics Association, 2010, 17(5): 507-513.
[7] Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. A comprehensive study of named entity recognition in Chinese clinical text. Journal of the American Medical Informatics Association, 2014, 2I(5): 808-814.
[8] Wang Y, Yu Z, Chen L, Chert Y, Liu Y, Hu X, Jiang Y. Supervised methods for symptom flame recognition in free-text clinical records oftraditional Chinese medicine: An empirical study. Journal ofBiomedical Informatics，2014，47：91—104.
[9] 杨锦锋，关毅，何彬等. 中文电子病历命名实体和实体关系语料库构建. 软件学报，2016，27(11):2725-2746.
[10] 王卫平，孙锟，常立文. 儿科学.第 9 版[M]. 人民卫生出版社，2018.
[11] 沈晓明，桂永浩. 临床儿科学.第 2 版[M]. 人民卫生出版社，2013.
[12] Uzuner Ö, Mailoa J, Ryan RJ, et al. Semantic Relations for Problem-Oriented Medical Records. Artif Intell Med 2010;50:63–73.
[13] Xia, Fei & Yetisgen, Meliha. Clinical Corpus Annotation: Challenges and Strategies. In: Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining(BioTxtM2012)in conjunction with the International Conference on Language Resources andEvaluation(LREC), Istanbul, Turkey, 2012.
[14] Lipscomb C E . Medical Subject Headings (MeSH).[J]. Bull Med Libr Assoc, 2000, 88(3):265-266.
[15] Sundararajan V , Henderson T , Perry C , et al. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality[J]. Journal of Clinical Epidemiology, 2004, 57(12):0-1294.
[16] Nahler G. anatomical therapeutic chemical classification system (ATC)[M]. 2009.
[17] Hripcsak G, Rothschild A S. Agreement, the f-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association : JAMIA, 2005, 12(3): 296–298. [doi: 10.1197/jamia.M1733]
[18] Ogren P, Savova G, Chute C. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. In: Proceedings of the 12th World Congress on Health (Medical) Informatics. Marrakech, Morocco: European Language Resources Association (ELRA), 2008, 2325–2330.
[19] Artstein R, Poesio M.Inter-Coder Agreement for Computational Linguistics. Computational Linguistics, Dec. 2008, 34(4): 555–596.[doi:10.1162/coli.07-034-R2]

昝红英(1966-)，博士，教授，主要研究领域为自

昝红英(1966-)，博士，教授，主要研究领域为自然语言处理。

E-mail: iezanhy@zzu.edu.cn

刘涛(1996-)，通信作者，硕士，主要研究领域为自然语言处理。

E-mail: blalalt@163.com

陈俊富(1996-)，硕士，主要研究领域为自然语言处理。

E-mail: 893713591@qq.com