

# 面向机器学习的智慧诊疗语料库构建研究

凌 天,焦 阳,狄碧云,翁晓兰,李露芳

(浙江中医药大学图书馆,浙江 杭州 310053)

**摘要:**随着人工智能与大数据新兴理论技术发展,语料库由最初单语发展到双语语料库。语料内容由语言学扩展到文学、事实、政治、医学等建设。机器学习技术兴起降低了语料库获取足够规模语料的难度,并针对当下医疗行业医疗资源与需求不平衡问题提供有效解决方案。本文在动态分析语料库研究综述基础上,将复杂的疾病症状、真实的经临床病历、有效的治疗措施等等汇编成语料,以为社会提供智慧服务为目标,提出一种在面向机器学习的智慧诊疗语料库构建的思路。面向机器学习的智慧诊疗语料库构建过程,并探索可视化信息服务、智能语音病历、辅助治疗决策及风险预警应用场景。

**关键词:**语料库;诊疗;机器学习

**中图分类号:**TN911-34

**文献标识码:**A

**DOI:**10.3969/j.issn.1006-1959.2023.10.002

**文章编号:**1006-1959(2023)10-0006-05

## Research on the Construction of Intelligent Diagnosis and Treatment Corpus for Machine Learning

LING Tian, JIAO Yang, DI Bi-yun, WENG Xiao-lan, LI Lu-fang

(Library of Zhejiang Chinese Medical University, Hangzhou 310053, Zhejiang, China)

**Abstract:** With the development of artificial intelligence and emerging theory and technology of big data, the corpus has developed from the initial monolingual to the bilingual corpus, and the content of the corpus has been expanded from linguistics to the construction of literature, facts, politics, medicine, etc. With the development of machine learning technology, the rise reduces the difficulty of obtaining sufficient scale corpus from the corpus, and provides an effective solution to the current imbalance of medical resources and demand in the medical industry. Based on the review of dynamic analysis corpus research, this paper compiles complex disease symptoms, real clinical medical records, effective treatment measures and so on into idiom materials, aiming at providing intelligent services for the society, and puts forward an idea of constructing intelligent diagnosis and treatment corpus for machine learning; the construction process of intelligent diagnosis and treatment corpus for machine learning, and explore the application scenarios of visual information service, intelligent voice medical records, auxiliary treatment decision-making and risk early warning.

**Key words:** Corpus; Diagnosis and treatment; Machine learning

在健康中国战略背景下,效率较低的医疗体系、质量欠佳的医疗服务、看病难且贵的就医现状已成为社会关注焦点,反映医疗资源与需求不平衡的突出问题。在 2016 年中共中央政治局审议通过的《“健康中国 2030”规划纲要》中,强调全面建成统一权威、互联互通的人口健康信息平台,规范和推动‘互联网+健康医疗’服务。智慧医疗可提供更优质的医疗服务,保障人民健康。以语料库为支撑的辅助诊疗终端如医用机器人、虚拟家庭医生护理等方式提供智慧诊疗服务是当下智慧医疗发展的前沿趋势之一。而语料库是遵循特定标准采集而来的能够代表某种语言特征的数据集,可从规模化语料集中精确提取语料,挖掘出隐藏价值信息,联合定性与定

量方法研究关联数据组织成知识加以利用。随着人工智能与大数据新兴理论发展,机器学习技术得以兴起,通过学习样本数据内在关联与特征表现获得计算机语言所理解的文字、图像和声音等数据,其最终目标是让计算机像人一样具有分析学习能力,能够识别语音和图像等数据。将机器学习技术应用于语料库建设,可明显降低项目成本与工作量。因此,本研究将语料库与现代医学结合,构建基于机器学习的智慧诊疗语料库,将复杂的疾病症状、准确的临床检查、有效的治疗措施以及详实的随诊病历等汇聚成一体化的数据工程,以期让机器“学习”专家主任级医师诊疗经验,模拟诊疗时的思维逻辑,并在实际应用时给出可行性诊治方案,以智慧诊疗的方式解决医疗资源与需求不平衡问题等社会问题。

### 1 国内外语料库研究现状动态分析

**1.1 国外研究现状** 语料库起源于语言学研究,以单种语言—英语类为主。在 20 世纪 60 年代初,英语语言学家 Francis 和 Kucera<sup>[1]</sup>建立世界上首个英语文本语料库—布朗语料库。在 20 世纪 80 年代,随着科

基金项目:1.浙江中医药大学重点科研基金项目(编号:2021SZ05);

2.浙江省中医药科技计划项目(编号:2022ZA050)

作者简介:凌天(1990.7-),男,江苏东台人,硕士,馆员,主要从事智慧诊疗、中医药文化、智慧图书馆研究

学技术的不断发展,语料语言学研究领域扩展到基于平行语料库的英汉互译、文学作品和文学家语言风格甚至医学研究等。目前国外已建成且较有影响的主要有英国国家语料库 The British National Corpus(BNC)<sup>[2]</sup>与美国传统中介语料 American Heritage Intermediate Corpus(AHI)<sup>[3]</sup>,世界著名英语教学与英语字典语料库。医学研究主要有 Mollá D 等<sup>[4]</sup>提出了一个基于循证医学文本处理的语料库,该语料库是基于家庭临床杂志的临床查询部分文本信息。

1.2 国内研究现状 在 20 世纪 90 年代以来,国内专家基于语言学对语料库展开论证研究。1991 年国家语委文字应用管理司组织计算机专家对现代汉语语料库总体设计,选材原则,汉语语料库的规范和标准等关键性问题进行充分论证。2008 年刘泽权等<sup>[5]</sup>对语料库分词、标注方法进行研究,创建《红楼梦》中英双语语料库,系统全面的研究不同译本的《红楼梦》。近年来国内部分学者认识到语料库可以用于公共医疗卫生健康研究。2013 年李纲等<sup>[6]</sup>在充分回溯语料库研究的基础上,探索公共卫生突出事件动态监测系统语料库构建可行性方案。2019 年周永称等<sup>[7]</sup>自然语言标注工具 BRAT 人工处理预料,构建基于文本预料的精准医学文本语料库。2020 年刘一斌<sup>[9]</sup>在中文电子病历命名实体识别的基础上尝试引入中医命名实体,构建中医中文电子病历命名实体语料库。多个研究探索了在中医领域内更多的语料库应用场景<sup>[9-13]</sup>。2021 年林玉萍等<sup>[14,15]</sup>提出构建医学影像的多模态语料库,根据医疗检查影像实现甲状腺结节良恶性的精确分类识别。2022 年多个研究<sup>[16-18]</sup>基于中文预料将大量医学专业知识和医学术语融合,推进医学概念规范化,提高临床医学研究的效率。

纵观国内外研究,语料库起源于语言学与文学并逐渐拓展到不同学科领域研究,如公共医疗卫生健康等。构建以精准医学、影像医学、中文电子病例作为语料来源的语料库,具有一定的临床医疗效果,推动语料库在医学领域研究的发展。但仍存在不足之处,如预料采集方式单一、功能与应用场景较为稀少、采集学科领域较为局限等。基于此,本文采用机器学习技术搭建智慧诊疗语料库,将复杂的疾病症状、真实的经临床病历、安全有效的治疗措施等等汇编成语料,提出现在具备可行性的智慧诊疗应用场景,辅助医生选择最优治疗措施,降低医疗风险,同

时可降低医生工作时间成本,完善公共医疗体系,合理分配公共医疗系统资源,为患者提供优质便捷的智慧诊疗服务。

## 2 智慧诊疗语料库构建方法

2.1 需求调研阶段 需求调研是构建智慧诊疗语料库项目的前期基础。在明确实现特定类型功能的语料库前提下进行角色调研,收集整理角色用户自身需求与期望,以此为依据设计语料库搭建框架。而智慧医疗下语料库的构建最终目标就是实现辅助医生智慧诊疗的应用场景,因此在语料库构建过程中,研究者要清晰地认识到在智慧诊疗场景中的活动主体,分析医生、患者在诊疗过程中实际需求。

2.2 语料库设计阶段 智慧医疗语料库整体采用 B/S(浏览器/服务器)架构模式。在语料库设计主要包含 4 个方面:功能目标设计、技术路线设计、存储设计、数据分析与利用设计。①功能目标设计主要包含原始语料采集、数据清洗、TextDirectoryCorpus 字典调用、分词生词与标注等模块;②技术路线设计:以人工采集与基于 Python 的爬虫技术采集原始语料数据,再通过 Python、NLP(自然语言处理)技术进行语料预处理和复杂分析等;③存储设计:以 Oracle 数据库为存储模块存储语料元数据,由于各类语料库分析软件需要简单直观的可识别文本读取数据,因此同时需要再数据库所在服务器终端生成 TXT 格式的文本单元数据。语料库存储结构目录如图 1 所示;④数据分析与利用设计:通过开放数据库接口的方式拓展数据服务,利用大数据、机器学习、云计算等新兴技术实现数据分析与知识发现。

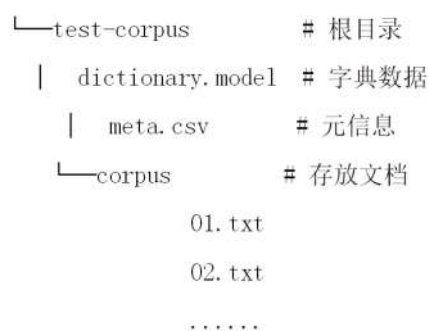


图 1 语料库存储结构目录

2.3 语料采集 原始语料是构建语料库的基础。在采集语料的过程中不仅要注重内容收集,还要收集内

容附属信息如内容来源、标题、时间等元数据信息。由于语料的规模与质量是实现智慧诊疗服务目标的前提,因此应以采集具有权威性、真实性、全面的诊疗知识为依据。智慧诊疗语料库语料可从循证医学数据库中采集,循证医学是利用现有最好的医学证据,同时结合医生临床经验和患者愿望作出医疗决策。采集方式主要有手工采集和自动采集。如图书、期刊、典藏古籍等没有数字文本化存档时,需进行人工采集,但手动采集往往工作量较大,且必须反复校对,这需要相当大的人力与时间投入。而自动采集可以以循证医学电子数据库(DynaMed、OVID EBM)等为采集对象,获取疾病临床知识等,但受限于采集字段标准多样性、网络质量、数据库源限制等影响因素,自动采集的语料会不同程度上存在字段不完整、信息缺失等情况,因此还需对所采集的语料进一步加工校对。典型自动采集语料工具有:Python、GooSeeker 等。

**2.4 语料预处理** 由于人工与自动采集的语料信息往往错综复杂,数据量庞大直接影响到语料库的分析、处理、使用。利用机器学习算法中无监督特征学习方式,通过已标注数据自编码器辨别区分无标注数据,选取合适的中英文字典实现对生语料的分词,还要使用除停用词、标注等方法才能形成可用语料信息。目前典型的文本预处理工具有:SnowNLP, OpenNLP, BosonNLP。机器学习工具有:基于 Python 的 Theano 机器学习库。

**2.5 数据库设计** 语料存储也是语料库建设的关键点,选择合理的数据库以及文本存储结构可直接影响基于语料库提供的诊疗服务的质量。选择合适的数据库结构,可有安全稳定的存储语料信息,也有效提供用户信息处理需求。本项目建设主要基于 Oracle 的数据库,关键表主要有 3 张:语料表、分词表、专用词表。而文本语料表主要存储语料的元信息包含医学名词、来源、证候、时间、方药等等;分词表主要记载所收集语料的词语信息;专用词表是根据语料库使用性质而确定,如收集的全部是疾病名称相关信息,则需要记录西医疾病名、对照的中医症状名专用词语,这有利于精准分析语料。根据以上不同阶段任务可以搭建面向机器学习的智慧诊疗语料库架构如图 2 所示。

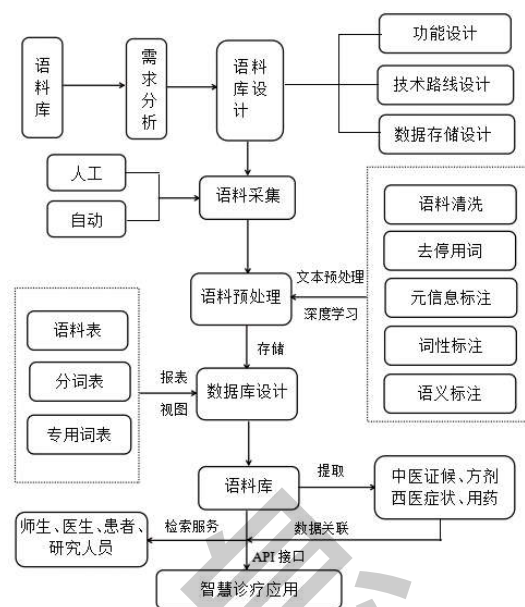


图 2 面向机器学习的智慧诊疗语料库架构

### 3 语料库建设成果及优劣势分析

**3.1 语料库建设成果** 按照以上构建方案,笔者项目团队从 2020 年 1 月起开始构建智慧诊疗语料库,截至目前已经收录中西医诊疗语料 2046 条,其中中医相关语料条数 866 条,西医相关语料条数 1180 条。分词后中文词语数 19 594 个,语句片段数约为 19.693 万条。其中中医语料条分为中医症状、中医病证名、西医疾病名、中医证候、中医医案、中医名家、所属流派、籍贯等类目。而西医语料条主要包括病因、症状、常用检查、治疗方案、常用药品及预防等 6 个类目。

#### 3.2 建设成果优劣势分析

**3.2.1 优势分析** ①智慧诊疗语料库可为实习、规培等新医生提供医学知识查询。所有知识均来自可循证的权威知识书籍、临床指南、医学数据库等,为医务人员提供实时可靠的医学知识,可根据需求学习科室总结的经典诊疗方案,满足不同科室、不同医生的个性化需求。其次这些医疗知识可整理成可共享开放数据集,提供给广大医学爱好者参考学习;②智慧诊疗语料库可作为临床医生辅助诊断的依据,基于主诉、现病史中提到的症状、疾病,以及相关检查、检验结果推荐相关的疾病、相关症状及体征,按照诊断结果由高到低匹配语料库中类似医案,推断潜在疾病可能性,根据患者基本信息、主诉、现病史等病历信息推荐合理的检查检验、用药及手术治疗等方



案,并提供对应的推断逻辑,辅助医生更好地决策;  
③智慧诊疗语料库可根据医院或者信息服务厂商的要求,可定制不同的接入方式,包括 API/BS/CS 应用程序接口,便于数据调用。

3.2.2 劣势分析 ①语料规模方面:由于本语料库致力于建设成为医务工作者医疗知识库以及临床辅助诊断参考库,并且语料库建设过程中涉及到中医、西医以及中西医结合等多维度,因时间仓促,医疗诊断语料库目前语料规模仍然较小,需继续建设;②中医语料方面:由于中医不同医家对于阴阳五行理论等理解不同,传承学术流派不同,因此对于病证的用药加减方案亦不同,需完善中医语料中同一证候的医案数量,不断搜集整理近代乃至古代我国传统名医介绍、医案、医著等,便于现代中医用药参考;③西医语料方面:医院语料主要通过 Python 等技术自动采集而来,但机器收集的资料往往良莠不齐,因此西医语料需邀医学类专业人士对其进行审核,去芜存精,提高智慧诊疗语料库的权威性。

## 4 面向机器学习的智慧诊疗语料库应用分析

4.1 可视化信息服务 可视化信息服务是智慧诊疗语料库应用最直观展示手段,传统语料库研究往往注重语料建设与语译应用,少有提供用户可视化功能展示的页面,用户通常无法直观的获取语料库包含的相关领域内知识,因此可通过一些可视化技术给语料信息搭建前端展示平台,实现包含语料信息的可视化信息服务。如本项目成果之一,近代浙派中医文献数据库,以智慧诊疗语料库中中医部分语料为基础,通过 H5 技术搭建的 Web 端、移动端一体化展示页面见图 3。从语料库中抽取的数据框架主要以浙派医学流派医家库、医著、医派为主系统整合相关信息。其中浙派医派医家库主要按浙江地域划分收录杭州、宁波、湖州、嘉兴、绍兴、金华等古今有影响力的典型人物传记 81 条,著录项包括医家名号、方剂、药物等。通过数据关联等方式为用户提供信息检索、浏览、知识图谱等多种信息服务,促进中医药文化研究,后期可以不断通过智慧诊疗语料库搭建中西医结合等多领域可视化平台。



图 3 近代浙派中医文献数据库

4.2 辅助治疗决策及风险预警 辅助治疗决策以循证医学知识库为支撑,结合医生诊疗经验,通过先进的机器算法对大规模临床诊疗数据和术后随访记录

数据进行训练,挖掘治疗方案和效果评价的隐性关联,寻找最佳治疗方案。随着语料收集技术不断升级,语料库中西医部分数据规模不断庞大,可以为辅

助治疗系统提供海量的临床指南、药典、病例、教材等医学知识库作为机器学习的数据集,提供程序(API)接口让机器“学习”专家主任级医师诊疗经验,模拟诊疗时的思维逻辑,并在实际应用时给出可行性诊治方案。打造遵循循证医学的临床辅助决策系统,从而协助医生为患者提供更精准优质的诊疗方案。这对于年轻乃至规培实习医生来说作用尤其明显,相当于把更多实战经验汇总,需要时自动调出,诊疗的过程也成了学习的过程。同时辅助治疗决策也具有风险预警的作用,如利用大样本临床诊疗数据构建风险预测模型,结合患者自身病情和特征,适时动态的给出规避风险的治疗方案,如在情况较为紧急时的急诊考虑手术治疗、术后并发症以及用药副反应等,及早预测患者不良反应并予以作出预防措施。

**4.3 智能语音病历** 语音识别技术将通过识别人类的语音中各种特征并转化成计算机可识别的二进制输入语言,是一项成熟稳定的声音特征提取技术,但医学往往存在诸多复杂晦涩的医学专业词汇,造成计算机识别程度低,也就很难帮助医生快速录入病历。而智能语音病历主要利用诊疗语料库汇聚海量的医学分词,搭建流式端到端语音语言一体化建模算法将语音快速准确识别为文字,支持智能手机系统语音交互、机器人语音沟通、多场景语音内容分析等。智能语音病历技术较为基础,但能帮医生减负不少工作量。据深圳市德信数据调查显示,我国50%以上的住院医生平均每天有4 h以上在写病历,而应用语音病历后,患者的主诉内容可以实时地转换成文字,效率明显提升,减少医生在诊疗过程中不必要的时间成本。语音识别技术是实现智慧诊疗的有效探索。

## 5 总结

随着数据挖掘、机器学习等技术的不断进步,语料库目前建设所面临的语料体量大、语料收集、语料分断困难等难点将得到有效解决。本研究在现有信息技术条件下,尽可能采集语料并进行规范化处理,设计数据库,合理存储预料表等数据,形成诊疗语料库,包含复杂的疾病症状、临床检查知识、随诊病历与治疗方案等。智慧诊疗语料库作为智能智能诊疗系统研究的基础与参考,未来对此进一步深入研究,可以帮助医生应用临床技能和经验迅速判断患者状况及疾

病诊断,选择最优治疗措施,降低医疗风险及医生工作时间成本,完善公共医疗体系,合理分配公共医疗系统资源,为患者提供优质便捷的智慧诊疗服务。

## 参考文献:

- [1]王天奇,管新潮.语料库语言学研究的拓展——《Python文本分析:用可实现的方法挖掘数据价值》评介[J].外语电化教学,2017(5):93-96.
- [2]张碧琼.英国国家语料库 BNC 在英语词汇教学中的应用[J].新丝路,2020(14):226-227.
- [3]Dictionaries EOTAH.The American Heritage Dictionary of the English Language[M].Houghton Mifflin,2000-09-14.
- [4]Mollá D,Santiago-Martínez ME,Sarker A,et al.A Corpus for research in text processing for evidence based medicine[J].Language Resources and Evaluation,2016,50(4):705-727.
- [5]刘泽权,田璐,刘超朋.《红楼梦》中英文平行语料库的创建[J].当代语言学,2008,10(4):329-339,379-380.
- [6]李纲,陈环浩,毛进.突发公共卫生事件网络语料库系统构建[J].情报学报,2013,32(9):936-944.
- [7]周永称,范少萍,晏归来,等.精准医学文本语料库构建研究[J].医学信息杂志,2019,40(12):41-47.
- [8]刘一斌.中医中文电子病历命名实体语料库构建及研究[D].广州:广州中医药大学,2020.
- [9]肖晓霞.基于机器学习的中医临床症状数据元研究[D].长沙:湖南中医药大学,2018.
- [10]游正洋,王亚强,舒红平.基于词性标注的中医症候名语料库[J].电子技术与软件工程,2017(21):177-178.
- [11]刘成,董益敏,王小芳.基于语料库的中医脉诊术语英译规范探讨[J].中华中医药杂志,2019(11):5064-5068.
- [12]王采薇.基于语料库方法的《千金方》中医理论创新研究[D].咸阳:陕西中医药大学,2019.
- [13]游正洋,王亚强,舒红平.基于词性标注的中医症候名语料库[J].电子技术与软件工程,2017(21):177-178.
- [14]林玉萍,龙红,李彪,等.基于医学影像和病历文本的甲状腺多模态语料库构建与应用[J].西北大学学报(自然科学版),2021,51(2):198-206.
- [15]林玉萍,郑尧月,郑好洁,等.基于医学影像分割方法的多模态语料库构建[J].模式识别与人工智能,2021,34(4):353-360.
- [16]程瑶.中文医学学术语资源对真实世界医学文档语料的覆盖度调研[D].北京:北京协和医学院,2020.
- [17]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.
- [18]易晓宇,易绵竹.基于中文语料的医学概念规范化研究[J].河南科技学院学报(自然科学版),2022,50(2):70-76.

收稿日期:2022-07-02;修回日期:2022-07-22

编辑/肖婷婷