• 医学信息技术 •

中文电子病历标注系统构建与应用*

赵琬清 胡佳慧 陈凌云 娄 培 方 安

（中国医学科学院/北京协和医学院医学信息研究所 北京 100020）

〔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** 分析系统需求，阐述系统架构，从数据层、服务层与功能层 3 方面对中文电子病历标注系统进行介绍，包括用户权限管理、实体和关系标注流程以及标注算法等。**结果/结论** 中文电子病历标注系统能有效满足电子病历标注任务的需求，目前已成功应用于垂体瘤电子病历语料构建工作。

〔关键词〕 中文电子病历；文本标注；医学标注系统；实体识别

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2025.06.012

Construction and Application of a Chinese Electronic Medical Record Annotation System

ZHAO Wanqing, HU Jiahui, CHEN Lingyun, LOU Pei, FANG An

Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

〔**Abstract**〕 **Purpose/Significance** To construct a Chinese electronic medical record (EMR) annotation system, so as to realize the automatic annotation of EMR texts. **Method/Process** The system requirements are analyzed, and the system architecture is elaborated. The system is introduced from 3 aspects: the data layer, the service layer and the functional layer, including user authority management, entity and relation annotation processes, and annotation algorithms, etc. **Result/Conclusion** The Chinese EMR annotation system can effectively meet the requirements of the EMR annotation task and has been successfully applied to the construction of pituitary adenoma EMR corpus.

〔**Keywords**〕 Chinese electronic medical record (EMR); text labeling; medical labeling system; entity recognition

1 引言

随着电子病历在医疗机构中的广泛应用，海量医疗诊断信息以自由文档格式被存储于文本中，亟待进一步挖掘与利用[1]。电子病历中包含大量在线或实时数据，以及支持临床决策的诊断和用药建议、各种结构化数据表、非结构化或半结构化文档等数据[2]。其中蕴含大量以自然语言文本形式记录的医疗知识，需要应用自然语言处理领域信息抽取技术将其中有价值的信息提炼为结构化形式，以便通过数据挖掘辅助临床科学研究、支持临床决策和护理健康评估等[3－4]。基于电子病历文本的命名实体识别是电子病历信息抽取和挖掘的关键技术与难点[5]。

本文旨在构建适配不同类型病历文本标注需求

〔修回日期〕 2025－03－19

〔作者简介〕 赵琬清，助理研究员，发表论文 3 篇；通信作者：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）。

的中文电子病历标注系统，辅助临床医生进行病历数据分析与信息抽取，将自动标注算法融入标注过程，减少人工重复劳动的同时提高标注准确性。用户在不具备系统部署与代码编写能力的情况下，可直接使用 Web 页面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（https://ccnlp.imicams.ac.cn/），并提供相关功能介绍视频（https://ccnlp.imicams.ac.cn/introduction）与试用功能。

2 国内外相关研究

目前国内外已有大量电子病历文本结构化研究与评测竞赛[6]，英文领域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，中文领域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。从测评中使用的电子病历数据来看，文本格式呈现模式化特点，如多次出现“发热”“头痛”等常用症状描述词，以及“精神可”“睡眠好”等例行检查文本。传统人工标注电子病历数据的过程非常低效，且无法完全保证此类文本标注的准确性。不同科室病历文本的医学描述存在一定差异，无法用统一标准对不同病历文本进行数据标注[10]。目前已有大量文本标注工具，旨在满足多种数据标注需求，包括 brat[11]、doccano[12]、Label Studio[13]、YEDDA[14]、Prodigy[15]、poplar[16]等，见表 1。其中 Prodigy 为商业标注工具；Label Studio 分为开源版与商业版，目前开源版中的自动标注功能需用户创建第三方标注应用程序接口（application programming interface, API）辅助使用。实际试用发现，商业版软件的易用性优于开源软件。例如，Prodigy 工具借助 spaCy 库[17]可快速识别英文文档中的实体，且界面友好；但其闭源、要求付费且 spaCy 库不支持中文。尽管 GitHub 平台上有基于 Prodigy 开发模式进行中文文本标注工具技术架构的探讨[18]，但目前尚未开展实际工具开发。在开源软件中，Label Studio 工具表现比较出色，部署和使用过程均较流畅；但其开源版功能有限，要求用户熟悉相关算法接口部署，并以接口形式嵌入工具平台系统，才能提供辅助标注服务。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，需用户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。Poplar 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，即用户可依据特定需求进行自主定制化开发。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，往往要依赖社区资源或第三方支持来解决问题。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 系统环境下部署，存在较高的使用门槛。总体而言，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，而商业版成本较高，难以普及。因此，开发一款无须用户部署、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服务，降低使用门槛，有助于医学科研人员自主生成医学数据集。

表 1 文本标注工具介绍

名称	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 库创建机器学习训练、评估数据
poplar[16]	开源标注工具，参考 brat 工具开发

3 系统需求分析

3.1 数据格式与标注需求

电子病历一般以 XML 形式存储于医院信息系统，可导出 XML 文件或 TXT 文件。XML 文件通常带有大量标签，会对标注过程造成一定干扰。因此一般使用 TXT 文件，这就要求系统支持 TXT 格式的原始数据上传。同时，为了在标注过程中区分不同用户的标注数据，采用 JSON 格式进行存储。

3.2 功能需求

3.2.1 自定义标注配置 不同医疗机构以及相同医疗机构不同科室电子病历的结构和内容均存在一定差异[2]，应针对不同电子病历文本构建不同标注配置文件。

3.2.2 用户权限管理 电子病历标注过程需要多人参与。由于不同标注人员的标注结果无法完全一致，需要具有更高权限的管理员进行审核，并确定统一的标注结果。

3.2.3 辅助标注 标注数据是一项重要的基础性工作，力求精准、高效，为后续数据处理奠定良好基础[3]。现有人工标注方式主要依赖专家，一般能保证准确性，但是效率较低。而面向中文电子病历的自动标注效果不理想，需要人工介入。可将人工与机器学习方法相结合，实现自动化辅助电子病历标注，提升标注质量和效率。标注系统应着重寻求标注效率与准确性之间的平衡，一方面确保标注的准确性，另一方面为用户提供清晰、及时的反馈信息，从而有效推动语料标注的迭代优化进程。

3.2.4 标注状态流转 数据文件标注通常要经历多次修改与审核，其标注文件的状态可分为标注未完成、标注完成待审核、审核中未锁定、已锁定。需要对这 4 种不同阶段的标注文件状态进行区分。

3.2.5 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。每轮标注完成后，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时可了解一份标注数据中多名用户的标注一致性，还可以通过标注一致性辅助判断标注存疑的实体。

4 系统设计与实现

4.1 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，见图 1。数据层为了支撑文本病历标注分析，引入语料、模型和字典数据，其中语料数据保存原始数据、标注数据和审核数据，记录电子病历数据在标注过程中的状态。服务层使用基于主动学习技术及条件随机场（conditional random fields，CRF）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，面向用户提供实体标注功能，包括实体配置、语料推荐、实体标注、用户管理、辅助标注等。

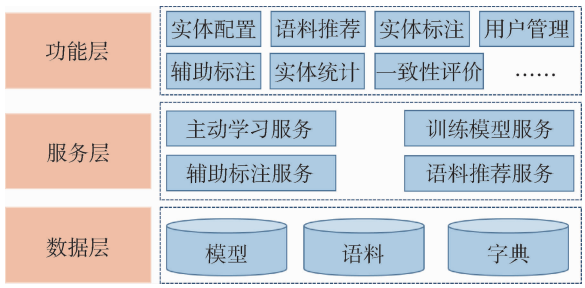

图 1 中文电子病历标注系统架构

4.2 用户权限管理模块

用户权限管理模块是中文电子病历标注系统的基础模块，包括多级角色设置，用户注册、激活、锁定、解锁，以及用户组分配等功能。用户分为系统管理员和普通用户，普通用户可拥有用户组管理员和普通成员身份。用户组由系统管理员创建，普通用户可同时拥有多个用户组身份，以参与不同项目，项目彼此之间数据隔离。系统管理员、用户组与用户的角色架构，见图 2。某用户可以是某用户组的成员，也可以单独存在，如用户 e、f。用户组 A 的管理员用户 a 可以同时为用户组 B、C、D 的成员，但同一用户组只能有唯一的用户组管理员，用户组管理员可以管理不同的用户组。

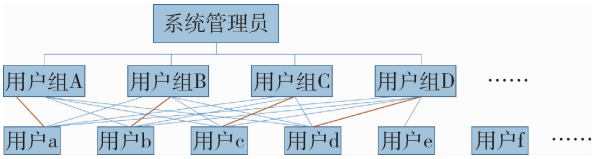

图 2 中文电子病历标注系统多级用户角色架构

4.3 实体和关系标注流程

参照流水线架构[4,19]设计实体和关系标注流程，主要包括项目创建、数据上传、实体和关系配置、语料标注与审核。用户按照推荐步骤操作，即可高效构建数据集。

4.3.1 项目创建 标注数据以项目形式呈现。用户首次登录后进入项目管理界面，浏览新手引导。创建项目时需填写项目名称、项目类型、项目描述、标注类型和推荐语料数量等信息。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。推荐语料数量指为标注者提供适量且高质量的标注任务，以优化标注流程并促进模型持续改进。系统用户项目角色权限，见表 2。

表 2 中文电子病历标注系统用户项目角色权限

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	√	√		
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）	√	√		

4.3.2 数据上传 项目创建者可上传 TXT 格式语料。上传后展示上传语料总文件数、上传失败文件数，超出屏幕范围的文件页面会自动折叠隐藏。

4.3.3 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对项目的实体关系配置进行更改，设置不同实体的名称、编码、颜色以及对应的快捷键位，点击保存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，以 JSON 编码形式进行批量更改，默认实体快捷键配置与 JSON 方式配置操作，见图 3。以“疾病”实体为例，JSON 方式配置中的“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”字段指定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。

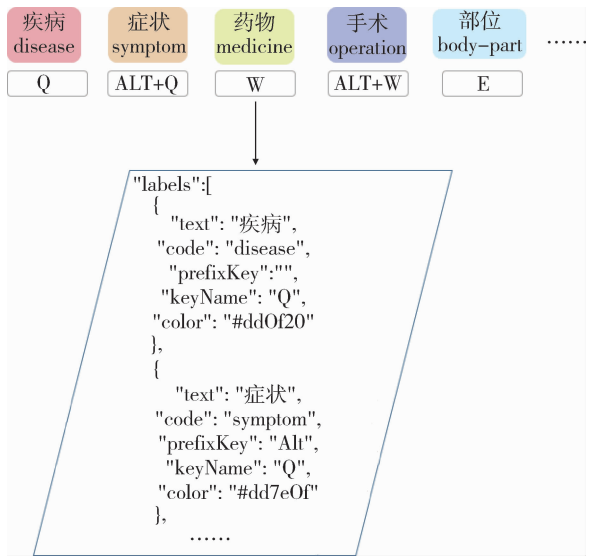

图 3 实体和关系快捷键配置

4.3.4 语料标注与审核 由项目创建者进行数据上传、实体配置，再由标注员对病历文本进行标注。系统标注流程中的数据状态变化，见图 4。

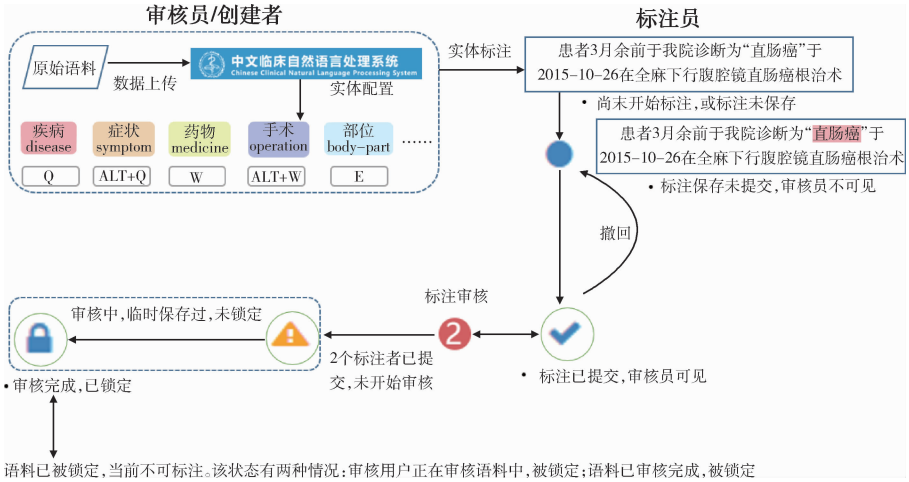

图 4 系统标注流程

4.4 关键技术

4.4.1 *主动学习* 基于电子病历实体识别的标注数据有限，但是医学相关数据标注对精度要求较高，因此采用主动学习方法，有目的地选择标注数据。标注系统从未标注数据中选择未充分训练的数据（即未标注语料中对标注价值最大的语料），交给用户进行标注，并将标注后的数据加入训练集中，进行下一次模型训练[21]。

4.4.2 *辅助标注* 辅助标注可减少重复性人工劳动，避免冗余标注，最大限度提升准确性与效率。先后采用两种不同的辅助标注方法，见图 5。第 1 种是基于字典匹配的辅助标注算法[22]，先人工标注一部分病历文本，审核完成后作为标准标注数据文件，并从中抽取已经标注的实体作为字典。部分词汇在不同文本中会被标注为不同实体，因此采用投票方式选取字典中的实体类型。第 2 种是基于 CRF 的辅助标注算法[23]，将第 1 种方法前期已标注的标准标注数据文件作为训练数据，训练完成后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进行修改，从而减轻负担，提高标注效率。基于实体识别性能较好的神经网络模型 CRF，采用上下文、字典、部首等作为特征；采用开源的 CRF + + 工具[24]作为算法实现工具；使用原始字、分词结果以及上下文（窗口大小为 5）中的信息训练 CRF 模型。

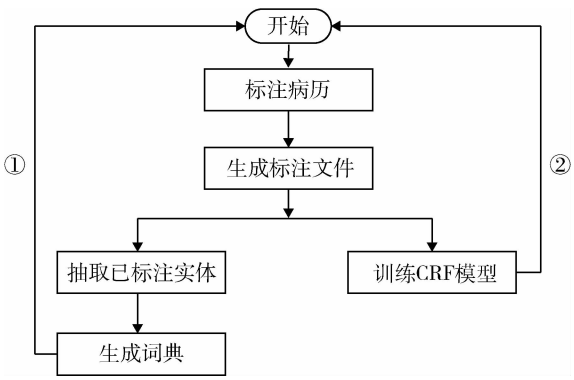

图 5 辅助标注流程

4.4.3 *一致性评价* 针对团队模式的项目，即多名用户标注同一语料的情况，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。系统根据标注数、一致数、P 值、R 值、*F*1 值指标给出一致性评价[4]，以表格呈现。标注者 a 与标注者 b 的 P 值、R 值、*F*1 值计算方式如下，最终以 *F*1 值作为多用户之间的一致性指标。

$$P = \frac{\text{a 和 b 标注一致数}}{\text{b 的标注数}} \quad (1)$$

$$R = \frac{\text{a 和 b 标注一致数}}{\text{a 的标注数}} \quad (2)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (3)$$

4.5 系统应用实践

中文电子病历标注系统上线至今，已完成 500 份电子病历现病史、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标注。实体标注的标注一致性为 0.86。采用基于 BiLSTM - CRF[26]、BERT - BiLSTM - CRF[27]的方法与本文方法（基于字典匹配和 CRF）进行对比，评估实体识别效果[24,28-29]。实验表明，基于 BERT - BiLSTM - CRF 的方法效果最优，但会占用大量计算资源，且训练时间长，不利于多轮标注迭代反馈。本文方法结合词性、部首和文档类型特征，*F*1 值仅比 BERT - BiLSTM - CRF 小 1 个百分点，但速度更快、系统资源占比更小。因此平台最终仍采用 CRF 算法实现辅助标注。

5 结语

本文构建可动态配置的中文电子病历标注系统，能够自动进行病历数据分析与信息抽取，在有效减少人工重复劳动的同时提高标注准确性。该系统可应用于医疗数据集构建、辅助疾病诊断相关分组（diagnosis related groups，DRGs）、临床路径优化等。目前系统仅提供实体和关系两种类型数据集构建功能，而完整的医学自然语言处理研究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大语言模型评测等[30-32]。未来将探索更多医学自然语言处理应用，通过大语言模型预先自动标注等方式提升数据标注的效率与准确性。

作者贡献：赵琬清负责需求分析、论文撰写；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；方安负责提供指导。

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 (1): 4302425.

3 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.

4 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.

5 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 (8)：1537 - 1562.

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 (8): 104432.

8 HAN X, WANG Z, ZHANG J, et al. Overview of the CCKS 2019 knowledge graph evaluation track: entity, relation, event and QA [EB/OL]. [2025 - 03 - 18]. https://arxiv.org/abs/2003.03875.

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 (3)：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 (10): 104535.

11 STENETORP P, PYYSALO S, TOPIĆ G, et al. Brat: a web - based tool for NLP - assisted text annotation [C]. Avignon: The Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012.

12 NAKAYAMA H, KUBO T, KAMURA J, et al. Doccano: text annotation tool for human [EB/OL]. [2025 - 03 - 18]. https://github.com/doccano/doccano.

13 TKACHENKO M, MALYUK M, HOLMANYUK A, et al. Label studio: data labeling software [EB/OL]. [2025 - 03 - 18]. https://github.com/heartexlabs/label - studio.

14 YANG J, ZHANG Y, LI L. YEDDA: a lightweight collaborative text span annotation tool [C]. Melbourne: Association for Computational Linguistics, 2018.

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 Poplar: a web - based annotation tool for natural language processing (NLP)[EB/OL]. [2025 - 03 - 18]. https://github.com/synyi/poplar.
17 SpaCy: industrial - strength natural language processing (NLP) in Python [EB/OL]. [2025 - 03 - 18]. 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 - 18]. 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 (8): 102573.
20 MERKEL D. Docker: lightweight linux containers for consistent development and deployment [J]. Linux journal, 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 J H, PRATT V R, et al. 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]. San Francisco: The Eighteenth International Conference on Machine Learning (ICML 01), 2001.
24 KUDO T. CRF + +: yet another CRF toolkit [EB/OL]. [2025 - 03 - 04]. https://taku910.github.io/crfpp/.
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 [EB/OL]. [2025 - 03 - 18]. https://arxiv.org/abs/1508.01991.
27 DEVLIN J, CHANG M W, LEE K, et al. Bert: pre - training of deep bidirectional transformers for language understanding [C]. Minneapolis: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.
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 - 18]. 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 - 18]. https://github.com/macanv/BERT - BiLSTM - CRF - NER.
30 ZONG H, WU R, CHA J, et al. Advancing Chinese biomedical text mining with community challenges [J]. Journal of biomedical informatics, 2024, 157 (9): 104716.
31 FLORIDI L, CHIRIATTI M. GPT - 3: its nature, scope, limits, and consequences [J]. Minds and machines, 2020, 30 (11): 681 - 694.
32 GUO D, YANG D, ZHANG H, et al. Deepseek - r1: incentivizing reasoning capability in llms via reinforcement learning [EB/OL]. [2025 - 03 - 18]. https://arxiv.org/abs/2501.12948.