面向医学文本处理的医学实体标注规范

张欢[1,2]，**宗源**[1,2]，**常宝宝**[1,2]，**穗志方**[1,2]，
眷红英[2,3]，**张坤丽**[2,3]
（1. 北京大学计算语言学教育部重点实验室，北京100871；
2. 鹏城实验室，广东深圳518055
3. 郑州大学信息工程学院，河南郑州450001）

摘要

随着智慧医疗的普及，利用自然语言处理技术识别医学信息的需求日益增长。目前，针对医学实体而言，医学共享语料库仍处于空白状态，这对医学文本信息处理各项任务的进展造成了巨大阻力。如何判断不同的医学实体类别？如何界定不同实体间的涵盖范围？这些问题导致缺乏类似通用场景的大规模规范标注的医学文本数据。针对上述问题，该文参考了UMLS中定义的语义类型，提出面向医学文本信息处理的医学实体标注规范，涵盖了疾病、临床表现、医疗程序等9种医学实体，以及基于规范构建医学实体标注语料库。该文综述了标注规范的实体体系、标注细则、混淆处理、语料标注以及医学实体自动标注基线实验等相关问题，希望能为医学实体语料库的构建提供可参考的标注规范，以及为医学实体识别提供语料支持。

关键词： 智慧医疗；医学实体；标注规范；标注语料

Medical Entity Annotation Standard for Medical Text Processing

ZHANG Huan[1,2]，**ZONG Yuan**[1,2]，**CHANG Baobao**[1,2]，
SUI Zhifang[1,2]，**ZAN Hongying**[2,3]，**ZHANG Kunli**[2,3]
（1. Key Laboratory of Computational Linguistics，Ministry of Education，
Peking University，Beijing 100871，China；
2. Peng Cheng Laboratory，Shenzhen，Guangdong 518055，China；
3. School of Information Engineering，Zhengzhou University，
Zhengzhou，Henan 450001，China）

Abstract

With the popularization of smart healthcare, the demand of applying natural language processing technology to identify medical information is increasing day by day. At present, there is no unified annotation standard for medical named entities in China, and the medical shared corpus is still in a blank state, which causes great resistance to the progress of medical text information processing tasks. How to judge different categories of medical entities? How to define the coverage of different entities? These problems lead to the lack of a similar mass of general scenario standard of medical text data. In view of the above problems, We referred to the semantic types defined in UMLS and proposed a unified medical entity annotation standard for medical text processing, covering 9 kinds of medical entities such as disease, symptom, medical procedure and so on，and constructed medical entity annotated corpus based on standards. This paper summarizes related issues such as the entity system, annotation principles, obfuscation processing, corpus annotation process, and medical entity automatic labeling baseline experiments, hoping to provide reference for medical entity corpus build annotating standard, as well as the medical support the corpus entity recognition.

第十九届中国计算语言学大会论文集，第562页-第572页，海南，中国，2020年10月30日至11月1日。

Keywords: Smart healthcare , Medical entity , Annotation standard , Annotated corpus

1 引言

近年来，互联网和数字化已为众多行业带来颠覆性变革，医疗健康领域也不例外。伴随着智慧医疗的到来，在很大程度上改进了医院的管理及运营模式、改进了对大众的医疗服务。

医学领域存在大量自然语言文献，例如医学教材、医学百科、临床路径、病历、医学期刊、检验报告等，这些医学文本中蕴含了大量的专业知识和丰富的医学信息。医学领域中的命名实体识别指的是将重要的医学实体，如疾病、症状等从医学文本中抽取出来，这个步骤也是医学关系提取等各项任务的基础。

命名实体识别的主要技术方法分为：基于规则和词典的方法、基于统计的方法、二者混合的方法。基于规则和词典的方法是命名实体识别中最早使用的方法，但规则往往过于依赖知识库，故而充满局限性。基于统计的方法利用人工标注的语料进行训练，现已成为目前研究的主流方法。对于医学实体而言，医学共享语料库仍处于空白状态，这对医学文本信息处理各项任务的进展造成了巨大阻力。目前针对不同的标注任务，其医学实体标注规范各有不同，医学实体的分类也是大不相同。如何判断不同的医学实体类别？如何界定不同实体间的涵盖范围？这些问题导致缺乏类似通用场景的大规模规范标注的医学文本数据。因此亟需建立医学实体标注的规范，并以此建立医学实体标注语料库。

基于这样的前提，本文做出的主要贡献有：

（1）制定了面向医学文本信息处理的医学实体标注规范。以更加细致的划分方式将医学实体划分为九大类，包含“疾病”、“临床表现”、“医疗程序”、“医疗设备”、“药物”、“医学检验项目”、“身体”、“科室”和“微生物类”。

（2）基于规范进行了约263万字的儿科类医学教材的语料标注，在标注质量评测方面，采取迭代标注、抽样检查等多种措施，提高标注效率和标注质量。

（3）构建了一系列医学实体自动标注基线实验，取得了F值为71.0%的标注性能，为后续的工作打下了基础。

2 相关工作

一体化医学语言系统（Unified Medical Language System；UMLS）是美国国立医学图书馆(NLM)自1986年起研究和开发的一体化医学语言系统 (Bodenreider O, 2004)，是对生物医学科学领域内术语词表的统一汇编，并提供了UMLS数据库，如超级叙词表、语义网络和专家词典，以及相关软件工具，如MetamorphoSys、MMTx等。2010 i2b2/VA challenge (2010)会议发布了电子病历命名实体的分类，该会议参考UMLS定义的语义类型，将医学实体分为3种：医疗问题（Medical Problem）、检查（Test）和治疗（Treatment）。Roberts et al. (2007)等人随即选择了50份临床记录、X射线和病理报告进行标注，将这些医学文本中的医学实体分为6种：状况（Condition）、药物（Drug）、干预（Intervention）、部位（Locus）、检查（Investigation）、结果（Result）。South et al. (2009)使用了316例炎症性肠病的临床记录进行标注，其中医学实体种类分为4种：体征或症状（Signs or symptoms）、诊断（Diagnoses）、程序（Procedures）和药物（Meditations）。

相比国外对医学语料库的构建以及相关任务的展开，国内没有公开可获得的面向医学实体识别的数据集。2014年Lei et al. (2014)等人使用北京协和医院2013年的电子病历进行标注，其中医学实体分为4种：医疗问题、治疗程序、药物和检查。2014年，Xu et al. (2013)使用一家中国医院提供的336个个出院总结进行标注，其中医学实体同样分为4种：医疗问题、治疗程序、药物和检查。2016年，哈尔滨工业大学团队 (2016)使用来自哈尔滨医科大学附属第二医院的122个科室的电子病历进行标注，并且制定了新的中文电子病历命名实体和实体关系标注规范。该规范将医学实体分为5种：疾病、疾病诊断分类、症状、检查和治疗。2019年，Gao et al. (2019)等人使用了255份来自中国湖南省某著名医院的真实入院记录进行标注，并在其论文

中提出的新的标注方案，将医学实体分为9种：医疗发现、症状、时间词、疾病、身体部位、药物、治疗、实验室检查和（非实验室）检查。

本文旨在提出面向医学文本处理的医学实体标注规范，并将医学实体划分为九大类，包含“疾病”、“临床表现”、“医疗程序”、“医疗设备”、“药物”、“医学检验项目”、“身体”、“科室”和“微生物类”。回顾以上中英文医学文本语料库标注准则，大部分都将“检查”和“治疗”视为医学实体。本文认为，“检查”和“治疗”应该是作为医学实体之间的某种关系，比如“某种药物治疗某种疾病”、“使用某种设备检查身体部位”或“检查哪种检验项目”等等，从而建立“底层”独立实体之间的“高阶”关系，“实体”和“关系”是相互独立的，这样的优势在于让标注者能够更好地理解实体的概念。本文引入的“医疗程序”和“医疗设备”实体就是将“检查”和“治疗”更加“实体化”和“概念化”。一种医疗程序既可以作为治疗某种疾病的过程，也可以作为检查某种疾病是否存在的过程；一种医疗设备同样可以用来治疗疾病，也可以用来检查疾病或身体；等等。此时，疾病、医疗程序、医疗设备和身体作为“底层”实体，而检查和治疗作为“高阶”关系，这样使标注者能够真正建立起医学的概念，在医学实体标注阶段仅仅集中研究实体本身的含义，而不关注实体之间的关系。这种实体的细致划分一方面是为了概念细粒度化，另一方面为后续的实体关系提取提供了良好的数据基础。

3 总体原则

3.1 简单性原则

本文将医学实体划分为九大类，包括“疾病”、“临床表现”、“医疗程序”、“医疗设备”、“药物”、“医学检验项目”、“身体”、“科室”和“微生物类”，并详细介绍了各个医学实体的涵盖范畴，阐述实体间的混淆处理，用大量示例举例说明。标注者无需太多专业知识，实体类别定义简易明了，方便标注者理解和区分。

3.2 易操作性原则

对于医学文本的医学实体标注，标注者需严格遵从本文提出的医学实体标注规范，使用“[named-entity]tag”的方式进行紧密标记（左右括号与标记实体首末字符之间无空格，括号需成对出现），若出现标注实体的英文缩写、中文简称或者俗称，均需要标注，各类医学实体标签如表1中所示。可嵌套标注的实体内部包含的实体应作为方括号嵌套成分，如“[[named-entity]tag XXXXX]tag”。

医学实体类别	标签/tag	备注
疾病	dis	disease
临床表现	sym	symptom
身体	bod	body
医疗程序	pro	procedure
医疗设备	equ	equipment
药物	dru	drug
医学检验项目	ite	item
科室	dep	department
微生物类	mic	microbes

Table 1: 医学实体标记

3.3 一致性原则

医学实体的标注包括实体类型和实体边界两个部分。医学实体的分词存在很多歧义，如何切分较长的疾病或药物名称等，这给标注工作带来了很多困难。对于除“临床表现”这个复杂的医学实体外，人们往往关注的是医学实体的含义，比如什么疾病、什么药物、哪个科室等等，这类医学实体内部无需分词，仅仅作为一个整体来看待。因此本文遵从以下统一原则：

1. “临床表现”实体内部允许分词，并且该实体内部允许嵌套标注，即若“临床表现”实体内部存在其他8种实体，标注者也应该将其标注出来，“临床表现”实体内部其他文本内容的分词

原则同3处理。

2. 除“临床表现”外的医学实体内部不允许分词。标注应当遵循“最大单位标注法”，即若一个实体内包含其他实体，则标注“最大”的实体，不做嵌套标注。.
3. 除以上实体之外的其他文本内容的分词原则遵从《北京大学现代汉语语料库基本加工规范（2002版）》 (2002)

另外，为保证实体意义的完整性和可理解性，所有的实体可以是一个词、短语和句子，实体中可包含标点符号。也就是说，当标点符号存在某种意义时，标注者同样应该将其标注。

4 医学实体体系

本文将医学实体划分为九大类，包含“疾病”、“临床表现”、“医疗程序”、“医疗设备”、“药物”、“医学检验项目”、“身体”、“科室”和“微生物类”，如图1所示。故本文提出的规范对于医学实体的划分上更加细致，这也有便于未来医学实体关系提取等各项工作的开展。本文借鉴UMLS语义类型界定实体涵盖的范围，但不局限于UMLS的定义。

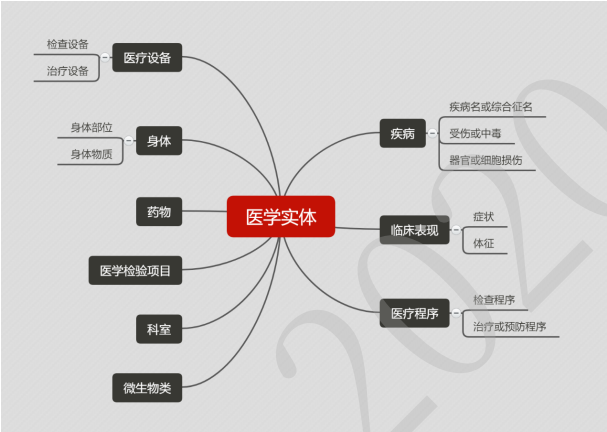

Figure 1: 医学实体架构图

第一类实体是疾病，疾病是指导致病人处于非健康状态的原因或者医生对病人做出的诊断，并且是能够被治疗的 (2016)。包括疾病或综合征、中毒或受伤、器官或细胞受损，其对应的UMLS语义类型有疾病或者综合征（disease or syndrome）、中毒或受伤（injury or poisoning）等；

第二类实体是临床表现，临床表现是疾病的表现，泛指患者不适感觉以及通过检查得知的异常表现。主要包括症状、体征，其对应的UMLS语义类型有症状或体征（sign or symptom）、异常检查结果（abnormal test results）等；

第三类实体是医疗程序，在本文中，医疗程序泛指为诊断或治疗所采取的措施、方法及过程。主要包括检查程序、治疗或预防程序，其对应的UMLS语义类型有化验过程（laboratory procedure）、治疗或预防过程（therapeutic or preventive procedure）、等；

第四类实体是医疗设备，在本文中，医疗设备泛指为诊断或治疗所使用的工具、器具、仪器等。主要包括检查设备、治疗设备，其对应的UMLS语义类型有医疗设备（medical device）、药物传输设备（drug delivery device）等；

第五类实体是药物，药物是指用来预防、治疗及诊断疾病的物质，其对应的UMLS语义类型有临床药物（clinical drug）、抗生素（antibiotic）等；

第六类实体是医学检验项目，医学检验项目是指检查涉及到的体液检查项目、重要生理指标以及其他检查项目，本文规定“医疗检验项目”主要针对人体而言，是能够通过设备或实验检测出的项目，并且是能够被量化，有其对应的测量值或指标值。其对应的UMLS语义类型有实验室检查（laboratory test）等；

第七类实体是身体，身体泛指细胞、组织、及位于人体特定区域的由细小物质成分组合而成的结构、器官、系统、肢体，另外包括身体产生或解剖身体产生的物质等。主要包括身体部位、身体物质，其对应的UMLS语义类型有身体部位（body part）、组织（organ）、组织成分（organ component）等；

第十九届中国计算语言学大会论文集，第562页-第572页，海南，中国，2020年10月30日至11月1日。

第八类实体是科室，科室主要是指医院或医疗机构所设有的科室其对应的UMLS语义类型有医疗保健相关组织（healthcare related organization）等;

第九类实体是微生物类，微生物类包括细菌、病毒、真菌以及一些小型的原生生物、显微藻类等在内的一大类生物群体，另外包括微生物类产生的毒素、激素、酶等，其对应的UMLS语义类型有细菌（virus）、真菌（fungus）等;

5 医学实体标注细则

5.1 疾病

5.1.1 疾病或综合征

疾病或综合征是指疾病或综合征名称。比如高血压、肺炎、心脏病、败血症、畸形等。

① 测定 结果 的 分析： [肺炎]dis 、 [败血症]dis 、 严重 [先天性心脏病]dis 或 [畸形]dis 影响 新生儿 代谢 和 循环 功能， 特别 是 [严重 感染]sym 时， 可 [微循环障碍]dis 和 [DIC]dis

5.1.2 受伤或中毒

患者在受伤或中毒后，对人体造成某种伤害，导致患者处于非健康状态。

① [局部灼伤]dis 处理 与 一般 [烫伤]dis 处理 相同

5.1.3 器官或细胞损伤

器官、细胞等发生异常或损伤后，如果能够危及人的机体，此时虽然它们属于身体的一部分，但是已成为一种致病因素，危害人体健康。

① 还有 [颅内出血]dis （ [产伤]dis 、 [外伤]dis ）， [颅脑损伤]dis ， [脑血管畸形]dis 。

在标注“疾病”实体时，需要注意:

（1）一般有些疾病名称很长，前面会有“XX性”、“XX状”、“XX型”等，以及身体部位（一个或多个）的修饰，在保证疾病完整性和具体性的情况下，在标注时应该与这些前缀一起标注。

① 有些 [急性病毒感染]dis 引起 明确 的 相关性 疾病， 如 [HAV感染]dis 与 [急性黄疸型肝炎]dis ； [轮状病毒感染]dis 与 [季节性婴幼儿腹泻]dis

③ 小儿 患有 [肝脏、肾脏、甲状腺疾病]dis

（2）大部分统称均标注，比如营养性疾病、代谢性疾病、化脓性和非化脓性综合征等，此类虽然是统称，但是有对应的疾病范畴，所以应该标注。特殊情况：像常见病、多发病、疾病等单独出现时，此类统称范围太广，不应标注。

① [高血压]dis 是 严重 危害 人类 健康 的 常见病 、 多发病 （3）当疾病有若干种分型时，“疾病+分型”或“分型+疾病”整体标注，分型单独出现不标注。

① 可将 [MPS]dis 分为 I ～ VII 型， 除 [MPSII型]dis 为 X 连锁 隐性 遗传 外

5.2 临床表现

5.2.1 症状

症状是指病人自己向医生陈述（或是别人代述）的不适或痛苦表现。通常是病人主观感觉的不适，如腹痛、头晕等，或是自己发现的病理改变，如血尿便血、活动障碍等。

5.2.2 体征

体征是指医生观察到的或者通过检查程序或设备检查到的发生于病人的异常变化以及异常检查结果。通常是指医师利用自己的感官（视触叩听）或者医疗器具（血压计叩诊锤等医疗设备）发现的病人的病理生理变化。

在标注“临床表现”实体时，需要注意:

（1）在“临床表现”实体内部，若包含除“临床表现”之外的其他实体（“疾病”、“身体”、“医学检验项目”等），内部实体应作为方括号嵌套成分（如“[[肺动脉干]bod 突出]sym”）。

① 如 出现 [[肺动脉]bod 高压]sym ， [[肺动脉干]bod 突出]sym

（2）如果前后文中有“表现为”、“表现有”、“有”、“不良反应有”、“症状有”、“等症状”、“反应”等描述症状出现的词，则标注对应实体为“临床表现”。

① 不良 反应 有 [[气管]bod 痉挛]sym 、 [[心功能]ite 不全]sym 、 [恶心]sym 、 [呕吐]sym

（3）对于“临床表现”的修成成分，通常表示其严重程度、频率等，为保证标注完整性，在标注时应该将修饰和症状一起标注。

① 遇 紧急 情况 ， [气管]bod 异物 导致 [严重 呼吸 困难]sym

（4）对于“体征”，一般都是通过医疗设备观察到的病理或生理改变等客观表现，因此多出现在表示检查的词后面，比如“见”，“可见”、“及”、“闻及”、“显示”等，这里有两种情况：一种是“医疗程序/医疗设备+（检查词）+体征”，则在标注时仅将对应的体征标注为“临床表现”，不用标注此类检查词；另一种是“身体部位/主体+（检查词）+体征”，为保证标注意义完整，则在标注时应该将“身体部位/主体+（检查词）+体征”作为整体标注成“临床表现”。

① 行 [头颅CT]pro 显示 [[双侧额部]bod 或 [额顶部]bod 有 [蛛网膜下腔]bod 增宽]sym

（5）描述“临床表现”时，通常是对病人（一个或多个）身体部位进行描述，在标注时应该与这些身体部位一起标注。若出现“部位/主体+有/无+临床表现”，应该整体标注为“临床表现”。若出现“无+症状/体征”，“无”作为描述临床表现的一种修饰成分，应该整体标注为“临床表现”。

① 常 伴有 [[膀胱逼尿肌]bod 无 抑制性 收缩]sym ， 其中 25 % 患儿 有 [尿失禁]sym

（6）如果临床表现实体后面紧跟“症状”、“体征”、“反应”，此时应该将实体和此类词语整体标注；若是非临床表现实体后面紧跟此类词语，则不标注；若单独出现“症状”、“体征”、“反应”表示的是一种统称含义，则不标注。

① [[中毒]dis 症状]sym 与 [[颅高压]dis 征象]sym 明显 、 [[神经系统]bod 局灶 定位 体征]sym 出现 ， [神经影像学检查]pro 帮助 诊断 。

（7）临床上的“指征”，一般是指手术指征。在标注工作中，指征标注为“临床表现”。

① 临床 指征 为 ： [血便]sym ； 有 [里急后重]sym ；

5.3 医疗程序

“医疗程序”泛指为诊断或治疗所采取的措施、方法等，包括检查程序和预防或治疗程序。

5.3.1 检查程序

检查程序包括通用检查方法、专项检查、医学影像检查等，检查方法是医生为达到化验目的而采取的某种手段；专项检查是病人通常情况下做的某种检查；医学影像检查是放射科或核医学部门的医疗程序。

① [肝活检]pro 应 争取 在 起病 后 4 ～ 5 日 内 进行

5.3.2 治疗程序

治疗或预防程序是医生为达到治疗目的而采取的某种手段，如化疗、放疗、手术、透析、紧急救治等。

① [静脉注射]pro 用 [丙种球蛋白]dru （ [IVIG]dru ） 对 部分 [狼疮]dis 有 一定 疗效

在标注“医疗程序”实体时，需要注意：

（1）当医疗程序前面有身体部位（对某部位进行检查或治疗），应该整体标注。

① 通常 需 经 [胸部X线平片]pro 进行 诊断

（2）辅助治疗和非药物治疗也标注为“医疗程序”。

5.4 医疗设备

“医疗设备”泛指为诊断或治疗所使用的器具、或仪器等，包括检查设备、治疗设备。

5.4.1 检查设备

检查设备通常指的是医院中用来检查或检验的仪器，比如血细胞分析仪、生化分析仪等。

① 通过 [血细胞分离仪]equ 可 分离 得 [白细胞]bod

5.4.2 治疗设备

治疗设备是医生为达到治疗目的而单独或者组合使用于人体的仪器、设备、器具，如注射器、供养面罩、呼吸器等。

① [血管内支架]equ 在 [先天性心脏病]dis 中 的 应用 ： 常用 [4通道测压导管]equ 。

在标注“医疗设备”实体时，需要注意：

（1）医疗设备的属性不可标注为“医疗设备”。

① 适用 的 [注射器]equ 范围 大 ， 包括 10 ～ 50 ml [注射器]equ 均 可 适用 。

5.5 药物

药物广义上是指用来预防、治疗及诊断疾病的物质，另外也包括临床诊断试剂。
在标注“药物”实体时，需要注意：：
（1）药物的属性不可标注为“药物”。

① 可用 [9α-氟氢可的松]dru 0.05 ～ 0.1 mg / d

（2）大部分药物的统称均标注，比如营养素、抗菌药物、急救药物等，此类虽然是统称，但是有对应的治疗范畴。像常用药、药物等单独出现时，此类统称范围太广，不应标注。

① 配备 [急救药物]dru 可 增加 安全性 。

5.6 医学检验项目

检查涉及到的体液检查项目、重要生理指标以及其他检查项目，本文规定“医疗检验项目”主要针对人体而言，是能够通过设备或实验检测出的项目，并且是能够被量化，有其对应的测量值或指标值。

在标注“医学检验项目”实体时，需要注意:
（1）测量值不应该标注为“医学检验项目”。

① [输血]pro 指征 ： [[心率]ite ＞ 110 次 / 分]sym （“＞110次/分”不标注）

（2）上下文有关键的提示文字的检查：“检查”、“检”、“查”、“测定”、“检验”等，其对应的检查项目标注为“医学检验项目”。

① 对 [腹泻]sym 较 重 的 患儿 ， 应 及时 检查 [血pH]ite 、 [二氧化碳结合力]ite 、 [碳酸氢根]ite 、 [血钠]ite 、 [血钾]ite 、 [血氯]ite 及 [血渗透压]ite 。（3）在“临床表现”中，有时会涉及到对医学检验项目的描述，比如一些生理指标等，标注时应该将这些生理指标嵌套标注为“医学检验项目”。

① 患者 的 临床 症状 有所 好转 ， [[血清CK]ite 下降]sym

（4）常规的检查项目：尿常规、血常规等标本采集类项目，应该标注为“医疗检验项目”。

5.7 身体

“身体”身体泛指细胞、组织、及位于人体特定区域的由细小物质成分组合而成的结构、器官、系统、肢体，另外包括身体产生或解剖身体产生的物质等，包括身体部位和身体物质。

5.7.1 身体部位

身体部位包括器官或器官组成、身体系统、身体位置或区域。

① 常 见于 尿布 包裹 处 及 光滑 [皮肤]bod 相互 直接 摩擦 部位 ， 如 [肛周]bod 、 [臀部]bod 、 [外阴]bod 、 [腹股沟]bod 以及 [腋窝]bod 、 [下颏]bod 等 处 [皮肤]bod

5.7.2 身体物质

身体物质可由组织、细胞（包括细胞成分、细胞结构）、生物大分子、以及身体或解剖身体产生的物质。

在标注“身体”实体时，需要注意:
（1） 表 示 身 体 部 位 具 体 位 置 的 方 位 词 或 者 数 量 词 ，如“上”、“下”、“左”、“右”、“部”、“侧”、“双”、“多”等应当一同标注。

① 尤 以 [双下肺]bod 明显 ， 严重 病例 可 合并 [[胸腔]bod 积液]sym 或 [脓胸]sym 。

（2）当出现多个身体部位的组合，分开标注会失去原本含义，应作为整体标注为“身体”。

① [指、趾甲]bod

（3）病变细胞亦标注为“身体”

① [肿瘤细胞]bod 、 [白血病细胞]bod 、 [狼疮细胞]bod

5.8 科室

“科室”主要是指医院或医疗机构所设有的部门以及科室。

① [外科]dep [血液/肿瘤科]dep 和 [放疗科]dep 为 基本 组成 单位

第十九届中国计算语言学大会论文集，第562页-第572页，海南，中国，2020年10月30日至11月1日。

5.9 微生物类

微生物类包括细菌、病毒、真菌以及一些小型的原生生物、显微藻类等在内的一大类生物群体。微生物类亦可分为致病微生物和非致病性微生物。

① 其他 [细菌]mic 也 可 产生 [肠毒素]mic ， 如 [耶尔森菌]mic 、 [鼠伤寒沙门菌]mic

在标注“微生物类”实体时，需要注意：

（1）出现“病毒+病毒的一部分（基因、DNA）”应该整体标注为“微生物类”。

① 此法 能 发现 不 完整 [病毒]mic 如 潜伏 [病毒DNA]mic

6 分类混淆处理

6.1 疾病(dis)和临床表现(sym)

疾病和临床表现的最大区别是：疾病是通过鉴别诊断的，疾病实质上就是身体受损；而临床表现实质上是身体受损后所表现出来的现象，比如说病人的不适感觉、身体出现的异常变化，但是这些往往是病人或者医生看到的表面现象。而作为医生则需要通过进一步的鉴别诊断来确认病人所患疾病，这也就说明疾病和临床表现存在着本质性的差异。

在遇到“感染”相关的实体时，有以下几点需要注意一下：

（1）若出现明确致病原因（病毒、细菌或身体部位等名称）与“感染”组合成词，则整体标注为“疾病”。如[HAV感染]dis、[球菌感染]dis、[上呼吸道感染]dis。

（2）当单独出现“感染”一词时，若上下文明显表示是对某种疾病的指代，则标注为疾病，否则标注为“临床表现”

（3）若“感染”一词前面的修饰词表明程度或频率时，则整体标注为“临床表现”

6.2 医疗程序(pro)和医疗设备(equ)

医疗程序是指检查过程以及预防或治疗过程，描述的是医生为诊断或者治疗而采取的一系列操作或过程。而医疗设备是指诊断或者治疗而使用的设备，描述的是具体的设备实体（工具、器具、仪器或机器），是医生进行诊断或者治疗的工具。标注者在标注时应该谨慎区分。

① 临床 发现 有些 [头颅]bod 较 大 的 婴儿 ， 行 [头颅CT]pro 和 [MRI检查]pro

6.3 药物(dru)和身体(bod)

标注实体是某种身体物质时，但是有“口服”、“注射”等字眼显式地表明该实体是一种药物（是外来的），此时应该将该类实体标注为“药物”。否则，如果只是表明该实体在人身体中的一种状态（是内在的），应该标注为“身体”。

① [糖尿病]dis 患儿 由于 [[胰岛素]bod 分泌 不足 或 缺如]sym

② 纠正 [高血钾]dis ： [葡萄糖]dru 0.5 g / kg 加 [胰岛素]dru 0.3 U / kg [静滴]pro

6.4 医学检验项目(ite)和医疗程序(pro)

医学检验项目是指体液检查项目、生理测量、重要生理指标以及其他检查项目，通常是名词。但如果医学检验项目名称后面紧跟着“检查”、“测定”、“诊断”、“分析”等，表明是医生为诊断或者治疗而采取的一系列操作或过程，故应将医学检验项目名称和这些词语作为整体一起作为标注为“医疗程序”。

① 监护 包括 [脉搏]ite 、 [血压]ite 、 [尿量]ite 、 [血乳酸含量测定]pro 和 [血气分析]pro

6.5 医学检验项目(ite)和身体(bod)

标注实体是某种身体物质时，但是在检查中涉及到对该实体的指标限定，显式地表明该类实体应该是某种检查项目，并且后面通常紧跟着测量值或者指标值，此时应该将该实体标注为“医学检查项目”。否则，如果只是表明该实体在人身体中的一种状态，则标注为“身体”。

① [输血]pro 指征 ： [[心率]ite ＞ 110 次 / 分]sym ； [[红细胞]ite ＜ 3 × 1012 / L]sym

7 医学实体语料标注

基于前文提出的医学实体体系和标注细则，本文制定了完整的医学实体标注规范。为检验规范的可行性、为医学实体识别提供语料支持，我们选择了约263万字的儿科类教材进行医学实体标注，选择儿科类的原因是儿科实质上是全科医学，医学知识涵盖范围广，具有代表性。

第十九届中国计算语言学大会论文集，第562页-第572页，海南，中国，2020年10月30日至11月1日。

7.1 语料标注过程

医学实体标注规范的制定难度较大，不仅涉及专业的医疗知识,而且涉及到对医学实体的定义和分类。我们制定出初步规范,然后采用多轮迭代的模式进行规范的修订和标注工作。主要分为三个阶段来进行：

在第一阶段，组织标注人员学习本规范，组织标注人员预标注，目的在于熟悉医学实体标注规范，以及收集在实际标注医学文本中发生的问题。两轮预标注后，经过与医学专家讨论，进一步对标注规范进行完善，使标注规范更贴近本次研究任务，为正式标注打下坚实基础。

第二个阶段，在标注平台上利用第一阶段形成的医学实体资源库进行医学实体标注。标注过程采取多轮迭代模式，即每个医学文本由两名标注人员负责。一标者完成标注任务后，记录存在疑问的地方，接着由二标人负责检查并记录下不一致和不确定的地方。与医学专家商量讨论后获得统一的解决方案。讨论之后再由一标者负责修改标注，形成最后的三标文件。在这个阶段，会根据标注人员标注时的反馈意见修改标注规范，使标注规范更加适用于医学文本。

第三阶段进行分词的校对。开发可以修改分词的标注工具，标注人员在新的标注平台上修改分词错误以及检查实体标注情况是否符合目前已经更新的规范，同时查看是否存在实体缺失、错位等问题，提升标注质量。

7.2 医学实体分布统计

鉴于标注语料中同一类别的重复实体较多，我们从例数和型数两个方面对各个类别的医学实体数量进行统计，其中例数包含同一类别的重复实体，型数不包含同一类别的重复实体。

首先，我们对标注完成的所有医学实体数量进行了统计,见表2。根据统计显示，在这9种医学实体中，“临床表现”实体总数最多，“疾病”实体次之；“科室”实体总数最少。其次，除了“临床表现”实体外，其他实体均不含嵌套实体，关于嵌套实体统计如表3，其中“嵌套类临床表现”实体型数占“临床表现”实体型数约三分之一。

医学实体类型	例数	型数
疾病	28913	10494
临床表现	22989	14482
身体	27078	7223
医疗程序	11545	5095
医疗设备	1836	851
医学检验项目	4570	1935
药物	7549	2714
科室	574	112
微生物类	3863	1036
总计	109097	43942

Table 2: 语料实体类型与数量统计表

医学实体类型	例数	型数
临床表现	22989	14482
嵌套类临床表现	5375	4749

Table 3: 临床表现及嵌套实体统计表

7.3 医学实体自动标注实验

我们将整个标注数据按照8:1:1的比例随即划分为训练集、验证集和测试集，并统计了对应集合的实体数量分布，如表4所示，例数为各类中包含重复实体的总数，型数为各类中不包含重复实体的总数。我们在数据集上展开一系列的基线实验，在实验中，为体现整体效果，我们将嵌套类临床表现均视为整个临床表现实体来对待。基线实验如下：

CRF：CRF通过引入自定义的特征函数，不仅可以表达观测之间的依赖，还可表示当前观测与前后多个状态之间的复杂依赖。本模型使用：“前一个词，当前词，后一个词；前一个词+当前词，当前词+后一个词”作为特征。

BiLSTM-CRF：在训练过程中，LSTM能够根据识别实体自动提取观测序列的特征，但是缺点是无法学习到状态序列（输出的标注）之间的关系。CRF的优点就是能对隐含状态建模，学习状态序列的特点。所以在LSTM后面再加一层CRF，以获得两者的优点。

医学BERT-BiLSTM-CRF：使用经大量医学文本预训练好的BERT模型作为预训练模型，将BERT预训练的输出输入到一个双向的LSTM网络，在双向的LSTM网络上层再叠加一个CRF层，能够对标签信息加以利用，最终得到输出标签序列。

医学实体类型	Train		Dev		Test	
	例数	型数	例数	型数	例数	型数
疾病	23297	8970	2794	1705	2822	1740
临床表现	18544	11892	2241	1787	2204	1792
身体	21887	6182	2720	1353	2471	1213
医疗程序	9315	4313	1071	763	1159	778
医疗设备	1500	717	139	98	197	130
医学检验项目	3748	1674	435	262	567	342
药物	6216	2375	639	441	694	433
微生物类	3214	925	342	150	307	163
科室	459	94	80	34	35	13
总计	88180	37142	10461	6593	10456	6604

Table 4: 实验数据统计

7.3.1 自动标注实验结果分析

不同模型的结果如表5所示。标注的数据集包含9种医学实体，涵盖广、类别多；对于所有的医学实体而言，同时存在较长或较短文本。以上存在的两个问题，给识别任务增加了一定难度，故准确率会明显低于召回率。

就医学Bert-BiLSTM-CRF模型而言，表6为分类统计的结果，可以看出“疾病”、“药物”实体的识别能力较强。结合数据特点和结果，存在以下几个问题：部分医学实体存在“碰撞”问题，也就是说相同实体在不同情况下会具有不同实体类别，比如“疾病”和“临床表现”，这在一定程度上影响了模型的识别能力；“临床表现”实体普遍较长，模型的识别能力有待提高；“科室”实体较少，模型识别能力不强；在基线实验中，嵌套类临床表现均视为整个临床表现实体，并未识别内部嵌套的实体。这些问题对未来的工作提出了新的难题和挑战。

Model	P(%)	R(%)	F (%)
CRF	62.05	55.94	58.84
BiLSTM-CRF	63.56	56.58	59.87
医学Bert-BiLSTM-CRF	69.6	72.5	71.0

Table 5: 实验结果展示

医学实体类型	P(%)	R(%)	F (%)
疾病	79.4	82.3	80.8
临床表现	56.9	61.2	59.0
身体	68.1	71.5	69.8
医疗程序	66.8	70.1	68.5
医疗设备	70.7	71.9	71.3
药物	83.3	85.7	84.5
医学检验项目	61.5	59.5	60.5
科室	61.1	66.7	63.8
微生物类	76.4	71.3	73.8
总体	69.6	72.5	71.0

Table 6: 各类医学实体的具体结果

8 结束语

本文提出的面向医学文本信息处理的医学实体标注规范，详细介绍了各个医学实体的涵盖范畴，阐述实体间的混淆处理，用大量示例举例说明，适用于多种类型的医学文本。并且详细描述了医学语料标注任务的过程以及基线实验。基于本规范我们将发布一定规模的标注样例，希望能为医学文本信息处理提供最基础的数据资源。

致谢

本文工作得到国家重点研发项目(2018AAA0102003)，特此致谢。

参考文献

2010 i2b2/VA challenge. 2010. *https://www. i2b2.org/NLP/Relations/assets/Concept Annotation Guideline.pdf.*

Bodenreider O. 2004. Nucleic Acids Res. *The Unified Medical Language System (UMLS):integrating biomedical terminology*, 2004(32): 267-270.

Gao, Yao Gu, Lei Wang, Yefeng Wang, Yandong Yang, Feng. 2019. BMC Medical Informatics and Decision Making. *Constructing a Chinese electronic medical record corpus for named entity recognition on resident admit notes*, 19(56):67-78.

Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H. 2014. J Am Med Inf Assoc. *A comprehensive study of named entity recognition in Chinese clinical text*, 21(5):808–14.

Roberts A, Gaizauskas R, Hepple M, Davis N. 2007. J Am Med Inf Assoc. *A comprehensive study of named entity recognition in Chinese clinical text*, 21(5):808–14.

South BR, Shen S, Jones M, Garvin J, Samore MH, Chapman WW, et al. 2009. BMC Bioinformatics. *Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease*, 10(12):1-32.

Xu, Yan Wang, Yining Liu, Tianren Liu, Jiahua Fan, Yubo Qian, Yi Tsujii, Jun'ichi Chang, Eric. 2013. Journal of the American Medical Informatics Association. *Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries*, 2013(21):84-92.

杨锦锋, 关毅, 何彬, 曲春燕, 于秋滨, 刘雅欣, 赵永杰. 2016. 软件学报. 中文电子病历命名实体和实体关系语料库构建, 27(11):27252746.

YANG Jin-Feng, GUAN Yi, HE Bin, QU Chun-Yan, YU Qiu-Bin, LIU Ya-Xin, ZHAO Yong-Jie. 2016. Journal of Software. *Corpus Construction for Named Entities and Entity Relations on Chinese Electronic Medical Records*, 27(11):2725-2746.

俞士汶, 段慧明, 朱学锋, 孙斌. 2002. 中文信息学报. 北京大学现代汉语语料库基本加工规范, 16(5):51-66.