

面向态靶辨治的语料库标注系统设计与实现

蒋欣然^{1,2}, 董富江¹, 苏雪^{1,3}, 苟春燕¹, 张文学^{1*}

¹宁夏医科大学医学信息与工程学院, 宁夏 银川

²宁夏医科大学中医学院, 宁夏 银川

³宁夏回族自治区吴忠市同心县豫海镇人民政府, 宁夏 吴忠

收稿日期: 2025年6月14日; 录用日期: 2025年7月11日; 发布日期: 2025年7月17日

摘要

针对中医态靶辨治领域非结构化文本阻碍知识挖掘的问题, 本研究提出一种面向对象的语料库标注系统解决方案。基于筛选自CNKI的32篇态靶辨治核心文献, 系统化定义10类实体(如疾病、标靶、症型态)与7类实体关系(如打靶、靶药、药理), 并设计多层次标注框架: 实体标注、实体属性标注、实体关系标注。系统实现了可视化的标注界面与BIE结构化的输出, 支持人工高效构建高质量标注语料库。本工作既可以为后续基于深度学习的态靶辨治文本标注提供训练数据集, 又能为中医同类标注系统研发提供参考。

关键词

态靶辨治, 语料库标注, 知识结构化, 面向对象设计, 中医人工智能

Design and Implementation of a Corpus Annotation System for State-Target Differentiation and Treatment

Xinran Jiang^{1,2}, Fujiang Dong¹, Xue Su^{1,3}, Chunyan Gou¹, Wenxue Zhang^{1*}

¹School of Medical Information and Engineering, Ningxia Medical University, Yinchuan Ningxia

²School of Traditional Chinese Medicine, Ningxia Medical University, Yinchuan Ningxia

³The People's Government of Yuhai Town, Tongxin County, Wuzhong City, Ningxia Hui Autonomous Region, Wuzhong Ningxia

Received: Jun. 14th, 2025; accepted: Jul. 11th, 2025; published: Jul. 17th, 2025

Abstract

To address the challenge of unstructured text hindering knowledge mining in the domain of State-

*通讯作者。

文章引用: 蒋欣然, 董富江, 苏雪, 苟春燕, 张文学. 面向态靶辨治的语料库标注系统设计与实现[J]. 计算机科学与应用, 2025, 15(7): 40-48. DOI: 10.12677/csa.2025.157178

Target Differentiation and Treatment (STDT) of Traditional Chinese Medicine, this study proposes an object-oriented corpus annotation system solution. Based on 32 core STDT literature pieces curated from CNKI, we systematically define 10 entity categories (e.g., disease, clinical indicator, syndrome state) and 7 entity relationships (e.g., precision targeting, target-specific herb, pharmacological mechanism). A three-tier annotation framework is designed: entity annotation, entity attribute annotation, and entity relationship annotation. The system implements a visualized annotation interface with BIE-structured output, enabling efficient manual construction of high-quality annotated corpora. It not only provides a training dataset for deep learning-based STDT text annotation, but also serves as a reference framework for developing analogous TCM annotation systems.

Keywords

State-Target Differentiation and Treatment, Corpus Annotation, Knowledge Structuring, Object-Oriented Design, Artificial Intelligence for Traditional Chinese Medicine

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来,随着我国医疗专业数据库建设流程的完善,海量医疗数据储存与调用的需求日益增强。然而,传统中医由于自身辨证视角与现代医学的差异,其数据挖掘的环节仍存在阻碍。中医全小林院士[1]针对中医药传承发展中存在的“态强靶弱”等问题提出“态靶辨治(State-Target Differentiation & Treatment, STDT)”,该理论作为一种中西医结合辨证论治理论的创新尝试,将调整人体内环境与发挥中药打靶作用相结合,取得了良好的临床疗效。然而,随着 STDT 临床应用与实验研究的深入,其现存文献中的非结构化文本增多的问题阻碍了领域知识储存与数据挖掘,需通过构建专业语料库来解决。而数据标注是语料库构建的核心环节之一,本研究通过设计一款面向中医态靶辨治语料库的标注系统,实现对文献数据的人工标注,旨在提高人工标注质量与效率。

本标注系统采用面向对象的系统构建方法,通过需求分析、系统分析、系统设计三个基本步骤,实现标注系统包括“实体标注、实体属性标注、实体关系标注”等核心功能的功能模块,辅助态靶辨治文本语义识别与知识抽取环节。不仅为后续基于深度学习的自动标注方法提供训练数据集,还为同类系统的开发提供一定参考。

2. 研究现状

李致重[2]认为“态靶辨证”弥补了中药治疗疾病态强靶弱的不足,促进“中医宏观”与“西医微观”的有力结合。林轶群等[3]构建态靶辨治体系进一步丰富传统中医对现代疾病的认识,并挖掘诊疗疾病的新模式,提炼新靶方靶药,加快疾病的治疗。何莉莎等[4]在中医治疗中,实施了“调态”策略,基于药物的独特性精准调整疾病造成的失衡状态,从而显著提升治疗的靶向性和效果。委李楠等[5]在中医语料库构建的研究上提到当前中医汉语语料库发展受到医古文与现代用语混用、专业术语不统一、通用词汇与教学用语有差异等阻碍,其自然语料库的构建仍需要多学科合作。刘丽红等[6]针对中医药古籍文献进行语义标注,以国标文件中的中医药学语言系统的语义网络框架为标准,探讨中医临床诊疗方案中术语间的关系,在制定标注规范和标注流程方面提供了相对完整且严谨的方法参考。张仕娜等[7]针对中医厥证这一具体的疾病进行领域本体的构建,利用本体技术推动中医领域知识的融合与推理,使得标注规范

的制定更具专业性。

标注的语料质量好坏对语料库的影响是巨大的，如何获得人工高质量的标注语料是当下亟需研究的问题[8]。标注系统在中医领域的应用旨在提供准确、高效的语义标注和知识抽取。通过使用先进的自然语言处理技术，标注系统能够自动化地识别和理解中医文献中的概念、关系和模式，从而形成结构化、规范化的知识体系[9]。随着大数据和人工智能技术的不断发展，标注系统在中医领域的应用前景将更加广阔，为中医的现代化和智能化发展提供重要支撑[10]。

3. 态靶辨治语料库的标注系统分析与设计

完善的中医药文本数据的标注系统助力于临床医疗决策，并为中医科研活动提供开放的数据处理系统，是中西医精准化、数字化的重要发展环节。态靶辨治语料库的标注系统的设计与实现是在态靶辨治中医文本标注规范制定以及人工标注的基础上，将标注过程及其结果可视化，有效提高人工标注的效率及准确性，实现医学文本数据在专业数据系统的集中展示及应用，为态靶辨治理论的发展提供临床决策支持以及科研数据的中转。

3.1. 态靶辨治语料的获取与预处理

1) 语料获取。在中国知网平台，通过“态靶辨治”、“态靶”等关键词多次检索，最终筛选出 32 篇高度相关的文献。将这些文献下载并妥善保存，以备后续深入分析和使用。

2) 数据预处理。数据预处理环节包含文件格式转换，格式统一为 txt；清洗数据噪音，删除空白字符、多余标点符号、空白行等；文档语句分割，单一文本按 100 个句子分成若干个要注释的样例文件，其语句的切分识别码为句号、叹号、问号、双引号等。

3.2. 态靶辨治语料的标注体系设计

语料库标注体系主要分为实体类型、实体关系、实体属性 3 类。本研究围绕筛选出的数据以及态靶辨治语料特征筛选出 10 个具有代表性的个体作为核心实体。这些实体分别为：疾病、中药、症状(证靶)、症状-体征、症状-脉象、症型(态)、时间、功效、方剂以及标靶。这些实体被归类为八类主要的实体类别。其中，症状(证靶)因其复杂性和多样性，进一步细分为子类实体。详见下表 1。

Table 1. List of entity types for STDT
表 1. 态靶辨治实体类型列表

序号	英文名称	实体类	实体子类	示例
1	Disease	疾病		糖尿病、高血压、肥胖
2	Symptom	症状		乏力、多汗、手足心热
3	Tongue & pulse		舌脉	舌红脉滑数、舌淡苔白脉细
4	Sign		体征	血压偏高、血糖偏高
5	Effect	功效		清热化痰、疏肝理气
6	Prescription	方剂		大柴胡汤、葛根芩连汤
7	Medical	中药		黄连、生山楂、炒谷芽
8	Time	时间		10 年、2023 年 6 月 4 日
9	Syndrome type	态		热态、寒热错杂态
10	Target	标靶		糖化血红蛋白、血糖、血压

语料实体之间并非孤立存在，而是通过各种关系连接，例如“中药”治疗“疾病”；同时各实体有自身属性，例如实体“中药”有各自的“性味归经”，可以称之为“关系属性”。本研究筛选出的关系有“反应”、“治疗”、“打靶”等。详见表 2。

Table 2. List of entity relationships for STDT
表 2. 态靶辨治的实体关系列表

序号	关系名称	主体与客体	示例
1	反应	<症状, 疾病>	头晕反应高血压
2	治疗	<方剂, 疾病>	葛根芩连汤治疗糖尿病
3	打靶	<中药, 靶靶>	黄连打靶降血糖
4	靶药	<中药, 症状>	生姜是止呕靶药
5	组方	<方剂, 中药>	葛根芩连汤组方为葛根、黄芩、黄连等
6	药理	<方剂, 功效>	大柴胡汤疏肝理气
7	作用	<中药, 功效>	黄连清心火

3.3. 系统分析

系统用例是面向对象的系统构建方法的需求工程中用于明确系统边界、定义用户与系统交互逻辑的核心工具。态靶辨治语料库标注系统的系统用例详见下图 1，主要包含态靶辨治实体标准、态靶辨治属性标注和态靶辨治关系标注。

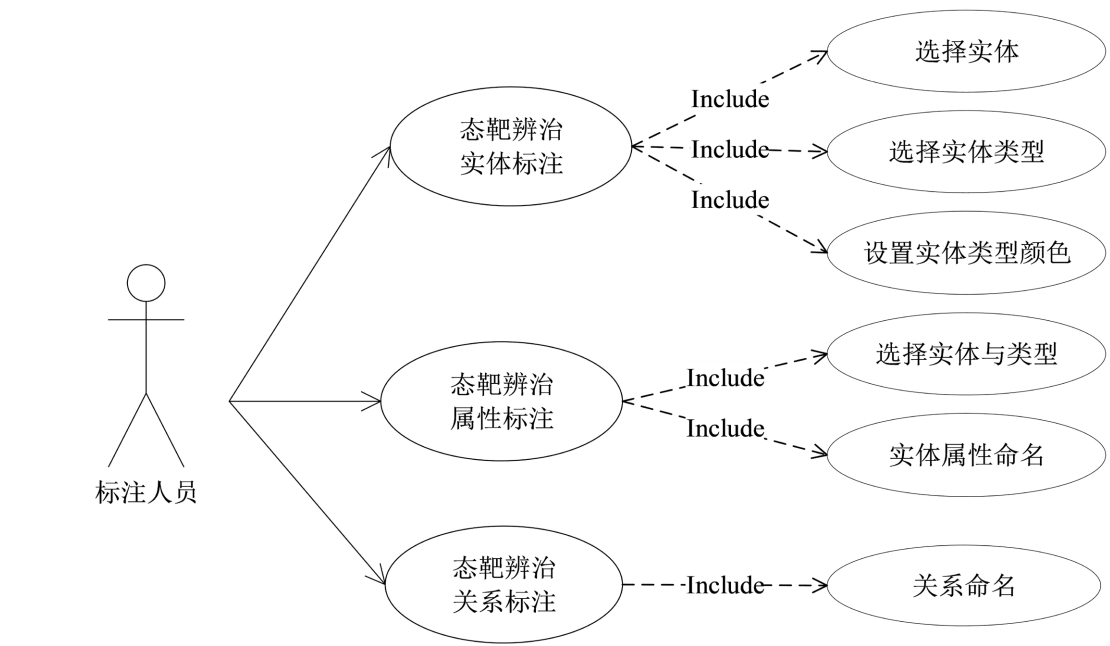


Figure 1. System use case diagram of the “annotation system for STDT”
图 1. “态靶辨治标注系统” 的系统用例图

3.4. 系统功能设计

在构建面向态靶辨治语料库的标注系统时，设计了 3 个核心模块：实体标注模块、实体关系标注模

块和标注结果输出模块。实体标注模块专注于实现精确的实体类型标注以及属性标注，实体关系标注模块能够实现标注态靶辨治实体关系，标注结果输出模块则负责将标注结果以直观的方式展现出来，如下图 2 所示。

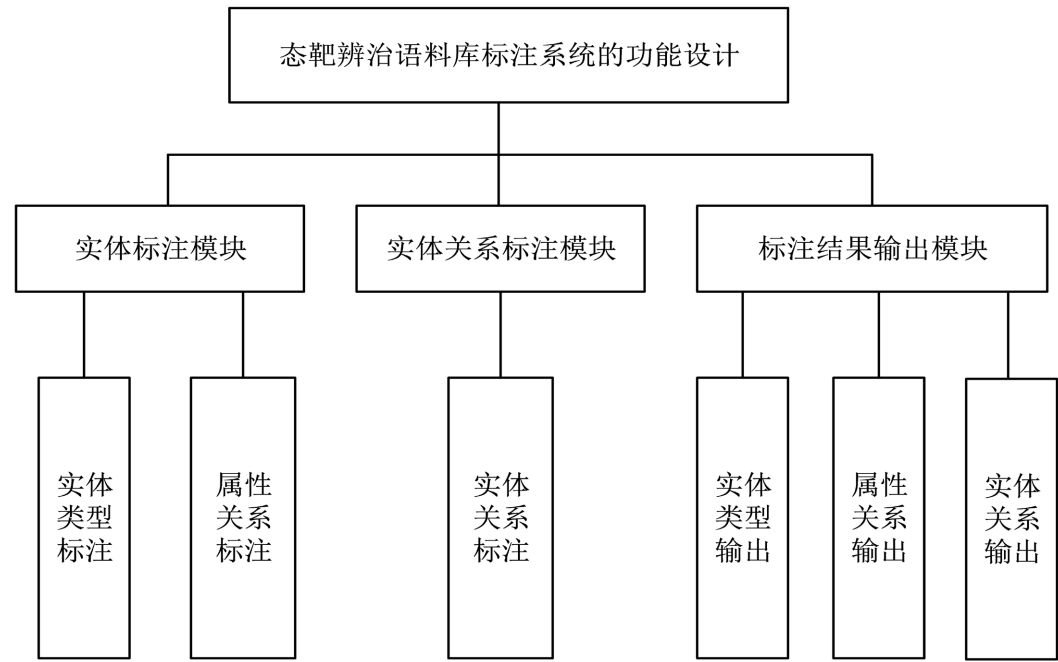


Figure 2. System function module design
图 2. 系统功能模块设计

4. 态靶辨治语料库标注系统的实现

4.1. 标注界面

在系统界面中，为了提升操作便捷性，设置了两个选择按钮，分别用于实现文件的导入功能和数据的复制进入功能。为了方便管理已标注的数据信息，我们将标注数据通过直观的页面形式展示，确保用户能够实时了解数据状态。

输入并标记文本后，选择特定的词语进行标注，通过不同的颜色来区分这些词语所具备的不同属性，以便更直观地呈现。实体类型方面选择了 9 种类型进行标注，分别为：“Disease (疾病)”、“Medical (中药)”、“Symptom (症靶)”、“Signs (症状 - 体征)”、“Tongue & pulse (舌脉)”、“Syndrome type (态)”、“Time (时间)”、“Effect (功效)”、“Prescription (方剂)”。具体的标注界面如图 3 所示。

4.2. 标注结果界面

本标注系统采用 BIE 标注方法，将标注出的实体用“B”、“I”、“E”来进行表示，例如选择的词语是“化纤散”，“B”是第一个字“化”，“I”是中间字“纤”，“E”是末尾字“散”，而“功效”是该实体的属性之一。详见下图 4。

建立两个实体类型之间的关系。标注体系模块选取了 7 个实体关系对文本进行标注，分别为：“Reflect (反映)”、“Treat (治疗)”、“Shooting (打靶)”、“Target drugs (靶药)”、“Composition of formula (方剂的组成)”、“Pharmacology (药理)”、“Effect (作用)”。详见下图 5。

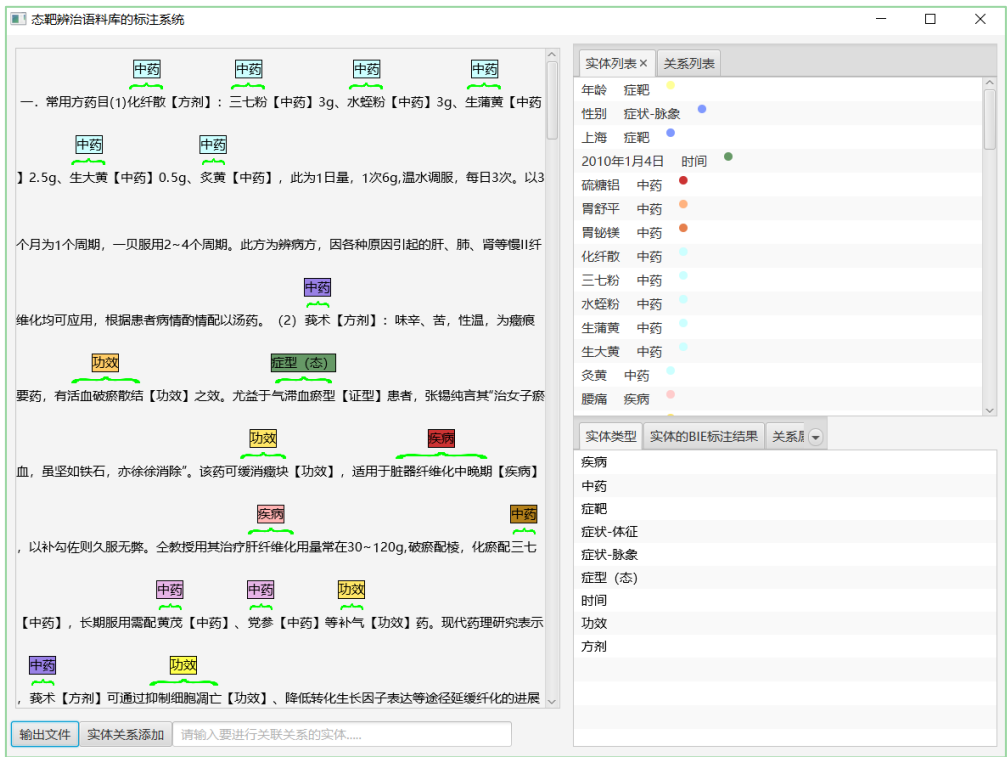


Figure 3. Main interface of the corpus annotation systems for STDT
图 3. 态靶辨治语料库的标注系统的主界面

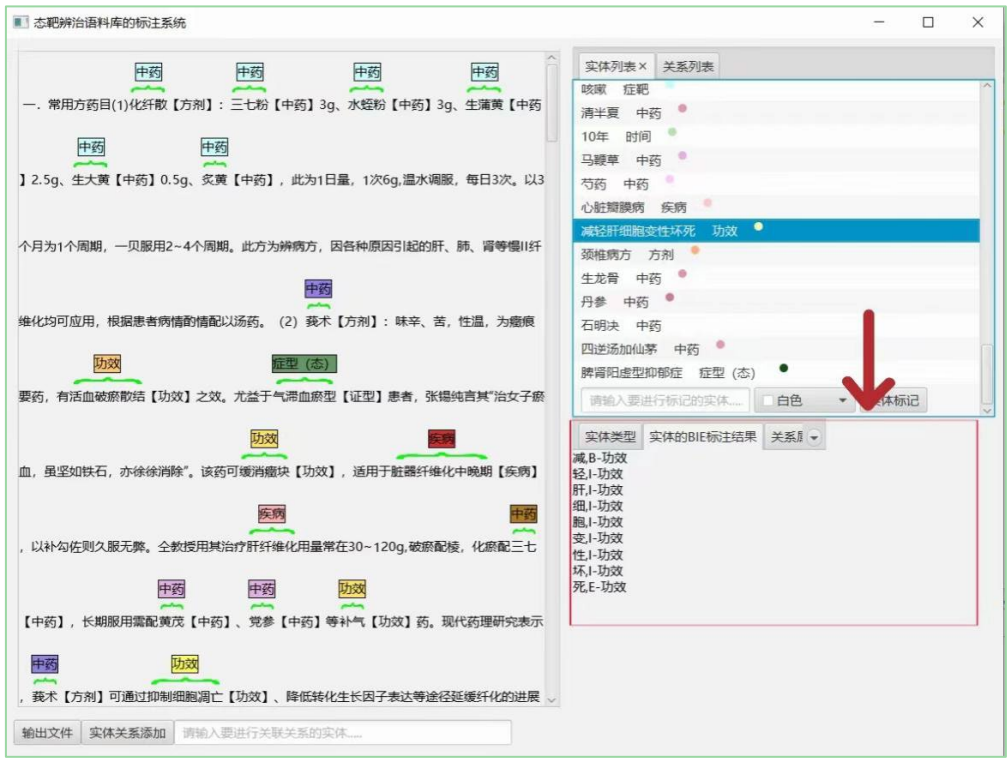


Figure 4. Results interface of annotation entity and entity attributes by BIE
图 4. BIE 标注实体与实体属性结果的界面

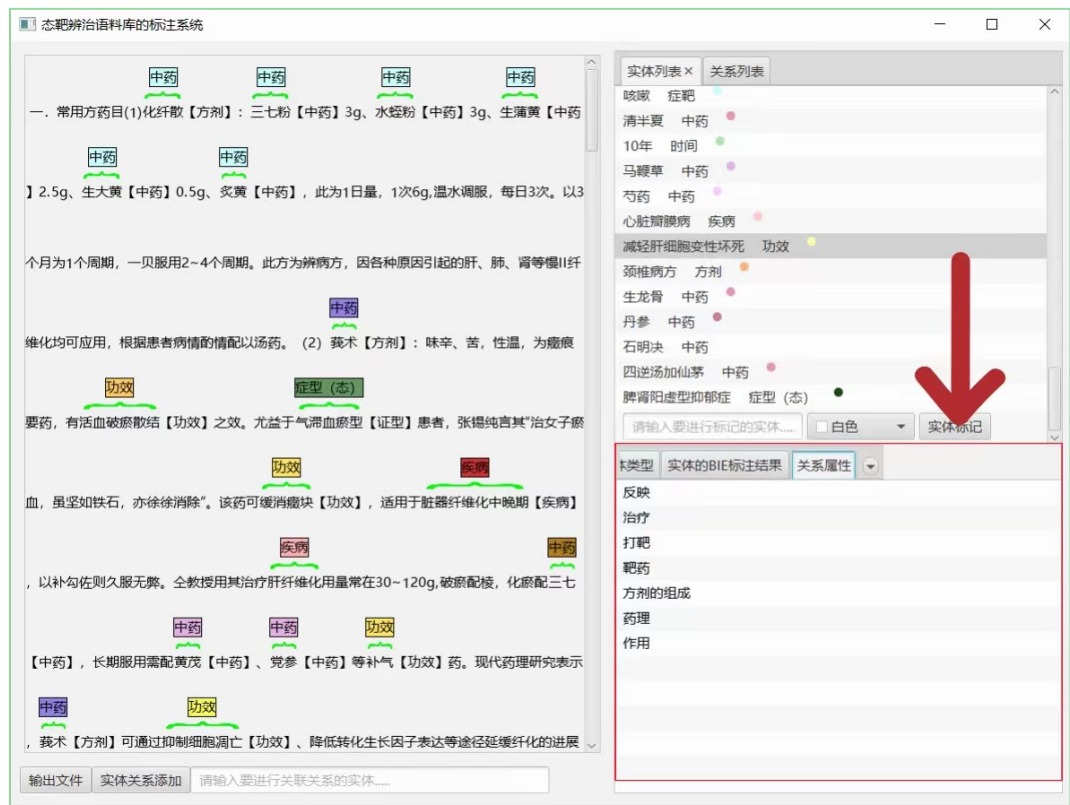


Figure 5. Results interface of annotation entity relationship
图 5. 标注实体关系结果界面

4.3. 标注结果输出界面

实体标注工作完成后，执行输出操作，以获取实体标注及其关系。输出的实体标注应包含输出的排列顺序、实体类型、该实体在文本中的起始和结束位置信息，以及实体的名称。详见图 6。

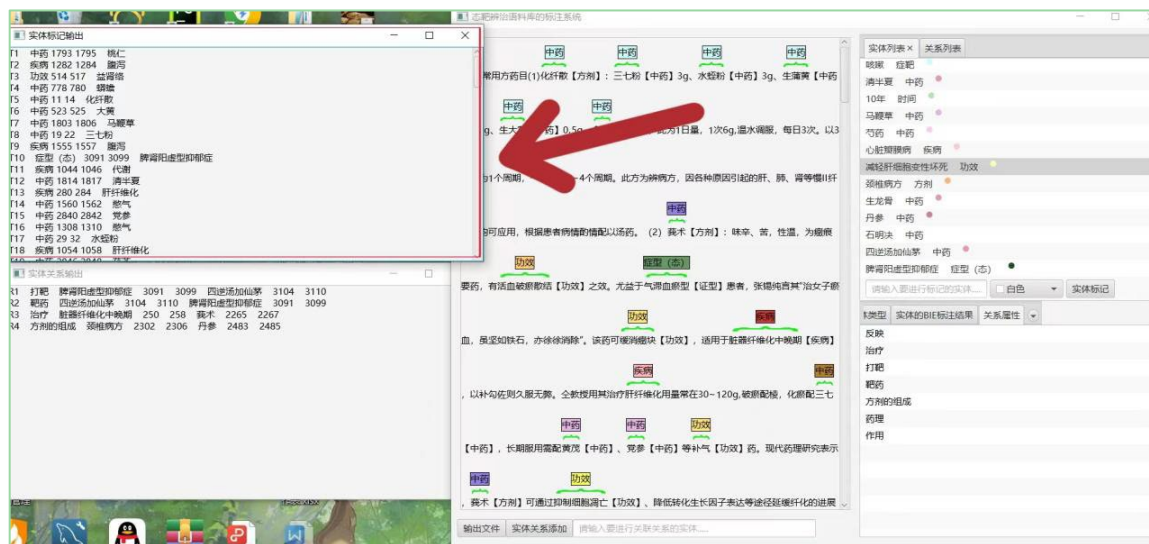


Figure 6. Output interface of the entity annotation results
图 6. 实体标注结果的输出界面

实体关系标注工作完成后，执行输出操作，以获取实体关系。详见图 7。

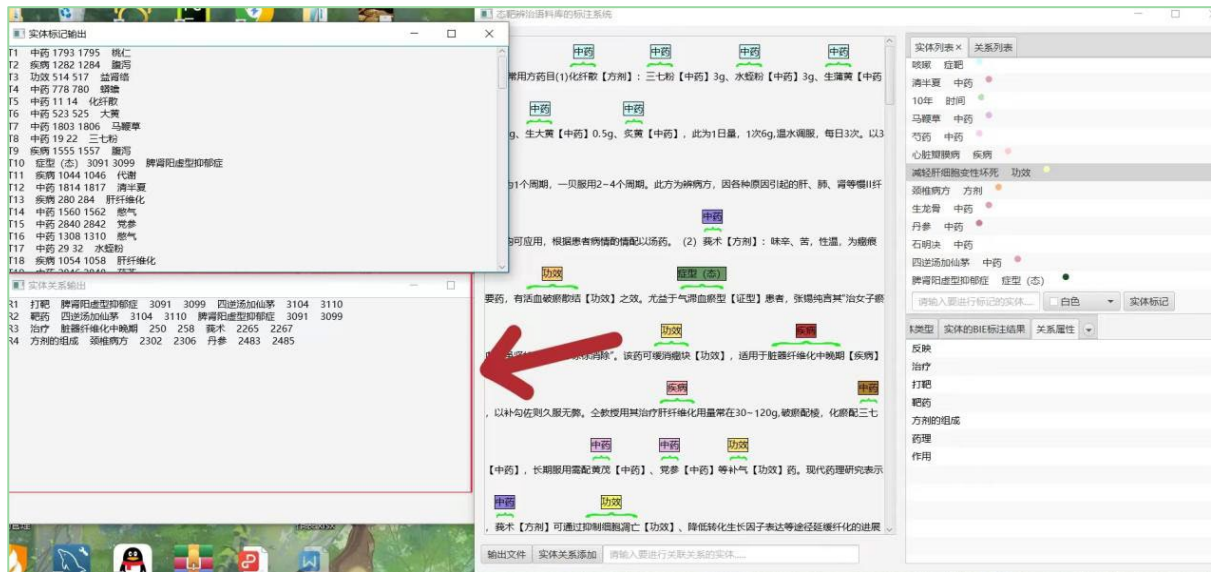


Figure 7. Output interface of annotation results for the entity relationship
图 7. 实体关系标注结果的输出界面

4.4. 讨论

当前主流标注工具(如 Brat、Prodigy、Label Studio)在中医文本处理中存在一些局限: Brat 支持实体识别、实体关系、事件抽取、方面级情感分析等标注任务,适配中文标注场景,但安装环境需要为 osx 或 linux 系统或 linux 虚拟环境,同时依赖静态配置、缺乏中医实体预设模板、需手动定义全部标签(如“靶靶”、“舌脉”);增加学习成本;Prodigy 的强大之处在于它有一套智能算法,可以减少人工的重复劳动,但是需付费授权且未适配中医语义关系(如“打靶 - 标靶”)层级关系);Label Studio 提供了多模态数据支持、丰富的可视化界面以及自定义标注模板的能力,但对中医嵌套实体(如“热态→手足心热”)支持不足,关系标注仅限二元组、缺乏属性标注。

本系统的创新性在于:预置中医态靶辨治领域的实体及其关系库,直接加载预定义实体类和实体关系;多层次动态标注框架,同步支持实体 - 属性 - 关系标注;BIE 结构化输出,兼容深度学习模型输入格式;采用 Java 语言开发,可跨平台部署,操作界面简单友好。

本系统的未来工作展望:借鉴 Prodigy 自动预标注优势,结合预训练模型并集成 Deepseek 大模型进行半自动标注。一是标注方案包括自然语言处理信息抽取智能标注、整合中医本体进行知识扩展、本系统的数据标注、模型微调、智能标注(含自动训练),有效扩展模型泛化能力。二是创新应用场景,拓展应用边界,形成数据标注 - 算法训练 - 应用场景,例如创新态靶辨治知识图谱构建、知识图谱补全、知识图谱质控等应用场景,通过态靶辨治临床深度耦合实现数据价值指数级释放。三是创新基于 Web 的高通量协作标注模式,模型部署后端 SDK 包装机器学习代码并将其转换为 Web 服务器,Web 服务器可以连接到正在运行的实例,以自动执行标记任务实现开拓数据生产新型组织方式。支持多人、多中心协同、数百人团队协作,解决大规模数据搬迁和协调效率问题。

5. 结束语

本研究面向中医态靶辨治的语料库标注系统,通过实体 - 属性 - 关系三层标注架构,解决了中医态

靶辨治文本结构化难题。该系统基于严谨的标注规范和面向对象设计，提升了标注效率与一致性，为中医态靶辨治知识图谱构建提供核心数据支撑。当前局限集中于人工标注成本及细粒度语义覆盖。未来工作将聚焦三方面：结合预训练模型和大模型开发半自动标注工具，整合中医本体深化语义关系进行知识扩展，对接临床决策系统以验证标注知识对靶方推荐的有效性。

基金项目

宁夏自然科学基金：面向全小林院士态靶辨治的中医知识图谱构建研究(2023AAC03165)，态靶辨治不确定性知识的语料库构建研究(2024AAC03214)。

参考文献

- [1] 全小林. 态靶医学——中医未来发展之路[J]. 中国中西医结合杂志, 2021, 41(1): 16-18.
- [2] 李致重. 中西医防治观之比较[J]. 中医药通报, 2023, 22(5): 1-4.
- [3] 林轶群, 赵林华, 王强, 等. 代谢综合征态靶辨治体系的构建[J]. 中医杂志, 2022, 63(13): 1223-1226.
- [4] 何莉莎, 顾成娟, 王涵, 等. 态靶结合辨治代谢性高血压病[J]. 中医杂志, 2019, 60(16): 1423-1424+1427.
- [5] 委李楠, 张丽, 薄彤. 浅谈中医汉语语料库的建设[J]. 中国中医药现代远程教育, 2023, 21(24): 7-9.
- [6] 刘丽红, 付璐, 姚克宇, 等. 中医药古籍文献实体标注规范探索[J]. 中华医学图书情报杂志, 2022, 31(12): 1-6.
- [7] 张仕娜, 高远, 郑爱华, 等. 中医厥证领域本体构建研究[J]. 湖南中医药大学学报, 2024, 44(3): 427-434.
- [8] 朱彦, 乔幸潮, 崔一迪, 等. 中医药文献语义标注系统研究与开发[J]. 中国中医药图书情报杂志, 2020, 44(3): 5-8.
- [9] 杨洋, 关毅, 李雪, 等. 中文医学细粒度知识表示体系与标注语料库构建[J]. 中文信息学报, 2023, 37(6): 52-66.
- [10] Li, H.L., Pei, X.M., Yu, H., Wang, W. and Mao, D.G. (2024) Autophagic and Apoptotic Proteins in Goat Corpus Luteum and the Effect of Adiponectin/AdipoRon on Luteal Cell Autophagy and Apoptosis. *Theriogenology*, **214**, 245-256. <https://doi.org/10.1016/j.theriogenology.2023.11.001>