基于医疗知识图谱的问答机器人

**一、引言**

大家都知道，人类历史发展到现在，已经经历了三次工业革命，每次工业革命都会带来翻天覆地的变化，生产力的进步带来了整个社会的变革，生活方式随之发生深刻变化。第一次工业革命使人类走向机械化时代，第二次是带来了电力，第三次是信息化革命。我们很幸运，今天处在第四次工业革命的开端，第四次工业革命的核心驱动力就是人工智能。当然，人工智能是为了模拟人的能力，需要包括语音、视频、图像、AR/VR 等感知方面的技术，也需要知识图谱、语言理解等认知方面的技术。当然，还有机器学习，以及最近这些年很重要的深度学习等等。

过去这些年，AI 技术通过深度学习取得了非常好的效果，尤其是在语音、视觉等感知技术上取得了非常大的突破，甚至在很多方面，深度学习达到的效果已经超过了人类。深度学习也给自然语言处理以及知识相关的技术带来了非常大的帮助，但是我们继续深入研究、应用实践的时候会发现，我们还需要更好地结合知识、推进知识图谱相关的工作，才能取得更好的效果。

基本的知识图谱，会涉及到实体的属性关系，每一个实体可能有若干个属性，实体和实体之间有很多关系，每一个关系基本上可以理解为是一个事实。因此知识图谱就是对客观世界的描述。本次案例的目的在于理解知识图谱，学习知识图谱构建。用途在于推进、结合知识图谱工作来加强人工智能的技术。

1. **案例内容**

在本次案例中首先描述了如何构建知识图谱框架以及基于知识图谱如何设计特定领域的框架，基于医疗知识图谱的问答机器人。首先以任务为驱动构建知识图谱框架，是采用自下向上的构建方式，主要由以下过程：

1. 对任务进行分类以确定领域。即先对将要进行的任务进行整理以及分析，从而确定任务领域的分类。
2. 模型的确定。进而结合各领域的特点归纳、总结，形成一个框架，也就是数据模型。在此阶段通过对数据的收集进行图谱的设计，然后进行知识的清洗，即进行各实体之间的关系融合，从而解决各实体之间不同构的问题。
3. 最后是进行应用，通过知识分析以及知识查询，实现自动问答。比如google，百度的知识图谱，是属于典型的公共领域知识图谱，现实中，使用他们的搜索工具进行内容搜索时，用户可能输入的内容千差万别，各个领域的问题都可能问到，也就使得他们的后台知识图谱内容也要覆盖所有知识，在构建他们这种公共领域的知识图谱过程中，随着数据的不断积累，才会对数据知识进行分类，慢慢呈现出知识架构。

具体如下图所示：

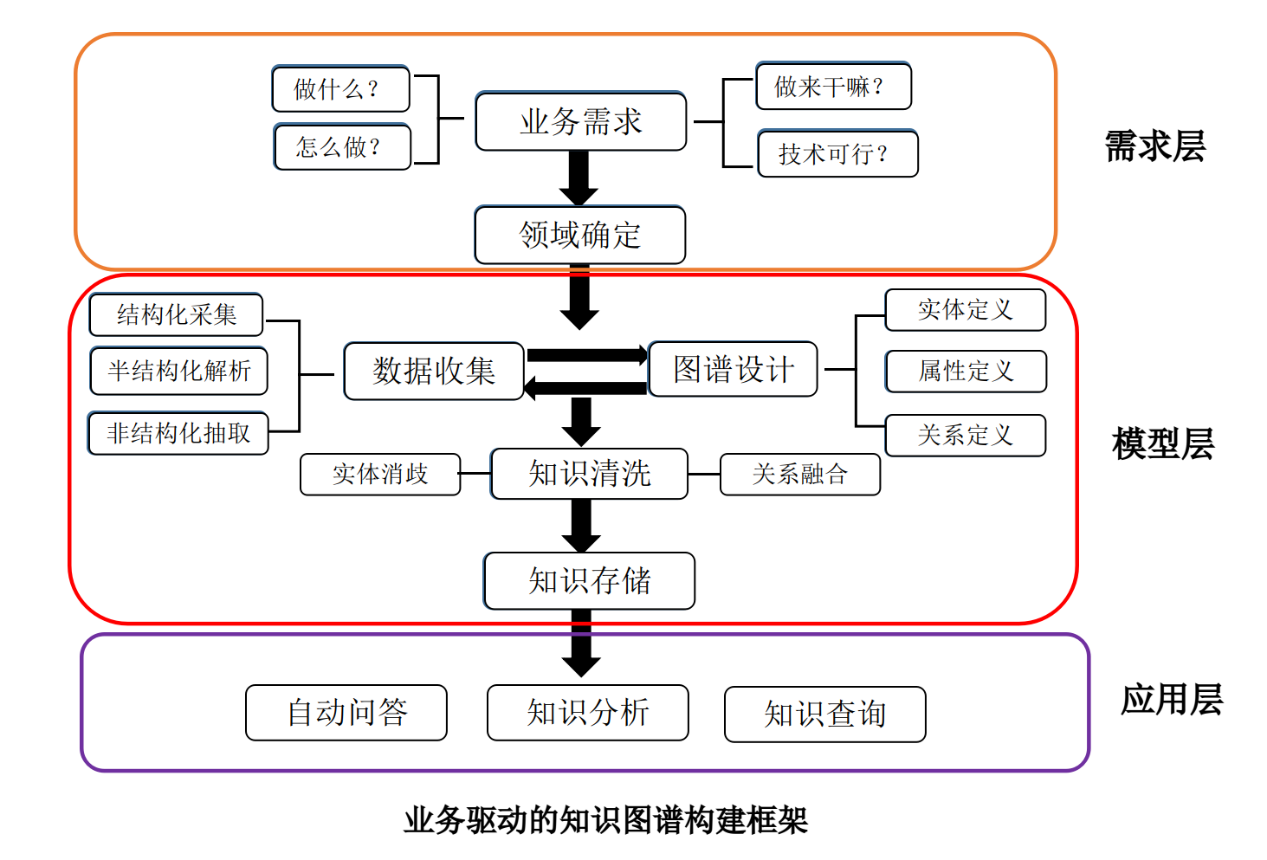


图 业务驱动的知识图谱构建框架

基于知识图谱的问答框架对输入的自然问句通过特定组件进行问句的分类、解析以及语句转换的操作，最后和存储在neo4j中的数据进行匹配，然后返回结果。

具体过程如下图2基于知识图谱的问答框架所示：

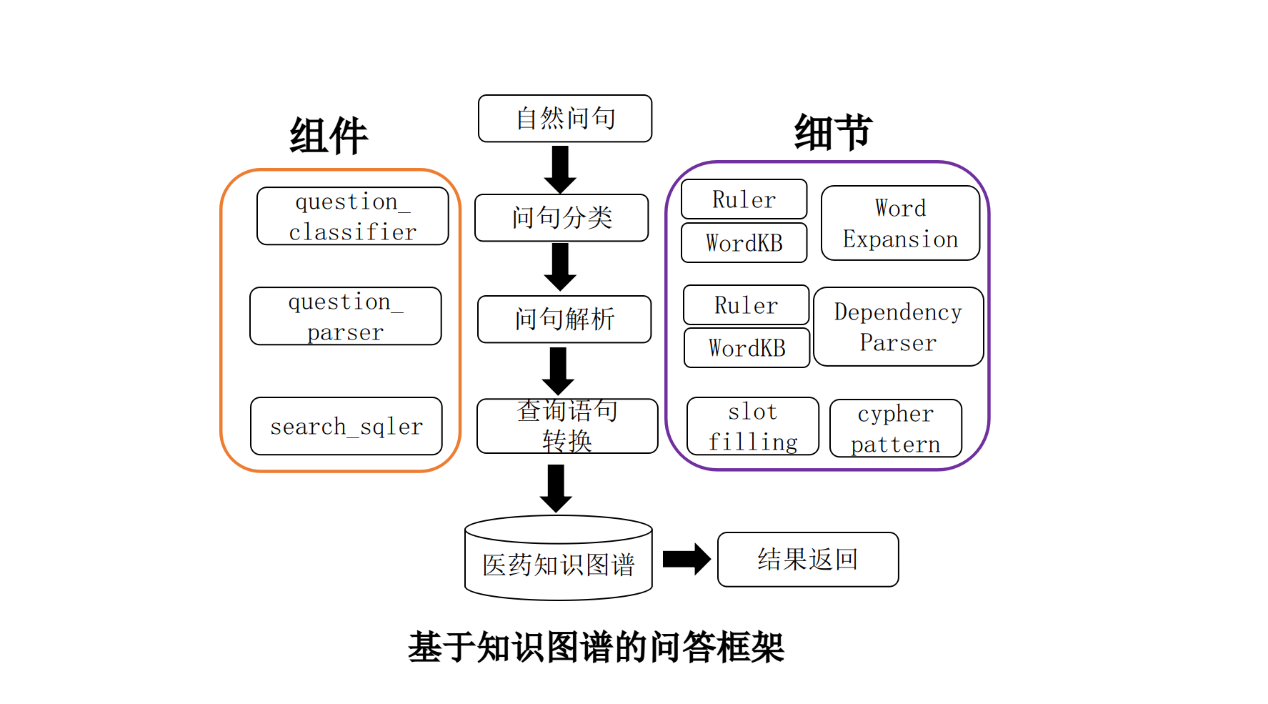


图 基于知识图谱的问答框架

**三、案例解决方案与分析**

知识抽取：

通过爬虫（lxml-etree、xpath）来提取页面中相关的信息，进行持久化，并将持久化的数据进行整理与清洗，得到格式化的的数据

图谱构建：

如何去构建图谱，各实体间的关系怎么去确定？都是影响图谱执行效率的因素。在这里，实体设计与数目如下图所示：



图3 医疗知识图谱实体

实体关系类型的设计与数目如下：



图4 医疗知识图谱实体关系

属性类型的设计与数目如下：

图5 医疗知识图谱属性

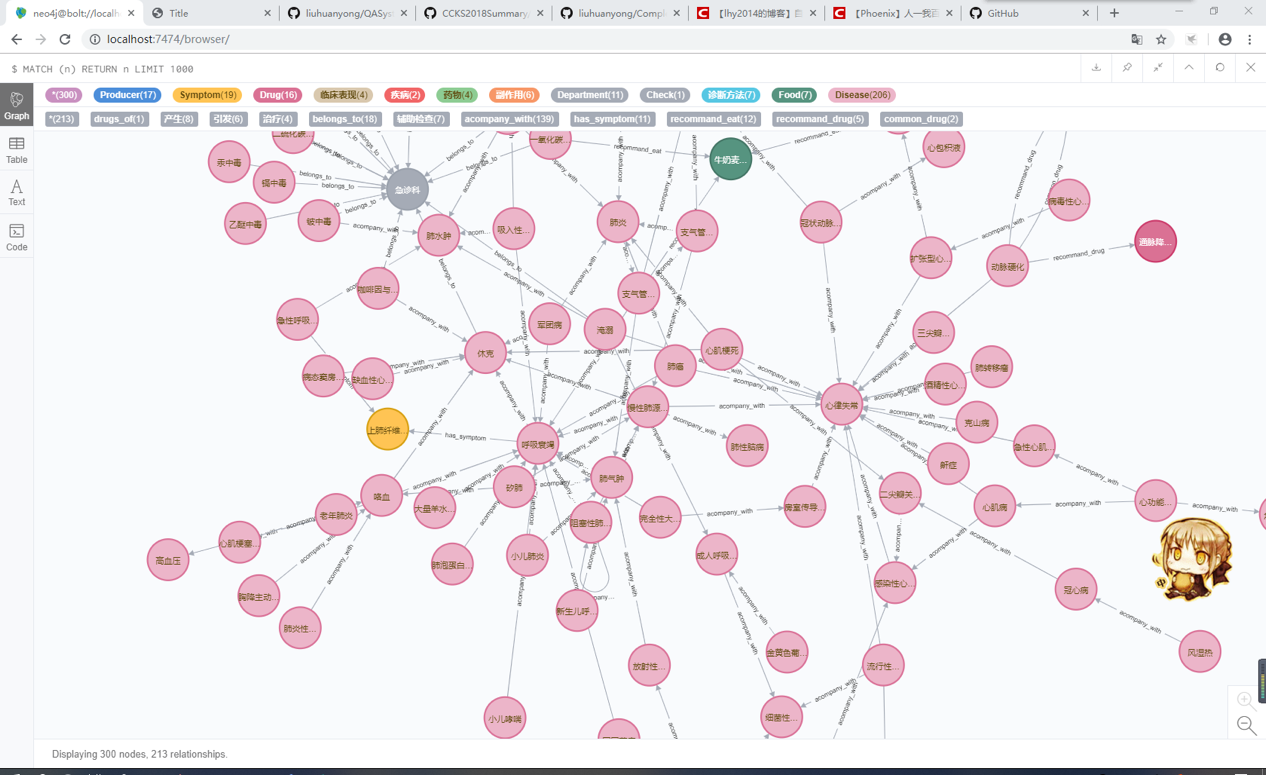


图6 本地搭建的依赖图谱

基于规则匹配实现问答：

本项目的问答系统完全基于规则匹配实现，通过关键词匹配，对问句进行分类，医疗问题本身属于封闭域类场景，对领域问题进行穷举并分类，然后使用cypher的match去匹配查找neo4j数据，根据返回数据组装问句回答，最后返回结果。

流程如下：

1. 对问句中的关键词进行匹配：

包括最大向前匹配、最大向后匹配、双向最大向前匹配

1. 根据匹配的关键词进行分类：

包括创建特征词路径、加载特征词、构造领域、构建词典、问句疑问词标记

1. 问句解析：

包括构建节点实体关系、建立节点、创建知识图谱中心疾病的节点

1. 查找相关关系：

包括执行cypher查询，并返回相应结果、根据对应的qustion\_type，调用相应的回复模板

1. 根据返回的数据组装回答：

导出数据

脚本结构：question\_classifier.py：问句类型分类脚本

question\_parser.py：问句解析脚本

chatbot\_graph.py：问答程序脚本

支持问答类型：



图7 项目所支持的问答类型

部分问答结果：

用户:最近老流鼻涕怎么办？

小勇: 症状流鼻涕可能染上的疾病有：枯草热；副流行性感冒；急性上呼吸道感染；硫化氢中毒；小儿衣原体肺炎；风寒感冒；慢性额窦炎；鼻源性头痛；人禽流行性感冒；小儿流行性喘憋性肺炎；病毒性感冒；慢性鼻炎；风热犯肺；感冒；顿呛；小儿急性上呼吸道感染；嗜酸细胞增多性非变态反应性鼻炎；干酪性鼻窦炎；下呼吸道感染；麻疹

用户:感冒要多久才能好？

小勇: 感冒治疗可能持续的周期为：7-14天

用户:高血压要怎么治？

小勇: 高血压可以尝试如下治疗：药物治疗;手术治疗;支持性治疗

用户:白血病能治好吗？

小勇: 白血病治愈的概率为（仅供参考）：50%-70%

用户:什么人容易得高血压？

小勇: 高血压的易感人群包括：有高血压家族史，不良的生活习惯，缺乏运动的人群

用户:脑膜炎怎么才能查出来？

小勇: 脑膜炎通常可以通过以下方式检查出来：脑脊液钠；尿常规；Fisher手指试验；颈项强直；脑脊液细菌培养；尿谷氨酰胺；脑脊液钾；脑脊液天门冬氨酸氨基转移酶；脑脊液病原体检查；硝酸盐还原试验

用户:板蓝根颗粒能治啥病？

小勇: 板蓝根颗粒主治的疾病有流行性腮腺炎；喉痹；喉炎；咽部异感症；急性单纯性咽炎；腮腺隙感染；过敏性咽炎；咽囊炎；急性鼻咽炎；喉水肿；慢性化脓性腮腺炎；慢性咽炎；急性喉炎；咽异感症；鼻咽炎；锁喉痈；小儿咽喉炎；喉返神经损伤；化脓性腮腺炎；喉血管瘤,可以试试

用户:为什么有的人会失眠？

小勇: 失眠可能的成因有：躯体疾病和服用药物可以影响睡眠，如消化不良，头痛，背痛，关节炎，心脏病，糖尿病，哮喘，鼻窦炎，溃疡病，或服用某些影响中枢神经的药物。 由于生活方式引起睡眠问题也很常见，如饮用咖啡或茶叶，晚间饮酒，睡前进食或晚饭较晚造成满腹食物尚未消化，大量吸烟，睡前剧烈的体力活动，睡前过度的精神活动，夜班工作，白天小睡，上床时间不规律，起床时间不规律。 可能的原因有压力很大，过度忧虑，紧张或焦虑，悲伤或抑郁，生气，容易出现睡眠问题。 吵闹的睡眠环境，睡眠环境过于明亮，污染，过度拥挤。

用户:失眠的人不要吃啥？

小勇: 失眠忌食的食物包括有：油条；河蚌；猪油（板油）；淡菜(鲜)

用户:失眠有哪些并发症？

小勇: 失眠的症状包括：心肾不交；神经性耳鸣；咽鼓管异常开放症；偏执狂；十二指肠胃反流及胆汁反流性胃炎；腋臭；黧黑斑；巨细胞动脉炎；Stargardt病；抑郁症；腔隙性脑梗死；甲状腺功能亢进伴发的精神障碍；紧张性头痛；胃下垂；心血虚；迷路震荡；口腔结核性溃疡；痰饮；游走性结节性脂膜炎；小儿脑震荡

**四、启发**

通过复现一个基于知识图谱的医疗智能问答程序，我们可以大致的清楚所提及的知识图谱是什么东西。按照我个人的理解，他实际上是一个以图的结构进行存储的数据库，节点与边权各自有不同的含义，但是他比起其他最具有优势的地方，在于他的推理能力。

以图作为存储结构，比起常规的SQL数据库，没有行列的限制；比起No-SQL数据库，图数据库突出了存储的数据间的关系。我们后续的一些操作，都是基于这样的图数据库来进行实现。

我们可以依据实体、关系、属性，实现基于水务设备的物联网图谱服务，目前打算的是基于图谱的推理，实现设备运行状况的报告生成，后期会计划在此基础之上实现基于推理的设备智能控制。