

人工智能在医学诊断知识图谱构建中的应用研究^{*}

聂莉莉 李传富 许晓倩 朱川川 徐志鹏 武红利

(安徽中医药大学 (安徽中医药大学第一附属医院 (安徽中医药大学
合肥 230000) 合肥 230000) 合肥 230000)

〔摘要〕 介绍人工智能在医学领域的应用,通过自然语言处理方法自动构建基于“疾病-症候-特征”3层结构模型的医学诊断知识图谱,具体阐述智能构建方法与构建过程,将其运用到临床决策支持系统中检测临床疗效,结果表明通过该方法构建的医学知识图谱具有效率高、疾病诊断正确率高等特点。

〔关键词〕 人工智能;知识图谱;自然语言处理;正则表达式;隐马尔可夫模型

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2018.06.002

Study on Application of Artificial Intelligence in the Building of Medical Diagnosis Knowledge Graph NIE Li-li, Anhui University of Traditional Chinese Medicine, Hefei 230000, China; LI Chuan-fu, The First Affiliated Hospital of Anhui University of Traditional Chinese Medicine, Hefei 230000, China; XU Xiao-qian, ZHU Chuan-chuan, XU Zhi-peng, WU Hong-li, Anhui University of Traditional Chinese Medicine, Hefei 230000, China

〔Abstract〕 The paper introduces the application of Artificial Intelligence (AI) in medical domain, the medical diagnosis knowledge graph is built based on the 3-layer model of "disease-symptom-feature" through natural language processing, and it dilates upon the intelligent building method and its application in the Clinical Decision Supporting System (CDSS) for the detection of clinical efficacy, the results demonstrate that the medical diagnosis knowledge graph built with the method features high efficiency and high accuracy of disease diagnosis.

〔Keywords〕 Artificial Intelligence (AI); Knowledge graph; Natural language processing; Regular expression; Hidden Markov Model

1 人工智能在医学领域的应用

1.1 人工智能应用到医学领域是大势所趋

近年来医学领域的数字化进程不断推进,逐渐迈向智慧医疗阶段。人工智能(Artificial Intelligence, AI)作为智慧医疗的核心,在构建系统化、精准化医疗服务中起到关键作用^[1]。当前人工智能已广泛应用于医疗健康领域,如在医学影像的识别方面已经达到顶级医生的水平,甚至在某些疾病的诊断过程中智能决策支持系统对疾病的诊断正确率超过一般医生的水平^[2]。人工智能应用于医学领域

〔修回日期〕 2018-06-12

〔作者简介〕 聂莉莉,硕士研究生;通讯作者:李传富,教授,主任医师,发表论文40余篇。

〔基金项目〕 国家自然科学基金资助项目“面向全科医学诊断知识图谱的自动化构建方法研究”(项目编号:6187020839)。

将会推动传统医学向智能医学转型，为传统医学带来巨大变革，使其智能化^[3]。医学人工智能的发展可以更快、更准确地解决临床诊断问题，将人工智能应用到医疗领域是大势所趋^[4]。而医学知识是医学人工智能的核心内容，构建医学知识图谱是研究医学人工智能的首要任务。

1.2 人工智能在医学知识图谱中的应用

在国外已有较多重要的知识图谱研究成果，如 Google Knowledge Graph^[5]、DBpedia^[6]、YAGO^[7] 和 Freebase^[8] 等。而我国知识图谱的研究还处在初级阶段，主要有搜狗“知立方”、百度“知心”、上海交通大学的 Zhishi.me^[9] 以及复旦大学 GDM 实验室的中文知识图谱项目^[10] 等。在医学领域主要有中科院计算机研究中心研发的医学知识库^[11] (NKIMed) 和中国中医科学院中医药信息研究所构建的中医药知识图谱^[12]，但都还处在探索阶段。首先由于中文的语言特点，不能利用语义分析中处理英文文本的方式直接处理中文文本，需要对其方法进行改进^[13]；其次，自然语言处理技术的不成熟导致市场上的医学诊断知识图谱大多通过手动构建，工作量巨大，而且医学知识在不断地更新变化，难以满足人工智能研究的实际需求^[14]。知识的自动提取是目前业界的重点研究方向，也是将来知识图谱的发展趋势^[15-16]。医学知识图谱在构建过程中存在很多问题，如获取知识速度慢，多种疾病同时存在会造成噪音问题等，将人工神经网络应用于医学专家系统可以有效解决在构建过程中的噪音问题，提高知识获取的速率以及对知识的推理能力，从而提高对

疾病诊断的正确率。

2 医学诊断知识图谱的 3 层结构模型概念

从医学角度看，疾病的发生伴随着很多症候，相同的症候也可出现在不同的疾病中。支气管扩张症“疾病-症候”两层结构，见图 1。该疾病有咳嗽、咯血、咳痰等症状，将这种“疾病-症候”模型称为两层结构模型。仅根据咳嗽、咳痰和咯血症状不能对支气管扩张症进行详细表述。经过一系列相关研究，参考国内外构建知识图谱的成功经验，首次提出基于“疾病-症候-特征”3 层结构的常见疾病诊断知识图谱模型，通过专家人工构建该模型的呼吸系统疾病诊断知识图谱，在临床决策支持系统 (Clinical Decision Support System, CDSS) 方面取得良好效果^[19]。支气管扩张症“疾病-症候-特征”3 层结构，见图 2。该疾病表现为慢性咳嗽、咳大量脓痰、反复咯血，将这种“疾病-症候-特征”称为 3 层结构模型。3 层结构模型参考 Collins 疾病诊断决策树理论^[20] 构建，以常见疾病为基础建立两个有内在联系的知识库：疾病知识库和症候知识库。疾病知识库存储疾病信息以及该疾病相关的诸多症候和特征；症候知识库存储症候知识以及特定症候相关的各种具体特征选项，将两者融合形成“疾病-症候-特征”3 层结构模型的医学诊断知识图谱，是对“疾病-症候”两层结构模型的进一步提升，能够更加准确地表达疾病相关的各类结构化信息。3 层结构模型知识图谱的提出能够将知识表示得更具体、准确，提高对疾病诊断的准确性。

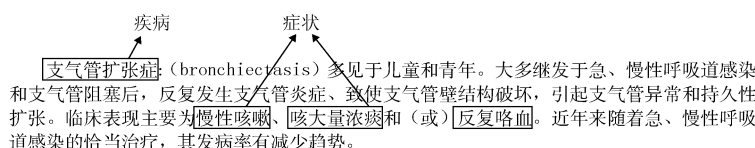


图 1 支气管扩张症“疾病-症候”两层结构

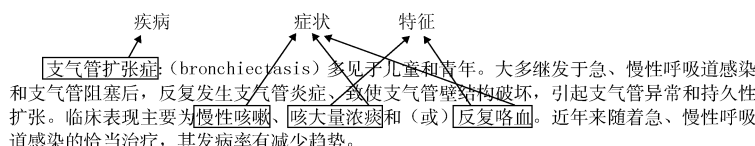


图 2 支气管扩张症“疾病-症候-特征”3 层结构

3 智能构建方法

知识图谱通常有两种构建方式：自顶向下和自底向上^[21]。自顶向下的方法是先构建本体，然后将抽取到的实体匹配到所构建的顶层本体中；自底向上的方法是直接从抽取到的数据中提取实体间的关系并更新到知识图谱中。本研究采用自底向上的方法，通过自然语言处理（Natural Language Processing, NLP）方法自动构建医学诊断知识图谱，具体构建流程，见图 3。以医学教科书、文献和专著作为知识来源；通过正则表达式^[22]获取含有实体对的句子；利用隐马尔可夫模型^[23]对获取到的句子进行分词，抽取症候、特征选项等实体以及重要性、特异性和频度等属性；通过构建疾病知识词典、症候知识词典和医学分词词典为后续句法分析做准备；通过依存句法分析^[24]获取实体与属性之间以及实体间的关系（3 元组），将构建好的疾病诊断知识图谱应用到 CDSS 中，用临床病例来训练数据，不断丰富和完善医学诊断知识图谱并优化 CDSS 算法和 NLP 算法。

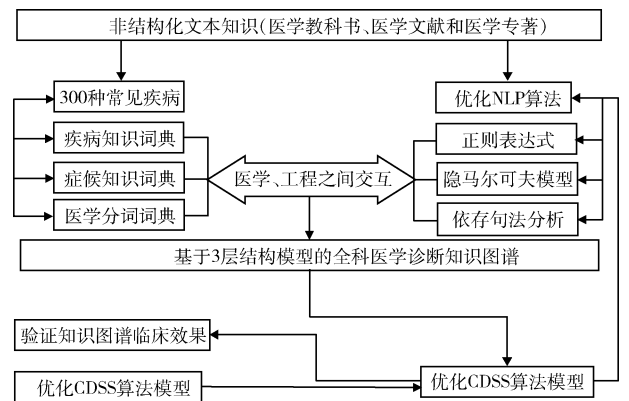


图 3 医学诊断知识图谱构建流程

4 智能构建过程

4.1 创建症候知识词典

医学专家通过对医学教科书、文献和专著进行研究，构建症候知识词典。根据特征属性将症候知识词典分为公有和私有词典，公有词典存储公有“症候 - 特征”属性及其属性值，如急性发病可用

于描述咳嗽的发病情况，也可描述咯血的发病情况；私有词典用来存储症状特有的“症候 - 特征”属性及属性值，如稽留热仅可描述发热的类型。该症候知识词典用于训练模型。

4.2 基于正则表达式进行分句及基于隐马尔可夫模型进行分词

搜集大量具有价值的医学文献和书籍并录入计算机中，利用正则表达式方法与已构建的症候知识词典进行匹配，从非结构化的医学文本中获取含有症候和特征选项的句子。采用 HanLP 分词器，通过隐马尔可夫模型算法实现分词，先进行原子切分，再进行初次分词，找出原子间所有可能的组合，通过遍历循环访问中文分词词典组成有意义的词组。但由于医学文本的复杂性和医学术语的多样性，基于中文分词词典的方法不能直接用于医学文本分词。因此在中文分词词典的基础上构建疾病知识词典、同义词词典、医学分词词典等，这些词典可通过不断手动添加逐步完善。

4.3 通过语义分析构建 3 层知识结构模型

通过依存句法分析抽取肺结核疾病、症候和特征之间的关系，见图 4。通过分词得到的结果是“肺结核/常/伴随/低热/咯血”，该句中共有 3 个实体，分别为肺结核、低热和咯血。在依存句法分析中核心谓语动词是支配其他成分的中心成分，而其本身不受其他成分的支配，所有受支配的成分都以某种特殊的依存关系从属于支配者。通过标记谓语动词来分析句中其他成分，如句中的肺结核是主语，低热和咯血均为宾语，伴随则是核心谓语动词（HED），肺结核和伴随之间形成主谓关系（SBV），伴随和低热、咯血形成动宾关系（VOB），低热为宾语，咯血也是宾语，两者是并列关系从而形成定中关系（ATT）。

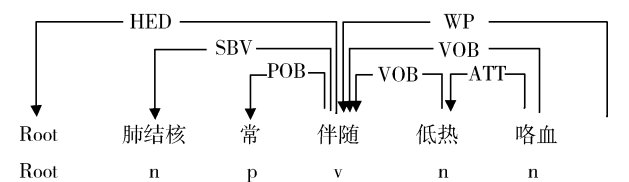


图 4 依存句法分析及词性标注

4.4 医学诊断知识图谱的自动构建

通过正则表达式获取含有实体对的句子，利用隐马尔可夫模型抽取实体，进行语义依存分析获取实体间的关系从而构建医学疾病诊断知识图谱。疾病部分知识图谱可视化，见图 5，可以看出支气管扩张症和肺结核都伴有咳嗽，支气管扩张

症表现为慢性咳嗽且反复发生，而肺结核则表现为干咳且程度较轻，通过咳嗽的发病情况、性质和严重程度来鉴别这两种疾病。在构建的疾病诊断知识图谱中有 3 个重要的特性词：频度、重要性和特异性，肺结核和支气管扩张症分别具有各自特有的症状和特征选项，但通过相同的症状连接形成医学知识图谱。

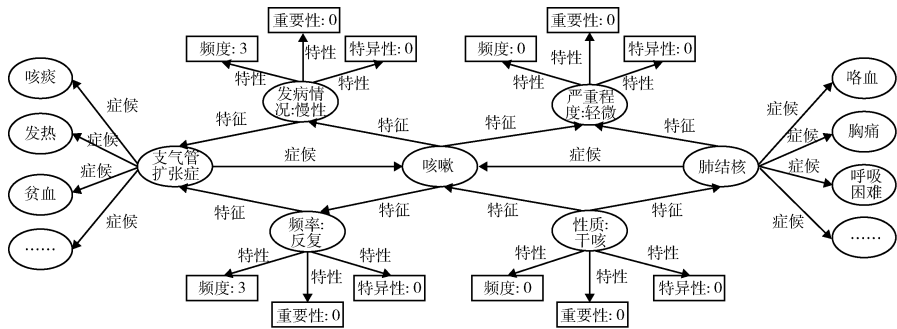


图 5 疾病部分知识图谱可视化

5 在临床决策支持系统中的应用

5.1 概述

CDSS 是通过计算机技术与某些特定的医学知识结合，对医学问题进行分析 and 解决的计算机应用系统。CDSS 通过人机交互的方式帮助医生进行临床决策，提高医生的诊疗水平。通过将医学诊断知识图谱应用到 CDSS 中用以检测其临床疗效并与前期构建的疾病诊断知识图谱进行比较。

5.2 有效性测试

5.2.1 方式 采用离线测试方法即选择临床病例进行诊疗测试，挑选 50 例临床病例，涵盖 10 种常见的呼吸系统疾病，临床病例分布，见表 1。

表 1 临床病例分布

疾病名称	病例数
急性上呼吸道感染	4
肺结核	6
肺脓肿	9
肺血栓栓塞症	4
支气管扩张症	8
支气管哮喘	4

续表 1

胸腔积液	1
原发性支气管肺癌	2
气胸	8
肺炎链球菌性肺炎	4
合计	50

5.2.2 标准 医学专家根据临床经验对测试结果确立一些衡量标准，方便对测试结果进行分析。根据推论概率的分布将系统结果分为 5 个类型：确定型、推荐型、建议型、可能型和误诊型，见表 2。

表 2 测试结果类型分布

类型	概率分布
确定型	0.8 ~ 1
推荐型	0.6 ~ 0.8
建议型	0.4 ~ 0.6
可能型	0.1 ~ 0.4
误诊型	< 0.1

5.2.3 过程 首先进行区域 1 部分的操作，输入患者姓名、年龄和性别等基本信息；然后在区域 2 输入患者主诉，系统内 65 部推理机根据输入的主诉进行推理并将可能要问的症候反馈给区域 3，再依次点击症候和特征选项，区域 4 给出推论结果并

通过条形图直观地显示在区域 5，观察系统推论结果并记录。测试过程，见图 6。



图 6 测试过程

5.2.4 结果 临床疗效评估结果，见表 3。误诊

表 3 临床疗效评估结果

项目	确定型	推荐型	建议型	可能型	误诊型	合计
衡量标准	0.8 ~ 1	0.6 ~ 0.8	0.4 ~ 0.6	0.1 ~ 0.4	< 0.1	-
病例数	6	13	23	7	1	50
所占百分比 (%)	12	26	46	14	2	100

5.3 与手动构建组模型的对比及结果

将前期手动构建的呼吸系统医学诊断知识图谱应用于 CDSS 中，将上述 50 例临床病例用以测试其效果，其测试过程和自动构建组测试过程相同，根据多症候贝叶斯算法^[25]，计算出某些症候特征下的可能疾病和其相应的概率并记录整个测试结果。两组结果对比，见表 4。可计算出 $\chi^2 = 1.325 < \chi^2_{0.05,4}$

表 4 基于手动构建模型和自动构建模型的结果对比

项目	确定型	推荐型	建议型	可能型	误诊型	总数
衡量标准	0.8 ~ 1	0.6 ~ 0.8	0.4 ~ 0.6	0.1 ~ 0.4	< 0.1	-
自动构建组	6	13	23	7	1	50
手动构建组	4	11	22	10	2	50
合计	10	24	45	17	3	100

6 结语

本研究通过 NLP 方法自动构建医学诊断知识图谱，在传统的两层模型（疾病 - 症候）基础增加特征属性，形成 3 层结构模型（疾病 - 症候 - 特征），

型的推论结果与临床诊断结果不相符，不能对医生的诊断起到辅助作用，则：

误诊率 (%) = (总诊数 / 总数) × 100%

由表 3 可计算出误诊率 = 1/50 × 100% = 2%，推算出正确率为 98%。可以看出系统的测试结果主要为推荐型和建议型，符合临床实际情况。系统通过与患者的交互过程及其主诉，将症状和特征选项输入到系统中，系统通过内部推理机制推导出结果并将结果反馈给医生，对医生的诊断决策过程给出建议和意见。医生根据推论结果让患者做相应的辅助检查从而对疾病进行确诊。

= 9.49, $p = 0.85712 > 0.05$ ，智能化构建的医学诊断知识图谱具有临床使用价值。医学诊断知识图谱涉及数千种疾病，若单纯依靠手动构建则工作量巨大，也不能及时更新医学知识，利用 NLP 方法构建的医学诊断知识图谱可以对大量数据进行挖掘，自动获取有效信息，构建的医学知识图谱全面、完善，可以及时更新医学知识，提升医生对疾病诊断的正确率。

将疾病知识描述更详尽。通过正则表达式对文本进行分句，利用隐马尔科夫模型对其进行分词，通过依存句法分析来构建 3 层结构知识模型。将人工智能应用于医学知识图谱的自动化构建中能有效解决构建医学知识图谱过程中存在的效率低、耗时长以及需要大量人工参与等问题。目前在 3 层结构模型

中只涉及症状和特征选项,未涉及辅助检查。在下一步的工作中将致力于这些方面的研究,构建更加全面、完备、准确的医学诊断知识图谱,解决人工智能在医学领域中的瓶颈问题。

参考文献

- 1 刘鹏宇. 人工智能:撬动智慧医疗 [J]. 互联网经济, 2017, 4 (3): 12 - 15.
- 2 袁冰, 范钢. 中医学如何走进人工智能时代 [J]. 中华中医药杂志, 2018, 33 (2): 698 - 703.
- 3 王锡山. 未来医学时代——人工智能诊疗 [J]. 中华结直肠疾病电子杂志, 2017, 6 (4): 349 - 352.
- 4 周路茜. 医学人工智能开始加速产业化 [J]. 新经济导刊, 2017, (7): 18 - 22.
- 5 Singhal A. Official Google Blog: introducing the knowledge graph: things, not strings [J]. Official Google Blog, 2012, (5): 1 - 8.
- 6 Paulheim H. Data - Driven Joint Debugging of the DBpedia Mappings and Ontology [C]. Cham: European Semantic Web Conference, 2017: 404 - 418
- 7 Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge [C]. Homburg, International Conference on World Wide Web, 2007: 697 - 706.
- 8 Bollacker K, Evans C, Paritosh P. Freebase: a collaboratively created graph database for structuring human knowledge [EB/OL]. [2018 - 05 - 10]. <https://doi.org/10.1145/1376616.1376746>, 2008. 06. 09.
- 9 刘知远, 崔安欣. 大数据智能: 互联网时代的机器学习和自然语言处理技术 [M]. 北京: 电子工业出版社, 2016.
- 10 程学旗, 靳小龙, 王元卓. 大数据系统和分析技术综述 [J]. 软件学报, 2014, 25 (9): 1889 - 1908.
- 11 周肖彬. 医学本体和医学知识获取的研究 [D]. 北京: 中国科学院计算技术研究所, 2003.
- 12 贾李蓉, 刘静, 于彤. 中医药知识图谱构建 [J]. 医学信息学杂志, 2015, 36 (8): 51 - 53.
- 13 鄂世嘉, 林培裕, 向阳. 自动化构建的中文知识图谱系统 [J]. 计算机应用, 2016, 36 (4): 992 - 996.
- 14 Shi L, Li S, Yang X. Semantic Health Knowledge Graph: semantic integration of heterogeneous medical knowledge and services [J]. BioMed Research International, 2017, (4): 1 - 12.
- 15 Ford E, Carroll J A, Smith H E. Extracting information from the text of electronic medical records to improve case detection: a systematic review [J]. J Am Med Inform Assoc, 2016, 23 (5): 180.
- 16 Rotmensch M, Halpern Y, Tlimat A. Learning a Health Knowledge Graph from Electronic Medical Records [J]. Scientific Reports [EB/OL]. [2017 - 07 - 20]. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5519723/>
- 17 沈乐天, 杜向慧. 人工智能在恶性肿瘤放疗领域中的应用与前景 [J]. 浙江医学, 2018, 40 (8): 783 - 795.
- 18 Fedorov A, Beichel R, KalpathyCramer J. 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network [J]. Magnetic Resonance Imaging, 2012, 30 (9): 1323 - 1341.
- 19 Jiang Y, Qiu B, Xu C. The Research of Clinical Decision Support System Based on Three - Layer Knowledge Base Model [J]. Journal of Healthcare Engineering, 2017, 2017 (2 - 3): 1 - 8.
- 20 Collins, Douglas R. Algorithmic Diagnosis of Symptoms and Signs: a cost - effective approach [J]. 2017.
- 21 刘峤, 李杨, 段宏. 知识图谱构建技术综述 [J]. 计算机研究与发展, 2016, 53 (3): 582 - 600.
- 22 罗粮, 朱儒明. 基于正则表达式的 Web 页面信息抽取技术研究 [J]. 现代计算机, 2017, 5 (15): 17 - 19.
- 23 王庆福. 隐马尔可夫模型在中文文本分词中应用研究 [J]. 无线互联科技, 2016, 7 (13): 106 - 107.
- 24 张丹. 语言学知识在依存句法分析中的应用研究 [D]. 沈阳: 沈阳航空航天大学, 2017.
- 25 沈宏伟, 邵堃, 张阳洋. 基于朴素贝叶斯的信任决策模型 [J]. 小型微型计算机系统, 2018, 39 (2): 275 - 279.