

## 中文电子病历标注系统构建与应用

赵琬清<sup>1</sup> 胡佳慧<sup>1</sup> 陈凌云<sup>1</sup> 娄培<sup>1</sup> 方安<sup>1</sup>

(<sup>1</sup> 中国医学科学院医学信息研究所, 北京 10020)

**[摘要]** **目的/意义** 构建中文电子病历标注系统, 实现电子病历文本的自动化标注。**方法/过程** 从数据特性与标注需求、功能需求方面详细分析了系统需求; 详细阐述了系统的完整架构, 从数据层、服务层与功能层三个方面对中文电子病历标注系统进行详细介绍, 包括用户权限管理、实体和关系标注流程以及相关关键技术。**结果/结论** 构建的中文电子病历标注系统能有效满足电子病历标注任务与需求, 目前已成功应用于垂体瘤电子病历语料的构建工作。

**[关键词]** 中文电子病历; 文本标注; 医学标注系统; 实体识别

### Construction and application of Chinese electronic medical record annotation system

ZHAO Wanqing HU Jiahui CHEN Lingyun LOU Pei FANG An

The Institute of Medical Information (IMI) & Library, Chinese Academy of Medical Sciences and Peking Union Medical College, BEIJING 10020, CHINA

**[Abstract]** **Purpose/Significance** We construct a Chinese electronic medical record(EMR) annotation system, aiming to annotate EMR automatically.

**Method/Process** We analyzed our system requirement from data features, annotate and functional requirements. We further introduce system framework, including data, service and function layers, with user authority managements, annotative pipelines with entity and relations labeling, and some key technologies.

**Result/Conclusion** The annotation system we constructed can fully meet EMR labeling tasks and requirements. It has been successfully applied to the construction of pituitary adenoma EMR corpus.

**[Keywords]** Chinese electronic medical record; text labeling; Medical labeling system; Entity Recognition

#### 1 引言

随着医疗机构中电子病历的广泛应用, 海量的医疗诊断信息被存储于自由文档格式的文本中亟待进一步挖掘与利用[1], 电子病历数据不仅包含大量的在线或实时数据, 同时包括临床决策支持中的诊断和用药建议、各种结构化数据表、非结构化、半结构化文本文档等多种数据[2], 其中基于电子病历文本的命名实体识别过程是对电子病历信息抽取与挖掘过程的关键技术与难点[3]。电子病历中的文本蕴含了大量的医疗知识, 包含着丰富的临床知识, 然而这些知识一般以自然语言文本形式记录的, 需要用到自然语言处理领域的信息抽取技术把其中有价值的信息提炼为结构化的形式, 方便后续的数据挖掘工作[4]。电子病历文本结构化研究属于自然语言处理技术在医疗领域中的应用, 结合通用领域的自然语言处理技术并针对医疗领域文本特点做出相应的改进与优化, 具有很强的实际应用价值, 辅助临床科学研究、临床决策支持、护理健康评估等[5]。目前国内国外已有大量电子病历文本结构化相关研究与评测竞赛[6], 英文以美国国家临床自然语言处理挑战 (Informatics for Integrating Biology & the Bedside/ National NLP Clinical Challenges, i2b2/n2c2) [7]为主要代表, 中文以中文信息学会主办的中国知识图谱与语义计算大会 (China Conference on Knowledge Graph and Semantic Computing, CCKS) [8]和中国健康信息处理大会 (China Conference on Health Information Processing, CHIP) [9]为主要代表。电子病历的文本较为模式化, 如“发热”、“头痛”等常用症状描述词以及“精神可”、“睡眠好”等模式化的例行检查文本的多次出现, 传统人工标注电子病历数据的过程非常低效, 同时也无法完全保证此类文本的标注准确性。不同科室病历文本在医学描述上存在一定的差异, 无法用同样的标准对不同病历文本进行数据标注[10]。

现有的文本标注工具主要包括 brat[11]、doccano[12]、Label Studio[13]、YEDDA[14]、prodigy[15]、poplar[16]等, 其中 prodigy 为商业标注工具, Label

**[修回日期]** 2025-02-27

**[作者简介]** 赵琬清, 助理研究员, 发表论文 3 篇。通讯作者: 方安, 研究馆员

**[基金项目]** 中国医学科学院医学与健康科技创新工程项目 (项目编号: 2021-I2M-1-056) 中央高水平医院临床科研业务费(项目编号: 2022-PUMCH-A-084)

Studio 分为开源版本与商业版本，目前开源版本中的自动标注功能需要用户创建第三方标注 API 辅助使用。文本标注工具简介见表 1。

表 1 文本标注工具介绍

文本标注工具	简介
Brat[11]	基于 Web 部署的文本结构化工具，包括实体与关系标注功能
Doccano[12]	提供文本分类、序列标记和序列到序列任务标注功能
Label Studio[13]	分为开源版与商业版，标记音频、文本、图像、视频和时间序列
YEDDA[14]	基于 Python 开发，提供单机版本部署，标记实体与事件
Prodigy[15]	商业版，基于 spacy 库用户创建机器学习训练、评估数据
Poplar[16]	森亿智能开源的标注工具，参考 brat 工具开发

在实际试用文本标注工具的过程中，从易用性角度，商业版本软件较开源软件更具优势。例如，Prodigy 工具借助 Spacy[17]工具可快速实现英文文档的实体识别，其设计方案以用户为中心，界面友好。然而，该工具闭源，且 Spacy 不支持中文，这限制了其在国内的应用。此外，付费要求也使许多入门用户望而却步。尽管 GitHub 上有一个专门讨论中文文本标注工具技术架构的代码库[18]，但目前尚未开展实际工具开发。在开源软件中，Label Studio 工具表现较为出色，部署和使用过程较为流畅。但其开源版本功能有限，自动化机器学习标注模型要求用户熟悉相关算法接口部署，并以接口形式嵌入系统提供辅助标注服务。Brat 工具作为使用最广的开源标注工具，不提供自动辅助标注功能，用户需自行导入数据进行标注[19]。Doccano 工具以 Docker[20]形式部署，对用户部署能力要求较高，不熟悉 Docker 环境的用户可能无法自行完成。Popular 工具参考 Brat 开发，采用 JSON 格式存储数据，可移植性强，但同样不提供自动标注功能。此外，开源工具的核心优势在于其源代码的开放性，用户能够依据特定需求进行自主的定制化开发。然而，这类工具通常缺乏稳定的支持与维护体系，在部署及使用过程中，用户往往需要依赖社区资源或第三方支持来解决遇到的问题。同时，开源工具的学习曲线相对陡峭，部分工具甚至要求在 Linux 系统环境下进行部署，这对于不熟悉代码操作和环境配置的用户群体而言，存在较高的使用门槛，不够友好。

综上所述，开源标注工具要求用户具备一定部署和代码能力，以便引入第三方辅助标注工具完成自动标注操作，而商业版本因高额成本，难以普及医学数据标注服务。因此，开发一款无需用户部署、免费可用的医学文本标注工具，可为广大用户提供医学数据标注功能，降低使用门槛，有助于医学科研人员自主生成医学数据集。

本研究旨在构建一个能适应不同类型病历文本标注需求的中文电子病历标注系统，辅助临床大夫进行病历数据的分析与信息抽取，将自动标注算法融合到标注过程背后，降低人工重复劳动的同时提高标注的准确性。同时，保障用户在不具备系统部署能力与代码能力的基础上，直接使用 Web 页面的中文电子病历标注系统完成自动化辅助标注电子病历操作。目前中文电子病历标注系统已上线<sup>1</sup>，系统提供相关功能介绍视频<sup>2</sup>与试用功能。

## 2 系统需求分析

### 2.1 数据特性与标注需求

电子病历中一般以 XML 的形式存储在医院 HIS 系统中，导出的 XML 文件通常带有大量标签，会对标注过程造成一定的干扰。由于标注主体为文本格式，一般可以存储至 TXT 格式文件中，因而系统需要支持 TXT 格式的原始数据上传。同时，在标注过程中需要对不同用户的标注数据进行区分，采用 JSON 格式进行存储。

### 2.2 功能需求

#### 2.2.1 自定义标注配置

不同医疗机构以及相同医疗机构的不同科室的电子病历的结构和内容都存在一定的差异[2]，标注系统需要针对不同的电子病历文本构建不同的标注配置。

#### 2.2.2 用户权限管理

电子病历标注过程需要多个标注用户的参与，同时由于不同标注人员的标注结果无法完全匹配，需要由更高权限的管理员用户对标注的病历数据进行审核，最终确定统一的标注结果文件。

#### 2.2.3 辅助标注

标注数据是一项重要的基础工作，标注数据力求精准，为后续的数据处理工作奠定了良好的基础[3]。标注数据首先要保证准确性，其次要保证效率，现有的人工标注数据方式主要由专业领域的专家进行标注，一般能满足准确性，但是时效性较差。而面向中文电子病历的自动标注效果不够理想，还需要人工进行协助标注。将人工与机器学习的方法结合，

1 <https://ccnlp.imicams.ac.cn/>

2 <https://ccnlp.imicams.ac.cn/introduction>

以实现自动化辅助电子病历标注，能够提升标注的质量和效率。系统应注重标注效率与准确性的平衡，保证标注准确性的同时能即时给用户标注反馈，有助于语料标注的迭代。

### 2.2.4 标注状态流转

一份数据文件的标注通常不是一蹴而就的，整体标注过程通常经历多次修改与审核，因此标注文件的状态需区分为标注未完成、标注完成待审核、审核中未锁定、已锁定四个阶段，需要对这四种不同阶段的标注文件状态进行区分。

### 2.2.5 一致性评价

在数据标注过程中，一致性评价是判断数据标注质量的重要指标，在一轮标注完成后，系统应给出多名标注人员的一致性评分，便于审核人员决定后续的优化方向，如细化标注规范、加强标注人员培训等。管理员在审核标注数据的过程中能够看到一份数据多名用户的标注一致性，标注存疑的实体可以通过标注一致性辅助判断。

## 3 系统设计与实现

### 3.1 系统架构

中文电子病历标注系统主要分为数据层、服务层和功能层，为中文电子病历标注系统的系统架构图见图 1 所示。数据层为了支撑文本病历标注的分析，引入语料数据，模型及字典数据，语料数据保存有原始数据、标注数据以及审核数据，记录电子病历数据在标注过程中的三个阶段。服务层基于熵的主动学习技术及条件随机场训练模型，提供主动学习服务、训练模型服务、辅助标注服务、语料推荐服务。功能层基于结构化框架需求，主要面向用户提供实体标注的功能。包括：实体配置，语料推荐，实体标注，辅助标注，实体统计，一致性评价以及用户管理等功能。

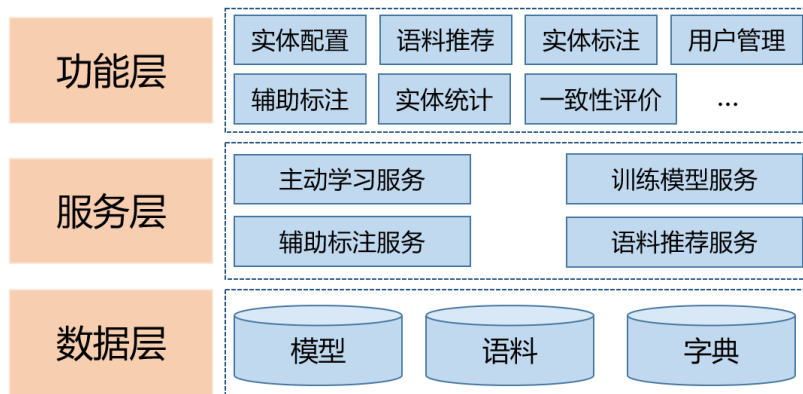


图 1 中文电子病历标注系统架构图

### 3.2 用户权限管理模块

用户管理模块是中文电子病历标注系统的基础模块，包括多级角色设计以及用户的注册、激活、锁定、解锁以及用户组的分配。标注系统用户可分为系统管理员和普通用户。其中普通用户又可拥有用户组管理员与普通成员的身份。用户组由系统管理员创建，普通用户可以拥有多个用户组的身份，从而参与不同的项目，项目彼此之间数据隔离。系统管理员、用户组与用户的架构见图 2 所示。用户可以是某个用户组的成员，也可以单独存在，如用户 e 与用户 f。用户组 A 的管理员用户 a 可以为用户组 B、C、D 的成员，但同一用户组只能有唯一的用户组管理员。

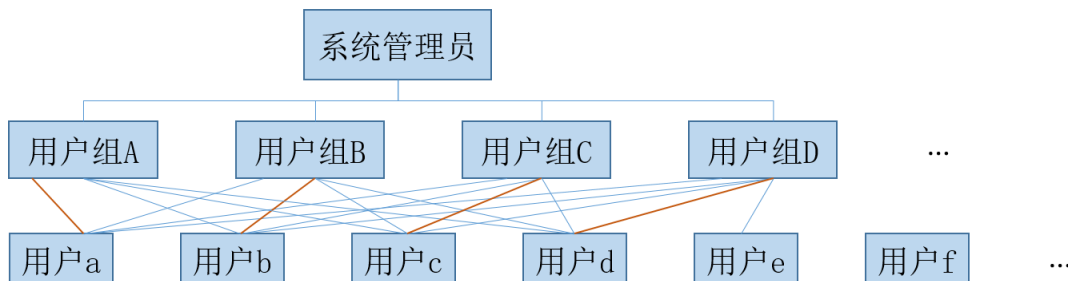


图 2 系统多级用户角色架构图

### 3.3 实体和关系标注流程

实体和关系标注流程主要包括标注数据从原始数据到标注完成的语料数据的全流程操作，包括项目创建、数据上传、实体和关系配置、语料标注与审核。本研究的标注流程设计参照了标注流程中所需的流水线架构，与文献[4][19]中提及的标注流程具有相似性。目

前，整体的标注流程已成功融入中文电子病历标注系统之中。用户在使用该系统时，只需遵循系统所推荐的步骤进行数据标注操作，即可实现数据集的高效构建。

### 3.3.1 项目创建

标注数据以项目的形式呈现，用户首次登录后进入项目管理界面，自动进行新手引导。用户创建项目包含项目名称、项目类型、项目描述、标注类型和推荐语料数量。项目类型包括团队模式和个人模式，团队模式为多人标注项目，而个人模式为单人标注项目。推荐语料数量为给项目标注者推荐最有标注价值的指定数量的语料。用户项目角色权限见表 2 所示。

表 2 系统用户项目角色权限列表

菜单功能	项目创建者（团队模式）	项目创建者（个人模式）	审核员	标注员
分配用户	√			
实体配置	√	√		
数据集维护	√	√		
语料上传、 下载	√	√		
数据标注与 审核	√	√	√	√
语料标注、 语料审核、 锁定	√	√	√	
一致性评价	√		√	
数据统计	√	√	√	√
数据导出	√	√	√	√
辅助标注配置	√	√	√	
知识图谱（实体和关系标注项目）	√	√		

### 3.3.2 数据上传

数据上传以后缀名为 txt 的语料进行上传，项目创建者及审核员可上传语料。在语料上传完成后展示上传语料总数量、上传失败文件数，支持超出屏幕范围的文件进行折叠隐藏。

### 3.3.3 实体和关系配置

在新增项目后，系统将生成默认的实体配置文件，后续可在实体关系配置功能中进行修改。用户可在实体和关系页面中对项目的实体关系配置进行更改，设置不同的实体名称、编码、实体颜色、以及对应的快捷键位，点击保存后即在服务器后台生成相应的 JSON 配置文件，也可以点击“JSON”按钮后 JSON 编码形式进行批量更改，默认实体配置与 JSON 配置操作见图 3 所示。以“疾病”实体为例，JSON 配置中的“text”字段对应实体的显示名称“疾病”；“code”字段取值为“disease”；表明该实体在系统内部以“disease”作为唯一编码存储；“prefixKey”字段定义了实体标注快捷键中的组合键部分，空值表示未启用组合键；“keyName”字段则指定了实体标注的快捷键字符为 Q，因此“疾病”的实际标注快捷键即为 Q；“color”字段设定了实体标记的颜色，此处为浅红色，对应 CSS 中的 RGB 编码“#dd0f20”。同理，“症状”实体的“prefixKey”配置为“Alt”，“keyName”仍为 Q，那么“症状”的实际标注快捷键组合为 Alt+Q。

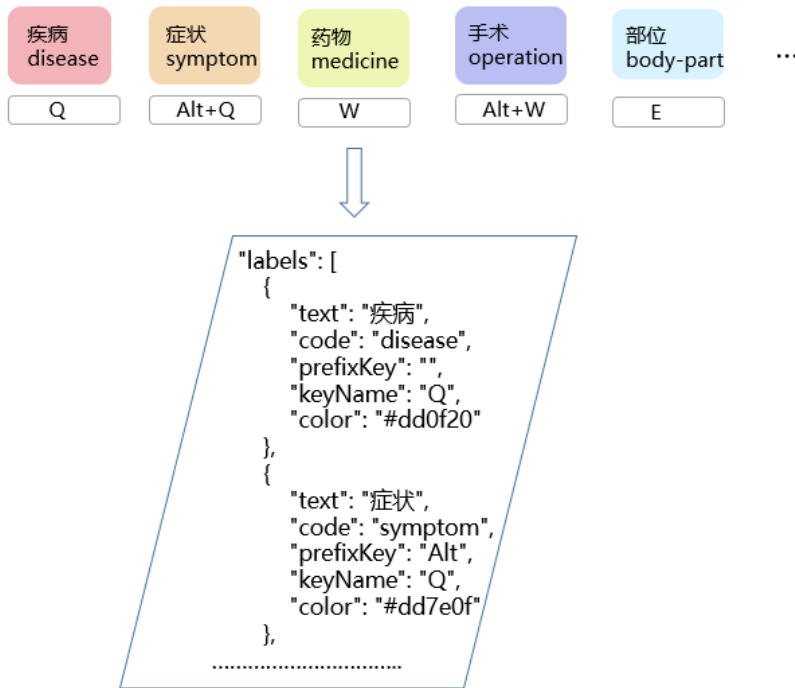


图3 系统实体和和关系配置示意图

### 3.3.4 语料标注与审核

在标注数据的过程中，由项目创建者或审核员对数据进行上传、实体配置，再由标注员对病历文本进行标注。标注审核过程中的数据状态流程图见图4所示。

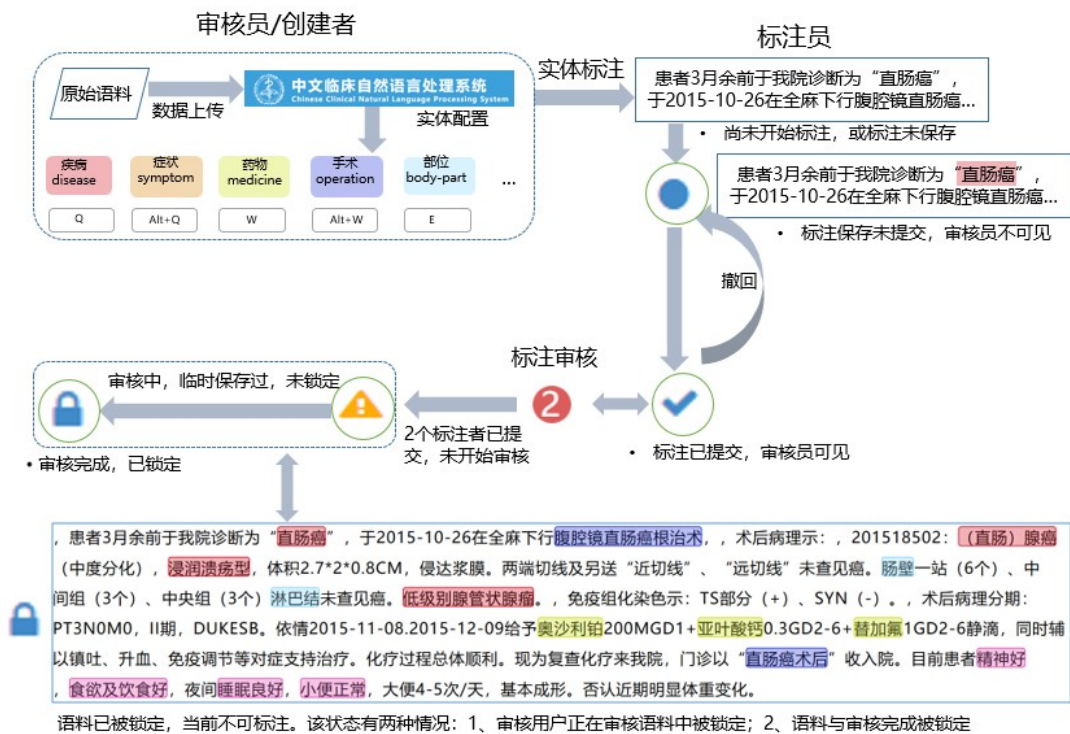


图4 系统标注流程图

## 3.4 关键技术

### 3.4.1 主动学习

基于电子病历实体识别的标注数据有限，但是对标注精度要求较高，因而利用主动学习方法，有目的地选择标注数据。从未标注数据中选择未充分训练的数据（我们定义为未

标注语料中对标注价值最大的语料），交给用户进行标注，并将标注好的数据加入到训练集中，进行下一次模型训练[21]。

### 3.4.2 辅助标注

实体和关系标注还提供自动化的辅助标注功能，旨在减少重复性的人工劳动，避免冗余标注，最大化标注准确性与效率。我们先后采用了两种不同的辅助标注方法，见图5所示。第一种方法为基于字典匹配的辅助标注算法[22]，首先人工标注一部分的病历文本，审核完成后作为标准标注数据文件，从这些标准标注数据文件中抽取已经标注的实体作为字典，由于有些词可以在不同文本中标注为不同的实体，因此采用投票的方式选取字典中的实体类型。第二种方法为基于条件随机场(Conditional Random Fields, 以下简称CRF)的辅助标注算法[23]，基于实体识别性能较好的神经网络模型条件随机场(CRF)，特征构造过程中采用的特征为上下文特征、字典特征、部首特征等。利用条件随机场来训练实体识别模型，并采用开源的CRF++工具[24]作为我们依赖的工具。使用原始字、分词的结果、以及上下文(窗口大小为5)中的信息作为特征，对CRF模型进行训练。前期的人工标注工作与第一种方法类似，不同的是将已经标注完成的标注数据文件作为CRF模型的训练数据，后续由训练完成的模型自动输出未标注电子病历文件的标注结果，用户在辅助标注的基础上修改标注结果，从而减轻负担，提高标注效率。最后实践表明，CRF模型辅助人工标注的效果更好，因而采用CRF模型进行辅助标注。

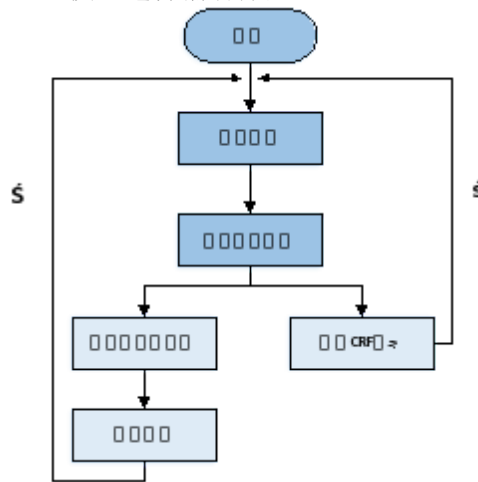


图5 系统辅助标注流程

### 3.4.3 一致性评价

一致性评价模块仅支持团队模式的实体和关系标注项目。在多名用户标注同一语料的情况下，项目创建者与审核员可以按项目或批量选择多个文件进行一致性评价。系统通过标注数、一致数、P值、R值、F值指标做出一致性评价[4]，以表格方式展现。

标注者a与标注者b的P值、R值、F值计算方式如公式(1)~公式(3)所示，最终以F值作为多用户之间的一致性指标：

$$P = \frac{\text{a和标注一致数}}{\text{b的标注数}} \tag{1}$$

$$R = \frac{\text{a和标注一致数}}{\text{a的标注数}} \tag{2}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{3}$$

## 4 系统应用与实践

已在中文电子病历标注系统中完成500份现病史、既往史以及病例特点的标注工作[25]，两名具有神经外科手术经验的临床医生，经过培训作为标注人员对数据集进行了注释。实体注释的整体注释者之间的协议(IAA)为0.86。我们在研究过程中采用基于字典、CRF[23]以及BiLSTM-CRF[26]、BERT-BiLSTM-CRF[27]的方法对实体识别算法进行了评估，算法具体实现参考[24][28][29]。实验表明CRF模型结合词性、部首和文档类型特征的算法

能达到距最优算法 BERT-BiLSTM-CRF 只相差不到 1% 的 F1 值。考虑到 CRF 算法的速度快、系统资源占比小，在实际的中文电子病历标注系统中，如果采用 BERT-BiLSTM-CRF 算法会占用大量的系统计算资源且模型训练时间长，不利于多轮标注迭代反馈，平台最终仍采用 CRF 算法实现辅助标注模块。

## 5 结语

本文构建了一个可动态配置的中文电子病历标注系统，自动进行病历数据的分析与信息抽取，有效降低人工重复劳动的同时提高标注的准确性。该系统可应用于医疗数据集的构建，用于辅助 DRGs 分组、临床路径优化等。

系统也有许多不足有待进一步改进，目前系统仅提供实体和关系类别的数据集构建功能，实际上在完整的医学自然语言处理研究中，除了信息抽取任务，还包括文本分类与文本相似度计算、知识图谱与问答、文本生成与知识推理、以及现在受到广泛关注的大语言模型评测[30]，如 GPT3[31]、Deepseek[32]等模型在医疗领域的进一步应用探索。未来我们将探索更多医学自然语言处理任务，并结合大语言模型对现有功能进一步的优化，通过大模型进行预先自动标注，提升数据标注的效率与准确性。

作者贡献：赵琬清负责论文构思与写作与需求分析；胡佳慧负责项目管理；陈凌云负责可视化设计与呈现；娄培负责数据分析；方安负责资源与项目指导。

利益声明：所有作者均声明不存在利益冲突。

## 参考文献

- [1] Heart T, Ben-Assuli O, Shabtai I. A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy[J]. *Health Policy and Technology*, 2017, 6(1): 20-25.
- [2] Sun W, Cai Z, Li Y, et al. Data processing and text mining technologies on electronic medical records: a review[J]. *Journal of healthcare engineering*, 2018, 2018(1): 4302425.
- [3] 杨锦锋,于秋滨,关毅,等.电子病历命名实体识别和实体关系抽取研究综述[J].*自动化学报*,2014,40(08):1537-1562.
- [4] 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 (11): 2725-2746.
- [5] Uslu A, Stausberg J. Value of the electronic medical record for hospital care: update from the literature[J]. *Journal of medical Internet research*, 2021, 23(12): e26323.
- [6] 赵琬清,胡佳慧,娄培,等.基于开放评测的临床信息抽取分析[J].*医学信息学杂志*,2020,41(10):30-36.
- [7] Mahajan D, Liang J J, Tsou C H, et al. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes[J]. *Journal of biomedical informatics*, 2023, 144: 104432.
- [8] Han X, Wang Z, Zhang J, et al. Overview of the CCKS 2019 knowledge graph evaluation track: entity, relation, event and QA[J]. *arXiv preprint arXiv:2003.03875*, 2020.
- [9] 熊英,陈漠沙,陈清财,等.CHIP 2021 评测任务 1 概述: 医学对话临床发现阴阳性判别任务[J].*医学信息学杂志*,2023,44(03):46-51.
- [10] Lloyd S, Long K, Alvandi A O, et al. A National Survey of EMR Usability: comparisons between medical and nursing professions in the hospital and primary care sectors in Australia and Finland[J]. *International Journal of Medical Informatics*, 2021, 154: 104535.
- [11] Stenetorp P, Pyysalo S, Topić G, et al. BRAT: a web-based tool for NLP-assisted text annotation[C]//*Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012: 102-107.
- [12] Nakayama H, Kubo T, Kamura J, et al. doccano: Text annotation tool for human[EB/OL].(2025-03-19) <https://github.com/doccano/doccano>, 2018, 34.
- [13] Tkachenko M, Malyuk M, Holmanyuk A, et al. Label studio: Data labeling software[EB/OL].(2025-03-19) <https://github.com/heartexlabs/label-studio>, 2020, 2022.
- [14] Yang J, Zhang Y, Li L, et al. YEDDA: A lightweight collaborative text span annotation tool[J]. *arXiv preprint arXiv:1711.03759*, 2017.
- [15] Xue L C, Rodrigues J P, Kastritis P L, et al. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes[J]. *Bioinformatics*, 2016, 32(23): 3676-3678.
- [16] Synyi. Poplar: A web-based annotation tool for natural language processing (NLP) [EB/OL].(2025-03-19) <https://github.com/synyi/poplar>, 2020.

- [17]spaCy: Industrial-strength Natural Language Processing (NLP) in Python[EB/OL]. (2025-03-19) <https://github.com/explosion/spaCy>
- [18]Chinese-Annotator: Annotator for Chinese Text Corpus (UNDER DEVELOPMENT) 中文文本标注工具[EB/OL].(2025-03-19) <https://github.com/deepwel/Chinese-Annotator>
- [19]Zhu E, Sheng Q, Yang H, et al. A unified framework of medical information annotation and extraction for chinese clinical text[J]. Artificial intelligence in medicine, 2023, 142: 102573.
- [20]Merkel D. Docker: lightweight linux containers for consistent development and deployment[J]. Linux j, 2014, 239(2): 2.
- [21] 胡佳慧,赵琬清,方安,等.基于主动学习的中文电子病历命名实体识别研究[J].中国数字医学,2020,15(11):6-9.
- [22] Knuth D E, Morris, Jr J H, Pratt V R. Fast pattern matching in strings[J]. SIAM journal on computing, 1977, 6(2): 323-350.
- [23] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Icml. 2001, 1(2): 3.
- [24] Kudo T. CRF++: Yet another CRF toolkit[J]. <http://crfpp.sourceforge.net/>, 2005.
- [25] Fang A, Hu J, Zhao W, et al. Extracting clinical named entity for pituitary adenomas from Chinese electronic medical records[J]. BMC medical informatics and decision making, 2022, 22(1): 72.
- [26] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
- [27] Devlin 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: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.
- [28]zh-NER-TF: A very simple BiLSTM-CRF model for Chinese Named Entity Recognition 中文命名实体识别 (TensorFlow)[EB/OL].(2025-03-19) <https://github.com/Determined22/zh-NER-TF>
- [29] BERT-BiLSTM-CRF-NER: Tensorflow solution of NER task Using BiLSTM-CRF model with Google BERT Fine-tuning And private Server services[EB/OL].(2025-03-19) <https://github.com/macany/BERT-BiLSTM-CRF-NER>, 2020.
- [30]Zong H, Wu R, Cha J, et al. Advancing Chinese biomedical text mining with community challenges[J]. Journal of Biomedical Informatics, 2024: 104716.
- [31] Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences[J]. Minds and Machines, 2020, 30: 681-694.
- [32] Guo D, Yang D, Zhang H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning[J]. arXiv preprint arXiv:2501.12948, 2025.