

【引文格式】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.

• 信息技术与中医药 •

中医药文献语义标注系统研究与开发

朱彦¹, 乔幸潮¹, 崔一迪¹, 高曼¹, 高博¹, 王俊慧², 尹仁芳^{1*}

1. 中国中医科学院中医药信息研究所, 北京 100700; 2. 中国中医科学院广安门医院, 北京 100053

摘要: 目的 研究和开发支持中医和现代生物医学本体和术语集的语义标注系统。方法 以 MedPortal 本体库和中医临床术语集等为术语资源库, 设计语义标注系统工作流程和功能框架, 并开发 Web 应用系统。结果 构建了一个基于 Web 的中医药文献语义标注系统, 支持语料库管理与维护、术语词典管理、语义标注和语义检索等功能, 既可以为基于机器学习的信息抽取算法研究提供训练集, 又能实现语义层面的多来源数据集成与知识融合。结论 该中医药文献语义标注系统设计方案已经过实际项目验证, 可为其他同类系统研发提供参考。

关键词: 中医药; 文献; 语义标注; 系统开发

中图分类号: R2-03 文献标识码: A 文章编号: 2095-5707(2020)03-0005-04

DOI: 10.3969/j.issn.2095-5707.2020.03.002

开放科学(资源服务)标识码(OSID):



Research and Development of Semantic Annotation System for TCM Literature

ZHU Yan¹, QIAO Xing-chao¹, CUI Yi-di¹, GAO Man¹, GAO Bo¹, WANG Jun-hui², YIN Ren-fang^{1*}

(1. Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China; 2. Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing 100053, China)

Abstract: Objective To research and develop a semantic annotation system supporting ontology and terminology of TCM and modern biomedicine. **Methods** Taking MedPortal ontology repository and TCM clinical terminology system as term resources, the workflow and functional framework of semantic annotation system were designed and a web application system was developed. **Results** A web-based semantic annotation system was built, supporting corpus management and maintenance, term dictionary management, semantic annotation and semantic retrieval, etc., which could not only provide training set for the research of information extraction algorithm based on machine learning, but also realize the multi-source data integration and knowledge fusion at the semantic level. **Conclusion** The design scheme of the semantic annotation system of TCM literature has been verified by actual projects, and can provide a reference for the development of other similar systems.

Key words: TCM; literature; semantic annotation; system development

基金项目: 国家自然科学基金(61701546); 中央级公益性科研院所基本科研业务费专项资金资助(ZZ13-YQ-126, ZZ13-YQ-127, ZZ13-YQ-021); 中国中医科学院基本科研业务费自主选题(ZZ130306)

第一作者: 朱彦, E-mail: zhuy@mail.cintcm.ac.cn

*通讯作者: 尹仁芳, E-mail: yinrf@mail.cintcm.ac.cn

随着生物医学的发展, 积累了越来越多的文献资料。文献数据是研究者获取知识的重要来源, 而方便快捷地从积累的文献中检索和获取知识至关重要。由于文献数据的增加, 且大部分为非结构化数据, 给知识检索和数据挖掘增加了很大难度。语义标注就是在本体和资源之间构建联系, 将数据智能化^[1], 是非结构化数据转化为结构化数据的重要过程。

在生物医学领域，已有不少支持英文的标注系统开发与应用的成功案例，例如，美国国立卫生研究院（USA National Institutes of Health, NIH）开发的 PubTator Central^[2]。另外，基于本体的维基百科知识库^[3]、临床文本语料库等^[4-5]也推动了本体在语义标注领域的应用。近年来，国内生物医学领域也在本体构建及社区建设方面发展很快，例如，“本体中国”社区及其本体库 MedPortal^[6]。进行中文语义标注的重要前提是中文的本体术语资源。中医药领域也在本体研究和术语构建方面取得了多项成果，包括中医临床术语系统^[7-8]、中医药学语言系统（Traditional Chinese Medicine Language System, TCMLS）^[9]等，这些本体术语资源为中文语义标注系统提供了很好的基础。

虽然国内对于语料标注领域一直都很关注并有一些研究^[10-14]，但现有开放的、可用的中文标注系统，尤其是支持中医和现代生物医学本体和术语集的标注系统未见文献报道。本研究以 MedPortal 本体库和中医临床术语集等为术语来源，研究构建基于自然语言处理的半自动化中医药文献标注系统。

1 中医药文献语义标注系统设计与实现

中医药文献语义标注系统（以下简称“本系统”）面向中医药领域文献自然语言处理的需求，构建基本语料库，同时也支持语料库的持续维护。可以支持导入 Excel 格式词汇表，也可以支持导入 OWL（Web Ontology Language）数据。本系统采用 MedPortal 的中英文本体及中医主题词表、中医临床术语系统^[7]等作为标注术语基础，共纳入本体 47 个，术语 100 多万条。

1.1 中医药文献语义标注系统工作流程设计

(1)文本处理：将要标注的文本按照篇章结构整理出目录，转换为标准的 Excel 格式的文档，导入到标注系统中。(2)词典管理：按照需要建立实体类型，并从 MedPortal 和中医领域的术语集中导入术语，建立词典。(3)标注：基于术语词典，使用自然语言处理算法（Natural Language Processing, NLP）进行自动标注，并进行人工校对及补充。(4)导出：标注结果可以导出开发格式的语料库，标注过程中识别出的新术语，可以扩充至系统的术语词典（见图 1）。

1.2 中医药文献语义标注系统功能设计

本系统具备以下 5 个功能：语料库管理、术语词典管理、文本标注、检索、后台管理（见图 2）。

1.2.1 语料库管理 包括文本管理及语料库导出。

(1)文本管理：本系统支持批量文本上传，原文档在准备前应将多级目录数据整理成 Excel 格式，然后上传到语料库。一个语料库中会包含若干个文档。导出的语料库作为后续机器学习的训练语料库。
(2)语料库导出：支持多种开放格式，如 xml 或 Brat^[15]等标准格式。

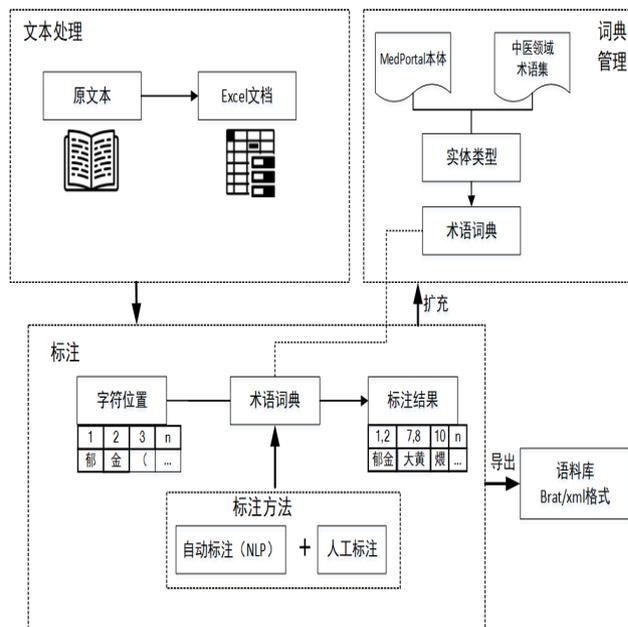


图 1 中医药文献语义标注系统工作流程设计图

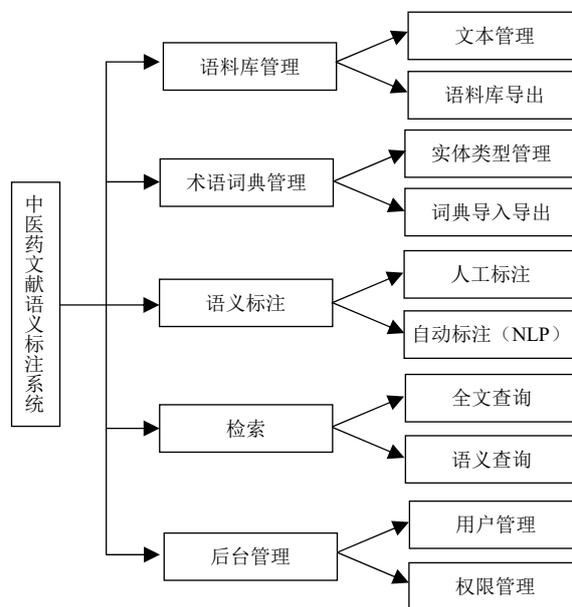


图 2 中医药文献语义标注系统功能模块图

1.2.2 术语词典管理 以语义类型单元来管理术语词典，基于医学一体化语言系统（Unified Medical Language System, UMLS）^[16]和 TCMLS^[9,17]中的语义类型，在系统中建立所需的语义类型，每个语义类型包含一个术语词典。

本系统支持 MedPortal 本体库的术语导入, 通过调用应用程序接口 (API) 与 MedPortal 建立链接。MedPortal 本体库基于美国国家生物医学本体中心 (National Center for Biomedical Ontology, NCBO) 的 BioPortal 系统的技术, 建立中国医学本体资源库, 整合中文与外文医学本体资源^[18]。目前, MedPortal 系统已整合生物医学本体 42 个, 建立了本体之间术语映射关系, 并通过页面和 REST API 方式, 提供术语检索、本体映射、数据标准化注释等本体应用服务。

用户选定要导入的本体或者本体中的某个实体

类, 就可以将所选本体或实体类下的所有子节点批量导入到系统中。为了方便用户查找, 本系统还支持基于术语或标识 ID 的检索、定位。例如, 用户输入 “cancer”, 能查询获得节点的 ID 地址为 “http://purl.obolibrary.org/obo/DOID_162”, 并展示出该 ID 的树形结构, 进一步将该实体及其子节点的实体循环迭代导入进来。另外, 与该术语为 “has_exact_synonym (有准确同义词)” 关系的同义词, 也可以全部导入进来。系统还支持 Excel 格式的术语批量导入。本系统中的术语词典包含术语 ID、语义类型、概念 ID 等信息 (见表 1)。

表 1 中医药文献语义标注系统术语词典包含的术语相关信息

术语 ID	语义类型	概念 ID	术语名称	是否为正式名称	网络链接	术语来源
1	中药	2027849999999104	莪术	是		中医临床术语系统
2	蛋白	PR_P29533	vascular cell adhesion protein 1 (mouse)	是	http://medportal.bmicc.cn/ontologies/PR?p=classes&conceptid=h ttp%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FPR_P29533	MedPortal
3	方剂	1881199999999101	安神补心汤	是		中医临床术语系统

1.2.3 语义标注 机器的初步标注都是依据词典自动标注完成的。在标注过程中, 可以选择某一类或几类实体, 得到机器自动初步标注结果。自动标注的结果是可视的, 会存在一些错误, 可以直接修改自动标注的结果, 也就是进行人工校对标注。人工校对的结果可以对词典进行更新, 用于下一个标注。这样的迭代对模型进行了训练, 对参数进行了调整, 提高了模型自动标注的效率。本系统也具有增加同义词的功能, 可以标注选择是否为首选词或同义词。

1.2.4 检索 除了传统的全文检索模式, 本系统还支持对标注结果进行语义检索。例如, 检索 “黄芪”, 也可以检索到标注了 “黄耆” 的文本段, 反之亦然。因为这两个术语对应的概念 ID 是相同的。

1.2.5 后台管理 包括用户注册、注销及权限管理等功能。

1.3 开发工具与开发环境

本系统开发使用 Java 计算机语言, 采用 MVC 设计模式和目前比较成熟的 SpringMVC+ Spring+ MyBatis 框架。前端使用 Html5 作为开发语言, 支持跨平台开发, 并使用了 Bootstrap 等开发框架和工具, 有效缩短了开发周期。

基于本系统设计方案, 实现了各个功能模块, 目前已投入应用到 “皮肤病古籍知识库” 建设项目中。基于该语义标注系统, 用户使用自动标注和人工审核相结合的模式, 提高了工作效率。不同实体的标注结果可视化效果如图 3 所示。



图 3 中医药文献语义标注系统语义标注可视化效果图

2 讨论

基于本体的语义标注有助于多源信息识别和处理^[19]。本体还使信息更加标准化,通过本体中的标准化术语,不同来源数据可以进行术语的统一,以消除认知差异,实现数据的整合与自动分析^[20]。

本文介绍了中医药文献语义标注系统的设计理念,并开发实现了该系统第一个版本,基本满足现有语义标注工作的需求。下一步将继续完善功能,包括支持实体关系的标注、实现供其他系统调用查询功能接口、支持领域内更多术语集和本体库等。

后续系统的应用场景包括:(1)面向自然语言处理的语料库构建。本系统能支持机器辅助人工的实体标注,并实现标注语料的管理功能,将标注好的语料按通用格式导出,可以为基于机器学习的信息抽取算法研究提供训练集,开发改进的算法又可以整合更新到标注系统中,提高机器标注效率,以此形成研究与应用的闭环。(2)多来源的知识库构建。在本体库和术语集的支持下,本系统可以对多来源、多领域的文献数据实现统一语义框架下的标注,包括中医药领域和现代生物医学领域的中英文文献及中医药古代文献等,从而真正实现语义层面的数据集成与知识融合。

参考文献

- [1] BANNOUR S, AUDIBERT L, SOLDANO H. Ontology-based semantic annotation: an automatic hybrid rule-based method[C]// Proceedings of the BioNLP Shared Task 2013 Workshop, 2013.
- [2] WEI C H, ALLOT A, LEAMAN R, et al. PubTator central: automated concept annotation for biomedical full text articles[J]. Nucleic Acids Research, 2019, 47(W1):587-593.
- [3] ONG E, HE Y Q. Community-based Ontology Development, Annotation and Discussion with MediaWiki extension Ontokiwi and Ontokiwi-based Ontobedia[J]. AMIA Joint Summits on Translational Science proceedings, 2016:65-74.
- [4] ROBERTS A, GAIZAUSKAS R, HEPPLER M, et al. The CLEF Corpus: Semantic Annotation of Clinical Text[J]. Amia Annu Symp Proc, 2007:625-629.
- [5] LÓPEZ-GARCÍA P, LEPENDU P, MUSEN M, et al. Cross-domain targeted ontology subsets for annotation: the case of SNOMED CORE and RxNorm[J]. Journal of Biomedical Informatics, 2014, 47(2):105-111.
- [6] HE Y Q, 余红, 杨啸林, 等. 本体:生物医学大数据与精准医学研究的基础[J]. 生物信息学, 2018, 16(1):7-14.
- [7] 朱彦, 贾李蓉, 高博, 等. 中医临床术语系统 V2.0 设计与构建[J]. 中国中医药图书情报杂志, 2018, 42(3):10-15.
- [8] 贾李蓉, 刘静, 高博, 等. 中医临床术语系统 V2.0 病证类概念选取及关系设定[J]. 中华医学图书情报杂志, 2017, 26(12):26-29, 55.
- [9] 贾李蓉, 于彤, 崔蒙, 等. 中医药学语言系统研究进展[J]. 中国数字医学, 2014, 9(10):57-59, 62.
- [10] 赵芳芳. 面向中文电子病历的词性标注技术研究[D]. 哈尔滨:哈尔滨工业大学, 2014.
- [11] 于晓繁. 基于本体和元数据的语义标注平台模型与系统架构研究[D]. 淄博:山东理工大学, 2012.
- [12] 杨舟. 基于自然语言处理的专利文档自动语义标注方法研究[D]. 杭州:浙江大学, 2011.
- [13] 窦玉萌, 赵丹群. 协作标注系统研究综述[J]. 现代图书情报技术, 2009(2):9-17.
- [14] 廖述梅. 基于本体的语义标注原型评述[J]. 计算机工程与科学, 2006, 28(9):123-125, 128.
- [15] TSENETORP P, PYYSALO S, TOPIC 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. Avignon, 2012.
- [16] KASHYAP V. The UMLS® Semantic Network and the Semantic Web[J]. Amia Annu Symp Proc, 2003:351-355.
- [17] 贾李蓉, 李海燕, 刘静, 等. 中医药学术语系统研究概述[J]. 中国中医药图书情报杂志, 2015, 39(5):7-10.
- [18] PAN H, ZHU Y, YANG S, et al. Biomedical ontologies and their development, management, and applications in and beyond China [J]. Journal of Bio-X Research, 2019, 2(4):178-184.
- [19] TORNIAI C, BRUSH M, VASILEVSHY N, et al. Developing an Application Ontology for Biomedical Resource Annotation and Retrieval: Challenges and Lessons Learned[C]// International Conference on Biomedical Ontology. Buffalo, 2011:101-108.
- [20] YANG X L, WANG Z, PAN H J, et al. Ontology: Footstone for Strong Artificial Intelligence[J]. Chinese Medical Sciences Journal, 2019, 34(4):277-280.

(收稿日期: 2020-04-08)

(修回日期: 2020-04-30; 编辑: 魏民)