基于知识图谱的医疗问答系统研究应用

黄承宁[1] 殷晓磊[2] 陈 武[3]
HUANG Chengning YIN Xiaolei CHEN Wu

摘 要 为了改善当前中文医疗领域高质量专业问答数据缺乏的问题，提出一种面向患者就医前自诊的基于医疗知识图谱的自动问诊系统。采用一种基于卷积神经网络（CNN）算法，解析用户以自然语言提出的问题，使用 Neo4J 进行面向医疗健康领域的知识图谱模型的构建存储，再使用基于卷积神经网络的知识库关系 / 属性映射算法进行语义解析，最后通过关联搜索算法完成用户问题关键字与知识图谱之间的匹配，得出患者问题的答案。实验证明，所设计构建的基于知识图谱的智能问答系统有效解决了相关病症咨询专业缺乏问题，有效实现了患者自助，具有较高的准确率和实用价值。

关键词 知识图谱；问答系统；深度学习；语义推理；智能医疗

doi：10.3969/j.issn.1672-9528.2024.05.048

0 引言

自动问答系统（question answering，QA）是一种将处理自然语言（natural language，NL）重新组织后对指定问题给出答案的智能程序[1]。以 2019 年第三季度的《中国网民科普需求搜索行为报告》的报告行文调查数据为例，中国上网人员的自我科普人数已经高达至 11 亿人，其中健康与医疗在科普主体搜索中环比增长 3.3%[2]。引入医学知识图谱概念成为通过人工智能技术处理生物医疗类数据方向的一个重要节点。知识图谱通常是对某一特定领域的知识进行抽取，经过处理后构成可以对该领域知识进行描述的概念、关系，将领域中不同实体之间的知识脉络连接构建成一幅图谱[3]。由于知识谱图在数据的组织和关联分析方面表现优异，可以反映出实体及其之间复杂的联系，因此在医疗领域，经常将互联网中海量的医疗健康知识通过关联点构建一幅医疗知识网络，即通过效果分析知识点与关系模型，构建准确而丰富的医疗知识图谱，并在此基础上进一步搭建设计实现医疗智能问答系统。

近年来，各大互联网企业加大了知识图谱构建工作的投入，知识图谱的构建得到了大力的推进。此类系统通常利用知识图谱（knowledge graph，KG）解析用户提出的问题[4]，经过对知识库的搜索后给出答案。知识图谱是一种表达实体之间关系的语义网络，它以实体作为节点、实体间关系作为边之间的属性，在自动问答领域发挥着重要作用，被越来越多地使用。

1 知识图谱关键技术相关理论

基于神经网络的深度学习模型有利于统计文本数据中语言的出现概率，更好地对大体量的文本数据进行建模。本文首先研究了神经网络的基本结构、基于深度神经概率语言模型以及基于语义相似度的向量空间模型。

1.1 知识图谱问答的定义

（1）知识图谱：一种基于图的数据结构，以符号的形式来描述现实中的知识信息及其之间的关系。它是结构化的语义知识库，用于以符号形式描述物理世界中的概念及其相互关系，其基本组成单位是“实体—关系—实体”三元组，以及实体及其相关属性—值对，实体间通过关系相互联结，构成网状的知识结构[5]。

（2）知识图谱问答：此模型可以解释为存在一个知识库网络，根据自然语言的问题，从图谱库中搜索相关知识答案给出反馈[6]。如图 1 所示，提出问题“巴拉克·奥巴马（Barack Obama）出生在哪里”的知识图谱，模型通过解析问题和推理等过程得出答案为“Honolulu”。文章把这种反馈提问建立在依赖知识图谱之上的问答定义为知识图谱问答。

1. 南京工业大学浦江学院 江苏南京 211222
2. 南京审计大学 江苏南京 211815
3. 中国建筑股份有限公司 北京 100029
[基金项目] 国家自然科学基金项目 (61702229)；江苏省高等学校自然科学研究项目（18KJD520001）；南京工业大学浦江学院科研重点培育课题（njpj2022-1-07）；南京工业大学浦江学院青年教师发展基金（PJYQ03）

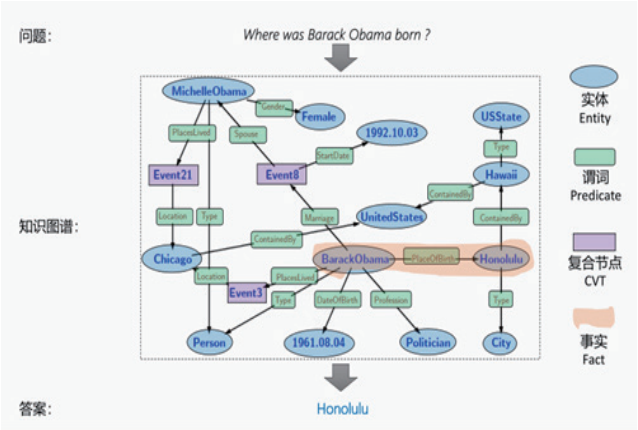

图 1 结构化知识图谱的问答实例

1.2 知识图谱构建概述

知识图谱的构建有一定的流程与步骤，一般而言首先要进行信息抽取（特征提取与建模）、相关网络知识融合、对知识库的处理与评价，再根据相关领域需求与经验对之前知识库进行优化调整。知识图谱的整体架构图如图 2 所示，图 2 中虚线框出的是知识图谱构建的过程，也是知识网络库建立和更新的相关流程描述，图中虚线内则是构建图谱的主要流程 [7]。

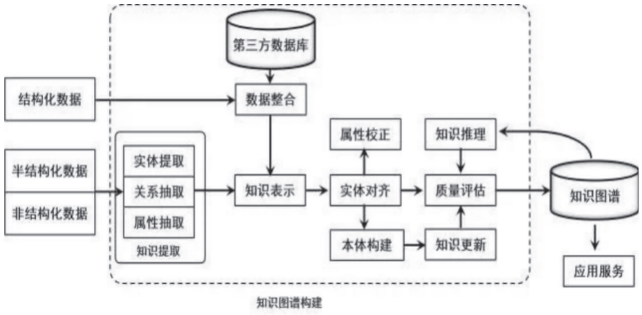

图 2 知识图谱的整体架构

通过使用半自动或自动方法，从原始数据提取实体关系并存储在基于知识的数据层中。为此，构建知识谱图一般需要两步：一是根据知识图谱领域，选择获取相关的知识信息，初步得到一个相关知识信息数据库；二是由于获取的信息数据格式不一，类型丰富，需要进一步对原始知识数据进行抽取与归纳处理 [8]。其中，知识融合和知识处理研究暂时不在本文进行赘述，此处本文针对信息抽取和知识存储进行相关工作的梳理。

1.3 知识抽取

知识提取是从不同的来源、类型、数据结构的数据源提取知识实体、关系和属性以形成知识源并存储在知识图中的技术 [9]。知识提取的任务包括实体提取、关系提取和属性提取。

实体抽取又称为命名实体识别（name entity recognition，NER）[10]，作用是从由自然语言构成的文本中抽取对象信息。关系抽取是为了得到对象语义信息，再从相关语料中提取出实体对象之间的关联矩阵，通过关系网络将实体建立联系，从而进一步形成知识网状结构。通过从数据源中提取实体属性和其值之间的关联，完成对属性名和值的定位。

1.4 知识存储

目前，应用较为广泛的知识存储方案分为单一式知识存储和混合式知识存储 [11]。为了方便知识图谱的快速搭建及对知识库检索的效率，本文采用混合式存储的方式。对应图数据库，其存储模型一般分为资源描述框架（resource description frame，RDF）三元组 [12]、分布式属性图 [13] 和超图三种。

本文使用 Neo4J[14] 对基于图的知识库进行存储。相较于传统的关系型数据库，图数据库在抽取信息和对网络关系的关联查询能力上有着天然的优势。尤其是当业务需要进行多维度的查询时，基于图的知识库的查询效率有时会高出关系型数据库的数万倍。另外，图数据库的设计也很灵活，无需进行大幅度的修改。针对大规模数据的存储，图数据库相较于关系型数据库更加受推荐。

1.5 知识图谱推理问答的方法

1.5.1 键值记忆网络

记忆网络（KV-MemNN）[15] 模型在处理各类文档及问题回答任务时有着很高的优势。该模型寻址基于键存储器高速缓存（查找表），读取基于值存储器高速缓存，不但为行业从业人员带来了可为以前知识更加灵活编码的模式，还使得模型对键值变化的表达力更强。在设计键值时，键负责匹配问题，而值负责答案响应。该模型有着整个模型可以通过键值变换进行训练，同时仍然通过随机梯度下降使用标准反向传播的特性 [16]。KV-MemNN 模型如图 3 所示。

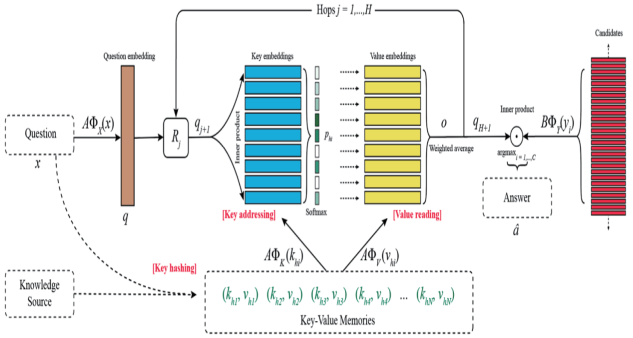

图 3 KV-MemNN 框架图

KV-MemNN 把各个存储槽定义为 (k_i, v_i) 向量对，以此来表示问题 x。key-value 的内存寻址和读取的过程包括以下三个步骤。

key-hashing：问题查询在使用键值对在一个非常大的文

件中找到自己的内存阵列。大于内存插槽大小的 KV 对可以随机删除到内存插槽大小，而小于内存插槽大小的 KV 对则满足。

key-addressing: 每一个候选记忆都被计算出一个概率：

$$p_{h_i} = \text{Softmax}\left(A\Phi_X(x)\cdot A\Phi_K\left(k_{h_i}\right)\right) \quad (1)$$

式中：$A\phi_X(x)$代表问题的表示，$A\phi_K\left(k_{hi}\right)$代表每个内存插槽的表示，两者点乘然后 softmax 得到每个内存插槽的概率。

value reading：最后，值被乘以它们的概率以获得输出：

$$o = \sum p_{h_i} A\Phi_V\left(v_{h_i}\right) \quad (2)$$

对于多层的 key-value memory network，得到输出的 O 以后，将循环：

$$q_2 = R_1(q+o) \quad (3)$$

根据 o 刷新 q_2, 反复重复 Value Reading 和 Key Addressing 两个操作，以实现内存访问过程，最后在第 H 跳后得到 q_{H+1}。算出全部可能为结果的参数：

$$\hat{a} = \text{argmax}_{i=1,\ldots,C}\,\text{Softmax}\left(q_{H+1}^{\cdot} B\Phi_Y\left(y_i\right)\right) \quad (4)$$

键值存储网络是由端到端存储网络改造而来的，其一方面是为了解决 KB 的结构不够灵活的问题，以构建一个可以直接读取文本的记忆网；另一方面解决端到端存储网络无法加入大规模已有知识的问题。

1.5.2 基于强化学习的多跳路径搜索

多跳路径搜索也称为多跳推理（Multi-Hop）[17]，是在不完全知识图谱（knowledge graph，KG）上解析回答（question answer，QA）的一种有效方法。这个问题可以由强化学习（reinforcement learning，RL）进行规划。本文选择以 Variational Reasoning Network（VRN）[18] 模型作为代表进行简单说明，模型框架如图 4 所示。

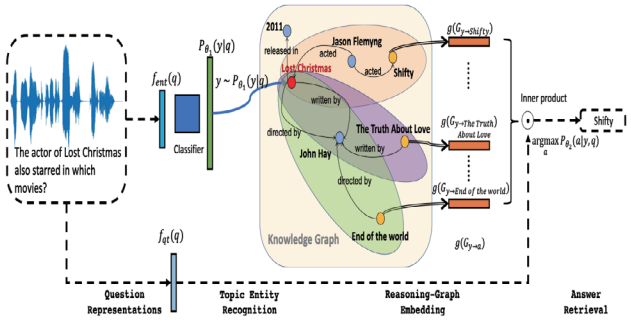

图 4 Variational Reasoning Network（VRN）框架图

该模型分为两个模块：主题实体识别模块$P_{\theta_1}(y|\ q)$和知识推理模块$P_{\theta_2}(a|\ y,q)$，将主体实体 y 看作隐变量。两个模块进行联合优化可得出以下结论：

$$\max_{\theta_1,\theta_2}\frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{y\in V(G)}P_{\theta_1}\left(y|\ q_i\right)P_{\theta_2}\left(a_i|\ y,q_i\right)\right) \quad (5)$$

主题识别模块：

$$P_{\theta_1}(y|\ q) = \text{softmax}\left(W_y\cdot f_{\text{ent}}(q)\right) = \frac{\exp\left(W_y\cdot f_{\text{ent}}(q)\right)}{\sum_{y'\in V(G)}\exp\left(W_{y'}\cdot f_{\text{ent}}(q)\right)} \quad (6)$$

知识推理模块：从边缘出发，找出所有可能在 T 跳之内到达实体（忽略边的方向）及对应的整个子图，用 G_y 表示，具体知识推理流程示意图如图 5 所示。

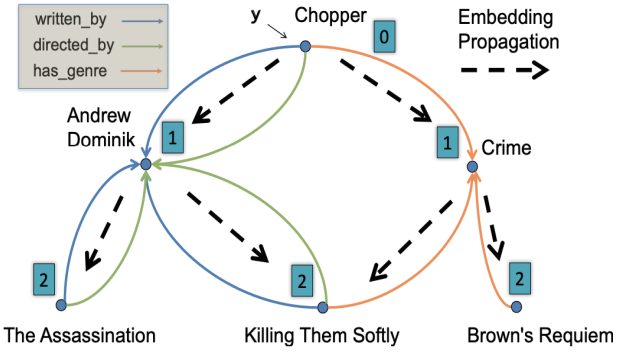

图 5 知识推理示意图

G_y 中包含的实体都是潜在的答案实体。对任意潜在答案 a，定义 $G_{y\to a}$ 为包含了所有 y 到 a 路径的子图。计算子图的向量表示如下：

$$g\left(G_{y\to a}\right) = \frac{1}{\#\ \text{Parent}\ (a)}\sum_{a_j\in\text{Parent}(a)}\ \text{or}\ \left(a,r,a_j\right)\in G_y \quad (7)$$
$$\sigma\left(V\times\left[g\left(G_{y\to a_j}\right),\vec{e}_r\right]\right)$$

将 $g(G_{y\to a})$ 作为 a 的特征向量用于计算概率训练：

$$P_{\theta_2}(a|\ y,q) = \text{softmax}\left(f_{\text{qt}}(q)\cdot g\left(G_{y\to a}\right)\right) = \frac{\exp\left(f_{\text{qt}}(q)\cdot g\left(G_{y\to a}\right)\right)}{\sum_{a'\in V(G_y)}\exp\left(f_{\text{qt}}(q)\cdot g\left(G_{y\to a'}\right)\right)} \quad (8)$$

目标函数的原始形式为：

$$\max_{\theta_1,\theta_2}\frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{y\in V(G)}P_{\theta_1}\left(y|\ q_i\right)P_{\theta_2}\left(a_i|\ y,q_i\right)\right) \quad (9)$$

使用辨分推断得到目标函数的下界：

$$\max_{\psi,\theta_1,\theta_2}\mathrm{L}\left(\psi,\theta_1,\theta_2\right) \quad (10)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}_{Q_\psi(y|\ q_i,a_i)}\left[\log P_{\theta_1}\left(y|\ q_i\right)+\log P_{\theta_2}\left(a_i|\ y,q_i\right)-\log Q_\psi\left(y|\ q_i,a_i\right)\right]$$

基于强化学习的多跳路径探索不仅具有强大的多跳推论能力，而且具有很好的解释可能性。但是，这不仅关系知识，属性和事实知识也可以处理，问题必须有一个主体，适用范围小，没有其他的推论能力。另外，在不完全的 KG 环境中，根据训练数据中错误的负示例，接收低质量的返回，降低测试阶段的一般化能力。

2 医疗知识图谱构建方法与实现

医疗知识图是智能查询的基础，可以为人们提供更高效、更准确的医疗服务。根据医疗数据的特征，采用了基于收集的结构数据设计和构建医疗领域知识图的半自动方法。结合知识图的构建方法和医疗现场的特性，设计了本文中医学知识图谱的方法。

2.1 医学知识图谱构建流程设计

基于医学知识的专业性，本文使用人工筛选结合深度学习的组合来构建知识图谱，架构如图 6 所示。通过对互联网医疗专家网站的数据进行克隆，结合专家医生的意见，正确且专业的医疗知识图将根据特定规则融合相关实体和相关属性数据而获得。

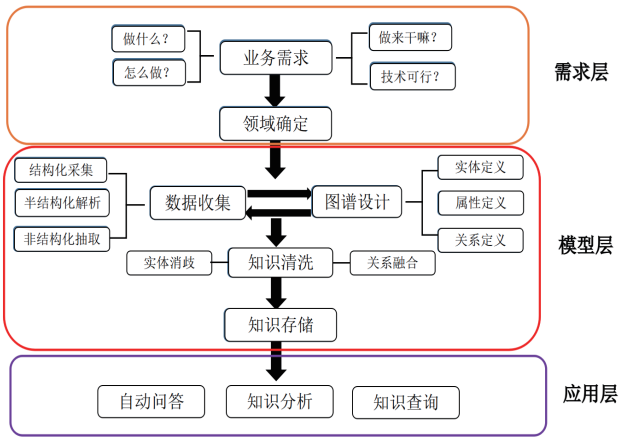

图 6 知识图谱构建架构

2.2 医学数据采集和知识提取

本文的医疗知识数据采集主要使用爬虫技术爬取包括丁香园、好大夫健康网、39 健康网等专业医学网站。爬取数据类型包括药品说明书、疾病描述、专家问诊数据等。

知识提取包括实体、属性和关系提取。通过前一节的说明，获得了主要的专业医学网站的医学数据。通过这些数据，需要提取实体信息，如实体、疾病、药物、人群、部门、实体属性、实体之间的关系等。

本文采用了一种基于字典和规则的命名实体提取方案。如流程图中说明的那样，规则制作是根据文本直接设定规则，另一个是根据单词分段后的单词和词性设定规则。

2.3 基于 Neo4J 的医学知识图谱的存储与可视化

本文采用图形数据库 Neo4J 对相关医疗数据和实体词表进行存储与组织管理 [19]，医疗知识图谱数据的来源主要有两个部分，一是前文中工作所积累的相关的疾病、症状、药品等医疗实体的词表；二是利用网络爬虫结合设定的规则从相关医学站点爬取的原始数据。这些数据可能是结构化数据、非结构化数据和半结构化数据，对于不同来源的不同类型数据，需要进行知识融合，也就是将代表相同概念的实体合并，将多个来源的数据集合并成一个数据集，之后建立知识图谱，进而通过推理获取新的知识，实现智能问答。

基于上述专家的医学网站，通过本文的实体提取、关系提取以及知识融合，最终得到完整的医学知识图，以及实体关系和实体关系属性。其中有 8868 个疾病实体，76 个部门实体，631 个人口实体，77 个部分实体，10 877 个征兆实体和 40 197 个药物实体。表 1 给出了各种实体之间的关系的数据结果。

表 1 各实体间关系数据结果

实体	科室	人群	部位	症状	疾病	药品
科室	*	*	*	*	16 163	*
人群	*	*	*	*	*	72 763
部位	*	*	*	14 579	9174	*
症状	*	*	14 579	72 844	90 841	209 217
疾病	16 163	*	9174	90 841	8868	79 394
药品	*	72 763	*	209 217	79 394	*

知识图谱最后被保存在 Neo4J 中。节点表示实体，实体之间的关系根据边连接。图是包括相应实体的属性信息的知识地图的节点的示例，但不包括与其他实体的关系。图 7 是 Neo4J 中医疗知识图谱的存储和展示。

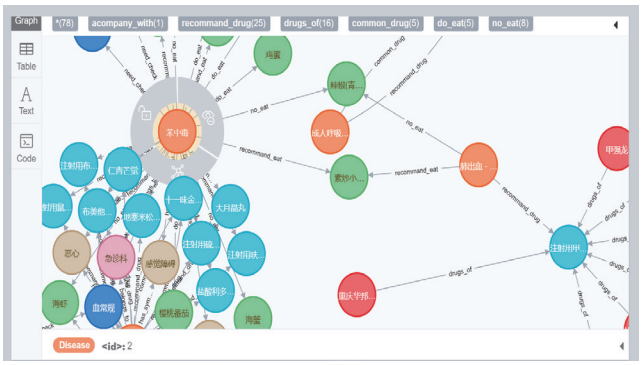

图 7 知识图谱展示

3 基于知识图谱的医疗问答系统实现

目前主流基于知识图谱的问答系统，基本都是通过关键字语义提取分析，之后遍历知识库进行匹配，反馈遍历结果。但是这种问答方式一方面依赖于知识库的语义解析和规则设定，另一方面对于复杂的自然语言问题处理效果往往不理想。除此之外，在医疗情境下，患者在表述自身症状时不一定准确，有可能出现缺失、心理臆想等问题。这在客观上给精准实现知识图谱问答系统实现智慧医疗情境带来了现实挑战。如图 8 所示，为本文阐述中所设计的医疗问答系统框架。

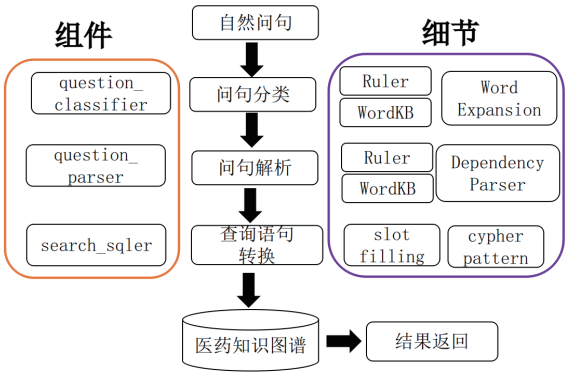

图 8 本文所述问答系统框架

3.1 基于卷积神经网络（CNN）的字 - 词分类模型

通过 n-gram 语言模型可知自然语言具有局部特征[20]。由于卷积神经网络在提取局部特征时表现优秀，局部特征所涉及的短文本在表达语义上也存在清晰明确的特点，因此卷积神经网络和短文本结合经常被用于自然语言的文本分类处理。

在本文中，使用卷积神经网络模型来有意义地表现患者的状态描述，并将其映射到基于知识的对应关系 / 属性中。集合特定医疗情境与具体科目，实现患者用户语义的分析提取。关系 / 属性映射文本分类任务 CNN 模型的特定方法如图 9 所示。

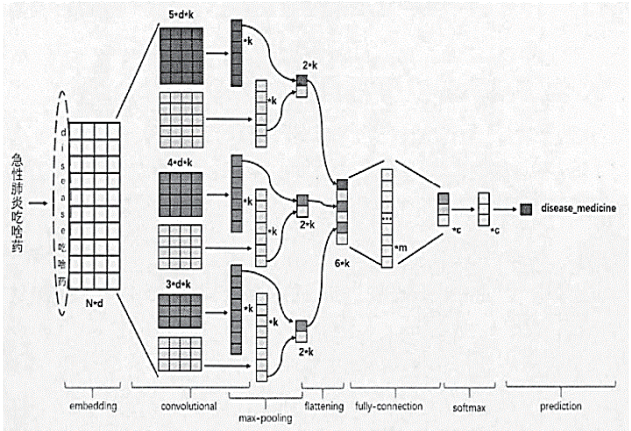

图 9 基于 CNN 的问句关系 / 属性映射模型图

3.2 信息归一化与查询问答逻辑转化

患者在相关情境下使用的自然语言词汇与相关医疗知识库中的语义实体存在一定区别和差异，因此在接收到患者相关自然语言语料之后，需要根据医疗知识库中的实体与关系存储情况对其进行归一化处理。之后通过实体识别模型提取出的医疗实体进行语义解析与遍历匹配处理，最后对知识库进行检索得出答案。

在患者相关问题语义被知识库正确解析之后，系统一般首先会根据语义提取出患者提问所属类型，具体涉及关系和属性的特征界定。患者提出的相关问题会在语义解析之后匹配对应的规则转换成 Neo4J 中的查询语句，通过提供的查询系统结构进行匹配对应知识库中的答案。整个过程其实是从实现自然问题语义学解析到计算机知识数据逻辑的转换处理。

3.3 系统部署与医疗问答系统实现

本系统目前只能在本地环境通过命令提示符运行，测试环境为 Windows10 64 位操作系统，系统部署流程如图 10 所示，先安装平台开发环境，再导入模型数据，之后构建知识图谱，最后启动服务。

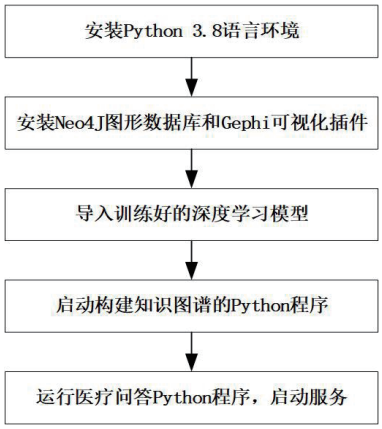

图 10 系统部署流程

本章节对基于知识图谱的智能问答系统选取部分案例对其功能进行测试。测试平台为 Windows10，使用 Python 版本为 3.8.3。根据前文构建的模型问答系统对在线用户咨询问题的回复效果进行测评，从用户日志中随机采样若干有效问题，最终搭建的医疗问答系统运行解答过程如图 11 所示。

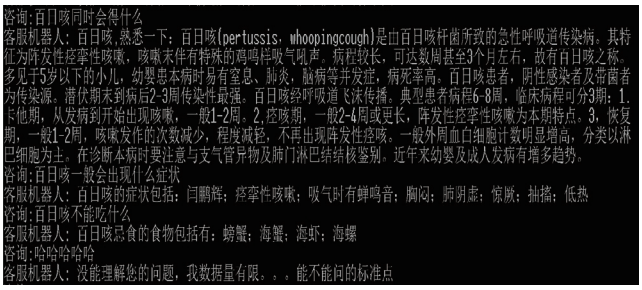

图 11 知识图谱问答系统演示

如图 11 所示，测试“百日咳一般会出现什么症状”，知识图谱给出答案“百日咳的症状包括：痉挛性咳嗽；吸气时有蝉鸣音；胸闷；肺阴虚；惊厥；抽搐；低热”。能够清晰详细地展示出以患者病情特征为核心实体为中心的知识图谱关系。

为了进一步验证本文设计实现的基于问答系统的有效性与精确性，笔者还特意邀请了不同相关患者用户针对自己就

医过程涉及的病症知识问题对系统提问。实践证明，本系统对于患者提出的医疗问题给出了科学合理、切合实际的回答，具有一定的实际应用价值。

4 结语

本系统结合目前用户的常见医疗问答情境需求和对应场景下医疗数据的特点，同时根据医疗知识图谱建设需求，开发了不同爬虫策略从医疗网站上获取相关生产医疗数据，进一步构建了医疗实体关系知识图谱。在存储和组织上选择了Neo4J 图数据库对图谱进行处理，并在此基础上构建实现了医学健康领域知识问答系统。从问答系统运行结果和响应时间来看，它能较为准确和及时地满足用户的需求。但从实际需求用户来说，本系统在数据提取和推理演绎上存在问句分类的算法不够高效、语义分析精准不足等问题，如何进一步提升医疗问答系统的精准性和增强用户体验，将是后续工作的研究重点。

参考文献：

[1]SINGHAL A.Official google blog:introducing the knowledge graph:things,not strings[EB/OL].(2015-01-02)[2024-02-18]. https://www.mendeley.com/catalogue/ca10852b-1f48-3172-9b4f-8b521878b39d/.

[2]AUER S, BIZER C, KOBILAROV G, et al.DBpedia:a nucleus for a web of open data[J]. Journal of web semantics, 2007, 7(3): 154-165.

[3]BOLLACKER K D, EVANS C, PARITOSH P, et al.Freebase:a collaboratively created graph database for structuring human knowledge[C]//SIGMOD ‘08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data.New York,United States:Association for Computing Machinery,2008:1247-1250.

[4]HAO Y, ZHANG Y, LIU K, et al.An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge[C]//55th annual meeting of the Association for Computational Linguistics:long papers,vol.1. Stroudsburg,PA: Association for Computational Linguistics, 2017:221-231.

[5]MITCHELL T, COHEN W, HRUSCHKA E, et al. Never-ending learning[J].Communications of the ACM,2018,61(5):103-115.

[6]SWARTZ A.Musicbrainz:a semantic web service[J].IEEE intelligent systems,2002, 17(1):76-77.

[7]DODDS K.Popular geopolitics and audience dispositions: James Bond and the internet movie database(IMDb)[J].Transactions of the institute of british geographers,2006, 31(2):116-130.

[8]BODENREIDER O.The unified medical language system (UMLS):integrating biomedical terminology[J].Nucleic acids research, 2004,32(1):D267-D270.

[9]WISHART D S, FEUNANG Y D, GUO A, et al.DrugBank 5.0:a major update to the DrugBank database for 2018[J]. Nucleic acids research,2018,46(D1):D1074-D1082.

[10]LICATO J, BRINGSJORD S, GOVINDARAJULU N S.How models of creativity and analogy need to answer the tailorability concern[M]//Computational Creativity, Concept Invention, and General Intelligence. [S.l.]:[s.n.], 2013:9-16.

[11]SCHRIML L M, ARZE C, NADENDLA S, et al.Disease ontology:a backbone for disease semantic integration[J]. Nucleic acids research, 2012,40(D1):D940-D946.

[12]LANDRUM M J, LEE J M, BENSON M, et al.ClinVar:improving access to variant interpretations and supporting evidence[J].Nucleic acids research,2018,46(D1): D1062-D1067.

[13]ZHAO C, JIANG J, XU Z, et al.A study of EMR-based medical knowledge network and its applications[J].Computer methods and programs in biomedicine,2017,143:13-23.

[14]YANG J, YU Q, GUAN Y, et al.An overview of research on electronic medical record oriented named entity recognition and entity relation extraction[J].Acta automatica sinica, 2014, 40(8): 1537-1562.

[15]KUHN M, LETUNIC I, JENSEN L J, et al.The SIDER database of drugs and side effects[J]. Nucleic acids research, 2016, 44(D1):D1075-D1079.

[16] 杜睿山 , 张轶楠 , 田枫 , 等 . 基于知识图谱的智能问答系统研究 [J]. 计算机技术与发展 ,2021,31(11):189-194.

[17] 黄承宁 , 李双梅 , 景波 . 基于深度学习表示的医学主题语义相似度计算研究 [J]. 计算机与数字工程 ,2022,50(6):1149-1152.

[18] 陈梅梅 . 基于知识图谱的医疗问答系统设计与实现 [D]. 厦门 : 厦门大学 ,2019.

[19] 张崇宇 . 基于知识图谱的自动问答系统的应用研究与实现 [D]. 北京 : 北京邮电大学 ,2019.

[20] 卢琪 , 潘志松 , 谢钧 . 融合知识表示学习的双向注意力问答模型 [J]. 计算机工程与应用 ,2021,57(23):171-177.

【作者简介】

黄承宁（1985—），男，江苏南京人，博士，副教授，研究方向：知识发现、人工智能、计算机教育应用。

殷晓磊（1987—），男，江苏南通人，硕士，工程师，研究方向：计算机技术应用、信息管理与信息系统。

陈武（1982—），男，海南海口人，硕士，审计师，研究方向：大数据与审计、内控审计。

（收稿日期：2024-03-14）