DOI：10.61189/409428oucija

·论著·

面向专病电子病历的实体语料库构建方法

陈思旭[1]，刘独玉[1*]，谭小琴[1]，齐　星[1]，罗　彬[2]
1. 西南民族大学电气工程学院，成都 610041
2. 四川互慧软件有限公司，绵阳 621000

［摘要］ 本研究针对医疗领域中电子病历命名实体识别任务资源匮乏问题，在医学专家的指导下制定了统一的专病实体语料库标注方法，并构建了2种专病实体语料库——儿童支气管肺炎实体语料库和糖尿病实体语料库。在BERT-BiLSTM-CRF和ERNIE-BiLSTM-CRF模型上，将儿童支气管肺炎实体语料库与公开数据集进行比较，验证本文提出的专病实体语料库标注方法的有效性；再将专病实体语料库标注方法重新应用于糖尿病电子病历，以评价模型鲁棒性。模型验证结果显示：2种自建专病实体语料库的*F1*值均优于公开数据集，说明本文提出的专病实体语料库标注方法的鲁棒性。

［关键词］ 电子病历；命名实体识别；语料库构建；儿童支气管肺炎实体语料库；糖尿病实体语料库
［中图分类号］ R-3/TP391　　［文献标志码］ A

Construction methodology of entity corpus for special diseases electronic medical records

CHEN Sixu[1], LIU Duyu[1*], TAN Xiaoqin[1], QI Xing[1], LUO Bin[2]
1. College of Electrical Engineering, Southwest Minzu University, Chengdu 610041, Sichuan, China
2. Sichuan Huhui Software Co.,Ltd., Mianyang 621000, Sichuan, China

［**Abstract**］ Addressing the issue of resource scarcity for named entity recognition tasks in the medical field, a unified annotation methodology for special diseases entity corpora was formulated under the guidance of medical experts, and two special diseases entity corpora were constructed, namely Pediatric Bronchopneumonia Entity Corpus and Diabetes Entity Corpus. To verify the effectiveness of the proposed special disease entity corpus annotation method, the Pediatric Bronchopneumonia Entity Corpus was first compared with the publicly available dataset using BERT-BiLSTM-CRF and ERNIE-BiLSTM-CRF models. Then, the methodology was reapplied to diabetes electronic medical records to evaluate the robustness of the model. The results showed that both special diseases entity corpora got higher *F1* scores than the public datasets, which suggests that special diseases entity corpus annotation methodology proposed in this paper has good robustness.

［**Key Words**］ electronic medical record; named entity recognition; corpus construction; Pediatric Bronchopneumonia Entity Corpus; Diabetes Entity Corpus

电子病历是医务人员在医疗活动过程中，使用医疗信息系统生成的文字、图表、影像、数据等，并实现存储、管理、传输和重现的数字化医疗信息[1]。其中的文字记录大多是非结构化文本，如病程记录、出院小结等，蕴含丰富的医疗信息，能够帮助医生更好地了解患者情况，辅助临床诊断和决策。如果依靠人工分析电子病历需要耗费大量时间，目前常用自然语言处理（natural language processing，NLP）相关技术提取电子病历非结构化信息，以获得与患者相关的医学信息。

在NLP研究中，命名实体识别（named entity recognition，NER）是由Rau等[2]提出的一项基本信息提取任务，主要目的是从句子中提取预定义的特定实体并识别其正确类型，例如电子病历中的疾病、检查、药物名称等。NER模型构建需依赖标注语料作为训练数据，并采用监督或半监督学习方法，所以NER模型的效果及性能与标注语料库的规模和质量密切相关。

由于电子病历对标注人员的专业性要求较高，使得中文医学共享语料库资源仍十分匮乏，并且标

［收稿日期］ 2024-07-07　　［接受日期］ 2024-09-02
［基金项目］ 国家重点研发计划（2021YFF0704100），西南民族大学中央高校基本科研业务费专项资金（2023NYXXS016）. Supported by State Key Research and Development Project（2021YFF0704100），the Fundamental Research Funds for the Central Universities，Southwest Minzu University（2023NYXXS016）.
［作者简介］ 陈思旭，硕士研究生. E-mail：chen1296647721@163.com
*通信作者（Corresponding author）. Tel:15828397145, E-mail: liuduyu10000@163.com

注存在一定不足。因此,本文旨在构建专病电子病历实体语料库,期待能够提高NER任务的性能,并为医学领域的信息抽取和知识发现提供更丰富和深入的数据资源。

1 相关研究

目前,国外的医学电子病历NER语料库构建已取得一定进展。Meystre[3]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均标注其修饰词信息。2008年,美国梅奥诊所[4]构建了包含160份文档的命名实体语料,并首次对实体和实体关系的修饰信息进行细致分类。Roberts等[5]于2009年首次将实体类型扩展至6类;Campillos等[6]构建了法语命名实体及实体关系语料库;Byun等[7]利用ChatGPT构建了韩国第1个开源生物医学语料库(KBMC)。语料库由6 150个句子组成,包括4 162个疾病名称、841个身体部位和396个治疗方法,使医疗NER性能提高20%。

中文NER语料库也取得许多成果。Lei等[8]参考2010年Informatics for Integrating Biology & the Bedside(I2B2)的实体分类,将治疗细分为药物和过程,在2名医生的参与下构建了包含800份文档的命名实体语料库。杨锦锋等[9]结合中文电子病历的特点,构建了包含922份病历文本的标注语料库,命名实体的标注一致性(inter-annotation agreement,IAA)达到0.922。昝红英等[10]在医学专家的指导下,制定了针对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,共标注23 603个命名实体和36 513个实体关系。常洪阳等[11]构建脑卒中专病电子病历实体及实体关系标注语料库(SEMRC),实体名标注一致率达到85.16%,且可以运用NLP技术实现命名实体信息提取。巩敦卫等[12]将500份糖尿病患者的中文电子病历中的实体划分为疾病、症状、身体部位、药品和检查操作5类,在BiLSTM-CRF中融入字词形嵌入和注意力机制,*F1*值达到97.54%。屈丹丹等[13]结合中医医案特点,将国医大师周仲瑛门诊治疗的2 177条肺癌患者临床数据中的实体划分为临床表现、舌象、脉象、身体部位和程度5类,实现了四诊信息实体抽取,5类实体的*F1*值均>93%。邓嘉乐等[14]使用脱敏的500位肝癌患者的电子病历数据,定义了23类相关实体,IAA达到88%,测试集*F1*值为85.46%。但这些中文电子病历NER语料库都是非公开的。

近年来,一些公开的语料库逐渐建立,丰富了NER和医疗信息提取的数据资源。目前为止,国内使用最普遍的公开数据集是全国知识图谱与语义计算大会(CCKS)数据集,其中CCKS 2017最为常用。CCKS 2017的语料来源于真实电子病历记录,其中包含1 596个带注释的实例,将医疗实体划分为症状、检查、疾病、治疗和身体部位5类。但是该数据集语料来源于多学科,划分的实体类型不适用于专病电子病历,临床应用价值不足。瑞金医院糖尿病数据集是目前最常用的公开数据集,其数据来源于糖尿病领域的中文权威杂志,共包含363篇文档,约250万字,医疗实体被划分为15种。但该数据集的数据来源并非真实电子病历,因此无法反映出临床医学文本的特点。CMeEE-V2数据集[15]是中文医疗信息处理评测(CBLUE)中所使用的数据集。CMeEE-V2来源于医学文本,较CMeEE原始版本,提升了语料质量,共包含训练集数据15 000条,验证集数据集5 000条,测试集数据集3 000条,实体共划分为9类。但其标注的实体中包含大量嵌套实体,增加了NER模型的识别难度。

基于对医院真实电子病历信息提取的需求,本文提出专病电子病历的实体语料库构建方法,并选取2个病种的电子病历作为标注对象,先后构建儿童支气管肺炎和糖尿病的真实电子病历实体语料库。通过与医学专家的合作和多轮迭代的标注流程,确保标注工作的专业性、准确性和一致性。与目前医学领域公开NER数据集相比,本文提出的专病实体语料库构建方法在语料库的标注规范和实体类型的划分上更加精准和深入。

2 中文专病医学实体语料库构建及评价

2.1 *实体语料库构建* 电子病历命名实体标注规范难度较大,既涉及专业的医疗知识,还涉及到对医疗实体的定义和分类。参考杨锦锋等[9]的标注流程,采用多轮迭代的模式进行电子病历命名实体的标注工作(图1)。

2.1.1 *第1阶段* 在医学专家的指导下,总结中文电子病历文本的特点,制定标注规范,并确定特定病种的实体类型。标注方法采取2名标注人员为主,2名医学专家从旁指导的模式。为了便于实体标注,使用阿里云的实体标注工具,标注示例如图2所示。

2.1.2 *第2阶段* 本阶段为预标注阶段,旨在培训标注人员,将标注规范和专业指导相结合便于规范的完善以及后续问题的修正。预标注一共包含90份电子病历文本,分为3轮,每轮30份病历。每轮由2名标注人员根据相同的标注规范,完成相同病

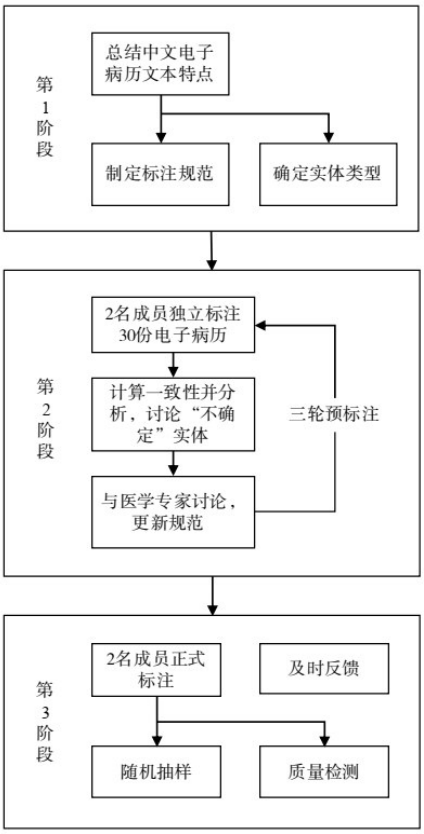

图 1　语料库标记流程

历的标注任务。

具体流程：①第 1 轮抽取 30 份病历，根据标注规范进行标注，对于不确定实体，建立“不确定”实体分类，方便后续讨论并统一标注。②评价标注结果一致性，分析不一致的标注，对“不确定”实体进行讨论。③与医学专家进一步讨论，完善标注规范。④基于新的标注规范，按照流程①至④，开始第 2 轮和第 3 轮标注。经过 3 轮预标注，标注规范趋于稳定，IAA 达到较高水平。

2.1.3　第 3 阶段　正式标注电子病历，其中包含预标注的 90 份病历文本。为保证标注进度及质量，采取如下措施：①及时反馈。标注人员及时反馈“不确定”的实体标签，避免相同问题的重复出现，经过讨论后再确定标注的实体类型。②随机抽样。医学专家根据 2 名标注人员的进度不定时随机抽取 20% 标注结果进行检查，及时纠正标注错误。③质量检测。在正式标注过程中，2 名标注人员在对方的标注结果中随机抽取 50 份病历文本重新标注，用于检测 IAA。命名实体的标注原则遵循实体间不重叠、不嵌套、实体不含表示停顿的标点符号（、，。：；）和连接词（或、和、以及）。

2.2　标注一致性　IAA 常用 Kappa 值[16]和 F 值[17]计算。Kappa 值的计算需要确定的负例数量，常用于分类实验的一致性评价。在 NER 任务中，负例指的是没有标记的文字，通常难以统计。因此，在负例很多的情况下，常用 F 值评价标注一致性。

将 2 名标注人员记为 A 和 B。将 A 的标注结果作为正确答案，计算 B 标注结果的精确度（precision，Pre）和召回率（recall，Rec），进而计算 F 值。计算公式如下：

$$Pre = \frac{\text{A和B标注一致数}}{\text{B标注总数}} \tag{1}$$

$$Rec = \frac{\text{A和B标注一致数}}{\text{A标注总数}} \tag{2}$$

$$F = \frac{(1+\beta^2) \times Rec \times Pre}{(\beta^2 \times Pre) + Rec} \tag{3}$$

其中，β 表示精确率和召回率的相对权重。$\beta>1$，表示召回率权重大；$\beta<1$，表示精确率权重大。本文不做召回率和精确率的权重设置，因此 $\beta=1$。

2.3　NER 模型及评价指标　本文采用的电子病历 NER 模型由嵌入层、双向长短期（bidirectional long short-term memory network，BiLSTM）层和标签的解码层采用条件随机场（conditional random field，CRF）层组成（图 3）。为了更好地理解字词含义、句法和语义信息，模型的嵌入层分别采用 BERT[18]和 ERNIE[19]预训练模型。BiLSTM[20]由前向 LSTM 和后向 LSTM 组成（图 4），用来获得更加全面的上下文信息并学习上下文之间的依赖关系。CRF[21]可以对标

图 2　标注示例

签序列的内部依赖关系进行建模，能够有效解决标签之间的冲突和歧义关系，并且在训练过程中学习标签之间的依赖关系，从而避免一些语法错误。

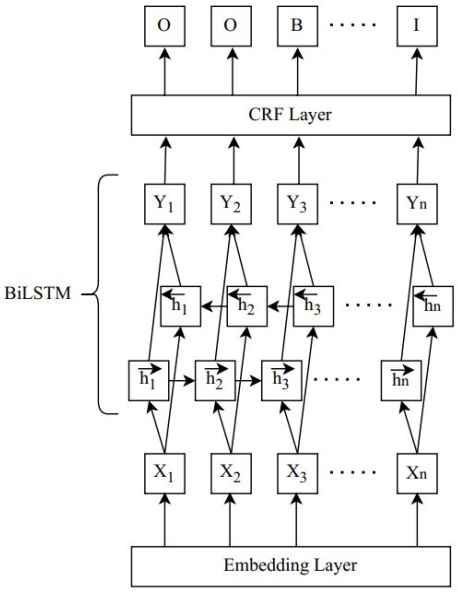

图3 模型结构图

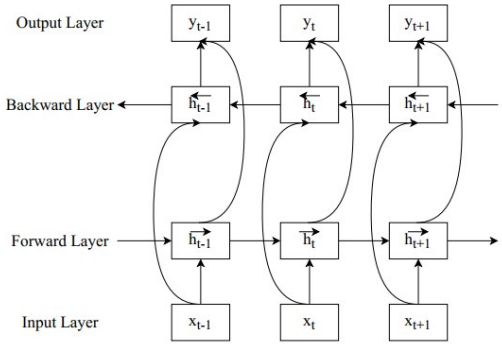

图4 BiLSTM网络结构

采用精确率P、召回率R和$F1$值评估模型的性能。计算方法如下：

$$P = \frac{T_P}{T_P + F_P} \tag{4}$$

$$R = \frac{T_p}{T_P + F_n} \tag{5}$$

$$F1 = \frac{2PR}{P + R} \tag{6}$$

其中，T_P表示真正例，F_P表示假正例，F_n表示假负例。

2.4 实验环境与设置 实验环境采用Ubuntu18.04操作系统，编程语言为Python3.9，GPU采用单卡NVIDIA GeForce 4090，深度学习框架采用Pytorch1.12.0。具体实验参数设置如表1所示。

3 实验结果与分析

3.1 儿童支气管肺炎实体语料库标注 儿童支气管肺炎实体语料库的电子病历文本来源于四川省绵阳市某医院儿科及呼吸内科的2 474条数据，从中随机选取1 000份首次病程记录文本进行标注。经过与医学专家的讨论，本文将实体类型划分为10类，实体标注数量为84 143个（表2）。

表1 实验参数设置

参数名称	参数值
训练集:测试集	7:3
最大序列长度	512
BiLSTM隐含层维度	256
Dropout rate	0.2
学习率	1×10^5
Batch size	128
Epoch	50
优化器	Adam

IAA结果（表3）显示：一致性随迭代标注流程递增。标注过程中的及时讨论，使得标注人员对规范的理解和运用趋于一致，说明通过预标注的方法进行修订标注规范是有效的。当IAA达到0.8时，可以认为语料库标注结果符合一致性要求[22]。基于此，本文构建的语料库标注方法是可靠的。

表2 儿童支气管肺炎实体语料库实体类型和数量

实体类型	示例	数量	实体类型	示例	数量
疾病	支气管肺炎	5 751	辅助检查	肺部超声	13 564
既往史	异物吸入史	3 135	药物名称	沙丁胺醇	2 336
症状	咳嗽	22 749	治疗手段	雾化	3 468
体征	肿大	15 467	身体部位	扁桃体	11 432
体格检查	脉搏	5 273	护理	二级护理	968

表3 儿童支气管肺炎实体语料库标注一致性

标注阶段	Pre	Rec	F值
第1轮预标注	0.75	0.81	0.78
第2轮预标注	0.83	0.89	0.86
第3轮预标注	0.94	0.96	0.95
正式标注	0.97	0.99	0.98

Pre：精确度；Rec：召回率。

3.2 儿童支气管肺炎实体语料库效果对比 为考察本文提出的专病医疗实体语料库标注方法对NER任务的有效性，采用BERT-BiLSTM-CRF和ERNIE-BiLSTM-CRF模型，将构建儿童支气管肺炎实体语料库与CMeEE-V2数据集进行比较。结果（表4）显示：儿童支气管肺炎实体语料库在NER任

务上的表现优于CMeEE-V2数据集。由于CMeEE-V2数据集的各类实体样本数量不均衡，并且包含大量的嵌套实体，增加了实体识别难度，在BERT-BiLSTM-CRF和ERNIE-BiLSTM-CRF上的*F1*值较低，分别为60.58%和62.03%。而自建的儿童支气管肺炎实体语料库中不同实体类型间样本分布均匀，并且在标注时采取实体不嵌套原则，所以在任务中取得更好的效果，*F1*值分别达到80.23%和80.92%。

表4　儿童支气管肺炎实体语料库效果对比结果

数据集	模型	*P*/%	*R*/%	*F1*/%
CMeEE-V2	BERT-BiLSTM-CRF	57.69	63.77	60.58
儿童支气管肺炎实体语料库		73.87	87.78	80.23
CMeEE-V2	ERNIE-BiLSTM-CRF	60.15	64.03	62.03
儿童支气管肺炎实体语料库		74.85	88.06	80.92

P：精确率；R：召回率。

3.3　*糖尿病实体语料库标注*　为了进一步验证专病实体语料库标注方法的鲁棒性，运用同样的标注方法重新标注糖尿病电子病历。电子病历文本来源于四川省绵阳市某医院内分泌科的443份首次病程和入院记录文本。糖尿病实体语料库中的实体类型划分为15类，实体标注数量为53 129个（表5）。IAA结果（表6）显示：糖尿病实体语料库正式标注阶段的IAA为0.95，该数据集的标注方法是可靠的。

表5　糖尿病实体语料库实体类型和数量

实体类型	示例	数量	实体类型	示例	数量
疾病	糖尿病	5 255	持续时间	10余年	1 751
既往史	糖尿病史	3 944	实验室检查	尿常规	4 347
症状	尿急	8 386	用药剂量	8单位	322
体征	反跳痛	10 237	程度	加重	465
体格检查	心率	3 112	用药途径	皮下注射	649
药物名称	利尿剂	1 276	用药频率	三餐前	170
治疗手段	降糖	1 089	检查结果	18次/分	6 236
身体部位	肝脾	5 890			

3.4　*糖尿病实体语料库效果对比*　将标注的糖尿病实体语料库和瑞金医院糖尿病数据集在NER任务上进行比较，结果（表7）显示：瑞金医院糖尿病数据集以中文糖尿病领域权威期刊内容作为语料，其临床医学价值不及真实电子病历，并且在BERT-BiLSTM-CRF和ERNIE-BiLSTM-CRF上的*F1*值较低，分别为66.09%和67.37%。而自建的糖尿病语料库根据真实电子病历的特点重新划分了15类实体类型，可以覆盖病历中绝大多数实体，*F1*值分别为76.68%和78.24%。因此，较瑞金医院糖尿病数据集，自建的实体语料库在糖尿病专病NER任务中的价值和适用性更高。

表6　糖尿病实体语料库标注一致性

标注阶段	Pre	Rec	F值
第1轮预标注	0.60	0.64	0.62
第2轮预标注	0.76	0.79	0.78
第3轮预标注	0.87	0.89	0.88
正式标注	0.93	0.97	0.95

Pre：精确度；Rec：召回率。

表7　糖尿病实体语料库效果对比结果

数据集	模型	*P*/%	*R*/%	*F1*/%
瑞金医院糖尿病数据集	BERT-BiLSTM-CRF	69.21	63.23	66.09
糖尿病实体语料库		69.71	85.19	76.68
瑞金医院糖尿病数据集	ERNIE-BiLSTM-CRF	69.56	65.31	67.37
糖尿病实体语料库		70.23	88.31	78.24

P：精确率；R：召回率。

4　结　论

针对医疗领域电子病历NER任务语料库匮乏的现状，本文构建了2个一致性较高的专病实体语料库——儿童支气管肺炎语料库和糖尿病语料库。在标注阶段，结合中文电子病历和专病病历的特点，划分了不同的实体类型，并制定了专病实体语料库标注规范。

将构建的儿童支气管肺炎实体语料库与CMeEE-V2数据集分别在BERT-BiLSTM-CRF和ERNIE-BiLSTM-CRF模型上进行比较，构建的实体语料库在NER任务中表现良好，说明NER模型可以根据电子病历进行准确的预测，证明了本文构建实体语料库方法的有效性。

按照相同的专病实体语料库标注方法重新构建糖尿病实体语料库，与公开数据集的比较分析结果提示，自建实体语料库的标注质量优于所对比的公开数据集，说明本文提出的专病电子病历实体语料库构建方法能够适应不同的特定病种，在增加实

体类型的情况下，仍然能够保持较高*F1*值，进一步证实本文构建的专病实体语料库标注方法具有鲁棒性。

未来将针对自建语料库改进NER模型，如引入自注意力机制，捕捉句子中长距离的依赖关系，准确识别实体边界和类型，提高NER任务的准确率。

伦理声明 无。

利益冲突 所有作者声明不存在利益冲突。

作者贡献 陈思旭：论文选题、撰稿；刘独玉、谭小琴、齐星、罗彬：论文修改。

参考文献

[1] 中华人民共和国国家卫生和计划生育委员会. 电子病历应用管理规范(试行)[J]. 中国实用乡村医生杂志, 2017, 24(6): 1-2.

[2] RAU L F. Extracting company names from text[C]//[1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application. Piscataway: IEEE, 1991: 29-32.

[3] MEYSTRE S, HAUG P J. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation[J]. J Biomed Inform, 2006, 39(6): 589-599.

[4] SAVOVA G K, MASANZ J J, OGREN P V, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications[J]. J Am Med Inform Assoc, 2010, 17(5): 507-513.

[5] ROBERTS A, GAIZAUSKAS R, HEPPLE M, et al. Building a semantically annotated corpus of clinical texts[J]. J Biomed Inform, 2009, 42(5): 950-966.

[6] CAMPILLOS L, DELÉGER L, 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]. Lang Resour Eval, 2018, 52(2): 571-601.

[7] BYUN S, HONG J, PARK S, et al. Korean bio-medical corpus (KBMC) for medical named entity recognition[EB/OL]. [2024-06-03]. *http://arxiv.org/abs/2403.16158*.

[8] LEI J B, TANG B Z, LU X Q, et al. A comprehensive study of named entity recognition in Chinese clinical text[J]. J Am Med Inform Assoc, 2014, 21(5): 808-814.

[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[J]. 软件学报, 2016, 27(11): 2725-2746.

[10] 昝红英, 刘 涛, 牛常勇, 等. 面向儿科疾病的命名实体及实体关系标注语料库构建及应用[J]. 中文信息学报, 2020, 34(05): 19-26.

[11] 常洪阳, 昝红英, 马玉团, 等. 脑卒中疾病电子病历实体及实体关系标注语料库构建[J]. 中文信息学报, 2022, 36(8): 37-45.

[12] 巩敦卫, 张永凯, 郭一楠, 等. 融合多特征嵌入与注意力机制的中文电子病历命名实体识别[J]. 工程科学学报, 2021, 43(9): 1190-1196.

[13] 屈丹丹, 杨 涛, 朱 垚, 等. 基于字向量的BiGRU-CRF肺癌医案四诊信息实体抽取研究[J]. 世界科学技术-中医药现代化, 2021, 23(9): 3118-3125.

[14] 邓嘉乐, 胡振生, 连万民, 等. 基于RoBERTa-CRF的肝癌电子病历实体识别研究[J]. 医学信息学杂志, 2023, 44(6): 42-47.

[15] ZAN H Y, LI W X, ZHANG K L, et al. Building a pediatric medical corpus: word segmentation and named entity annotation[M]//Lecture notes in computer science. Cham: Springer International Publishing, 2021: 652-664.

[16] JEAN C. Assessing agreement on classification tasks: the kappa statistic[J]. Comput Linguist, 1996, 22(2): 249-254.

[17] HRIPCSAK G, ROTHSCHILD A S. Agreement, the f-measure, and reliability in information retrieval[J]. J Am Med Inform Assoc, 2005, 12(3): 296-298.

[18] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2024-06-03]. *http://arxiv.org/abs/1810.04805*.

[19] ZHANG Z Y, HAN X, LIU Z Y, et al. ERNIE: enhanced language representation with informative entities[EB/OL]. [2024-06-03]. *http://arxiv.org/abs/1905.07129*.

[20] WANG Z H, YANG B. Attention-based bidirectional long short-term memory networks for relation classification using knowledge distillation from BERT[C]//2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). Piscataway: IEEE, 2020: 562-568.

[21] CHEN W L, ZHANG Y J, ISAHARA H. Chinese named entity recognition with conditional random fields[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: Association for Computational Linguistics, 2006: 118-121.

[22] ARTSTEIN R, POESIO M. Inter-coder agreement for computational linguistics[J]. Comput Linguist, 2008, 34(4): 555-596.

引用本文

陈思旭，刘独玉，谭小琴，等. 面向专病电子病历的实体语料库构建方法[J]. 元宇宙医学，2024，1(3)：41-46.

CHEN S X, LIU D Y, TAN X Q, et al. Construction methodology of entity corpus for special diseases electronic medical records[J]. Metaverse Med, 2024, 1(3): 41-46.