项目编号：_____

吉林大学“大学生创新创业训练计划”

创新训练项目

申请书

项目名称 基于知识图谱的语义搜索及其应用

项目负责人 李超

所在学院、年级、专业 计算机科学与技术学院

2017 级 计算机科学与技术

联系电话 18843148082

电子邮箱 lichao0528@163.com

指导教师姓名 孙延风 职称 副教授

填表日期 2019 年 4 月 23 日

吉林大学教务处制表

填表须知

一、**本表适用于创新训练项目**。本科生个人或团队，在校内导师指导下，自主完成创新性实验方法的设计、设备和材料的准备、实验的实施、数据处理与分析、总结报告撰写等工作。

二、申报书请按顺序逐项填写，实事求是，表达明确严谨。空缺项要填“无”。

三、申请参加大学生创新训练项目团队的人数为 3—5 人。

四、申请项目，必须聘请教师作为指导老师，并请指导教师在申请书中的指导教师意见栏中签署意见。

五、填写时可以改变字体大小等，但要确保表格的样式不变；不得随意涂改；A4 纸正反面打印，左侧装订。

六、本表由项目负责人报所在学院初审，学院签署初审意见后报送教务处实践教学科（一式 3 份原件）。

七、“项目编号”由教务处填写。

八、申报过程有不明事宜，请与教务处实践教学科联系，电话 85166413。

<table>
<tr><td colspan="2">项目名称</td><td colspan="8">基于知识图谱的语义搜索及其应用</td></tr>
<tr><td colspan="2">项目起止时间</td><td colspan="8">2019 年 5 月 至 2020 年 5 月</td></tr>
<tr><td rowspan="6">负责人
项目组成员</td><td>姓名</td><td>学院</td><td>专业</td><td>教学号</td><td>联系电话</td><td>E-mail</td><td>QQ</td><td colspan="2">各类实验班</td></tr>
<tr><td>李超</td><td>计算机科学与技术学院</td><td>计算机科学与技术</td><td>21172511</td><td>18843148082</td><td>Lichao0528@163.com</td><td>1141801706</td><td colspan="2">是□ 否√</td></tr>
<tr><td>唐思源</td><td>软件学院</td><td>软件工程</td><td>55170902</td><td>15567079395</td><td>1063071473@qq.com</td><td>1063071473</td><td colspan="2">是□ 否√</td></tr>
<tr><td>邓新柯</td><td>计算机科学与技术学院</td><td>网络信息与安全</td><td>21171204</td><td>15343296976</td><td>635477622@qq.com</td><td>635477622</td><td colspan="2">是□ 否√</td></tr>
<tr><td>高明龙</td><td>数学学院</td><td>统计学</td><td>10170548</td><td>18843146955</td><td>1506237994@qq.com</td><td>1506237994</td><td colspan="2">是□ 否√</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="2">是□ 否□</td></tr>
<tr><td rowspan="4">指导教师1</td><td>姓名</td><td colspan="4">孙延风</td><td>职务/职称</td><td colspan="2">副教授</td></tr>
<tr><td>所在单位</td><td colspan="7">吉林大学计算机科学与技术学院</td></tr>
<tr><td>联系电话</td><td colspan="3">18704482627</td><td>E-mail</td><td colspan="3">sunyf@jlu.edu.cn</td></tr>
<tr><td>对本课题相关领域研究情况</td><td colspan="7"></td></tr>
<tr><td colspan="2">项目性质</td><td colspan="8">1.小发明、小创作、小设计（ ）
2.开放实验室或实习基地中的创新性实验或新实验开发（ ）
3.基础性研究（ ）
4.应用性研究（√）
5.社会调研 （ ）</td></tr>
<tr><td colspan="2">项目选题来源</td><td colspan="8">1.自主立题（√）
2.教师科研课题的子项目（ ）</td></tr>
<tr><td colspan="2">项目学科类别</td><td colspan="8">计算机科学与技术</td></tr>
<tr><td colspan="2">项目受其他渠道资助情况（填“无”或具体资助来源和经费，包括获奖情况）</td><td colspan="8">无</td></tr>
</table>

一、立项背景和依据（包括研究目的、国内外研究现状分析与评价、研究意义，应附主要参考文献及出处）

1.基层医疗面临的困境

我国的基层医疗卫生机构包括社区卫生服务中心（站）、街道卫生院、乡镇医院、村卫生室、门诊部、诊所（医务室）等，占我国医疗卫生机构总数的93%以上[1]。为改善“看病难、看病贵”的现状，国家大力推行分级诊疗制度，改善基层医疗。但是由于基层医疗机构的资源条件和服务水平受到多种限制，很难为患者提供满意的治疗方案。基层医疗机构的设备数、业务用房面积和执业医师等基础条件占比均小于医院，导致患者就诊不便利，就诊体验差。在基层医生未接受过科学的医学技能培训的情况下，会出现大量根据“经验”诊疗缺少科学推测，分析疾病不全面、医嘱建议不完善等问题。

我们设想运用知识图谱、语义搜索等技术建立一个医疗咨询系统，为患者提供关于某种常见病的专业医学信息，让患者获取信息更加便利，疾病诊断更加科学、合理、全面。

2.国内外研究现状分析与评价

近年来，人工智能特色医疗技术已经取得了长足的发展，借助互联网发展而完善医疗体系是大势所趋。目前，国内外研究学者已经涉猎到智能问答系统、专家系统、辅助医疗分析系统等多项领域。与数据库、多媒体技术相结合让智能医疗更具多样性，深度学习的运用更是让该领域散发出新的活力。然而目前多数研究成果都将应用场景定位在大医院，比如基于病例的诊断系统和医学影像处理，受众基数最大的基层医疗机构却被人所忽视。可见，开发适合基层患者使用的医疗系统是当前的迫切诉求。

2.1 知识图谱

谷歌于2012年首次提出“知识图谱”概念[2]，是语义网络的一种表现形式。知识图谱以实体为节点，以实体之间的关系为边，核心思想是可视化地展现结构化的信息和信息之间的逻辑关系。通过知识图谱，可以获取知识和知识之间的逻辑关系，将抽象的知识可视化地展现出来，以及得到每一知识点全面的结构化信息。知识图谱初衷是用于提升搜索引擎返回结果的质量。各大搜索引擎的公司也推出了自己的知识图谱，例如百度的“知心”、搜狗的“知立方”。同时国内外现在已有很多规模巨大的知识图谱，如DBpedia已经包含约30亿RDF三元组，多语种的大百科语义网络BabelNet包含19亿的RDF三元组，Yago3.0包含1.3亿元组，阿里巴巴于2017年8月份发布的仅包含核心商品数据的知识图谱已经达到百亿级别。相较于通用知识图谱的发展，领域知识图谱的研究也是热点。医学是知识图谱应用最广的垂直领域之一，各大科研机构也构建了医学领域的知识图谱，如上海曙光医院构建的中医药知识图谱、本体医疗知识库SNOMED-CT，IBM Watson Health。

2.2 语义搜索

语义搜索分为语义理解和知识检索。通常认为语义分析存在两种研究思路:第一种是以语法分析为中心的语义分析，第二种是以语义为中心的句义分析。如美国语言学家菲尔默于 1968 年提出的格语法分析模式和美国学者 R．Schank 于 1973 年提出的概念依存理论都是典型的句义分析模式。而针对于语义解析，通常会采用信息抽取法、向量建模法和深度学习法等研究方式。针对本项目基于知识图谱的语义搜索系统，最前沿的智能检索方法[5]是使用深度学习方法训练检索模型。迄今为止效果最好的有两种方法，一是基于 Bi-LSTM 神经网络，直接用问题-答案对作为数据，训练端对端神经网络；二是基于记忆神经网络，利用记忆网络存储知识图谱内容，将用户的输入转化为内部的分布式表达，根据表达在记忆模块中寻找相关记忆，根据输出格式输出结果。

3.研究意义

我们以“常见病咨询”为基础模型，建立概念，意在建立一个基于常见病症信息的小型领域知识图谱，通过实现基于知识图谱的语义搜索功能，提供类似医疗咨询的服务模式。我们利用知识图谱来存储常见病的信息，患者向系统描述症状就可以得到精确的相关信息。我们会使用语义搜索技术来保证系统捕获精确的症状描述，从而准确返回信息。常规搜索引擎充斥着大量的无关信息与广告，很难检索到想要的专业医学知识。基层医疗机构就诊体验差，诊疗过程不专业。综合以上现状，我们选择基层医疗作为我们咨询模式的服务群体。

在医疗水准的提升、医疗资源的下沉等方面，人工智能或将是一味济世良药。传统“就医难”问题一直是各个机构国家头疼的事情，让计算机“学习”疾病的专业知识，作为广大群众的咨询“帮手”，将带动基层医疗发展，推动国家医疗卫生建设。

参考文献：

[1] 梁铭会，尹畅，董四平. 我国医疗质量监管体系制度变迁分析及思考[J]．中国卫生质量管理，2011，18（6）：13-17.

[2]AMIT S．Introducing the knowledge graph[R]．America：Official Blog of Google.2012．

[3]SUCHANEK F M，KASNECI G，WEIKUM G. YAGO：a large ontology from Wikipedia and wordnet[J]. Web Semantics Science Services & Agents on the World Wide Web，2007，6(3)：203-217．

[4]BIZER C，LEHMANN J，KOBILAROV G，et a1.DBpedia--a crystallization point for the Web of data[J]. Web Semantics Science Services & Agents on the World Wide Web，2009，7(3)：154-165．

[5] 郭嘉丰,范意兴.深度学习检索框架的前沿探索[J].计算机研究与发展,2018,55(09).

[6] 秦春秀;祝婷;赵捧未;张毅;.自然语言语义分析研究进展[J].西安电子科技大学, 2014(22).

[7] 牛亚冬;张研;叶婷;张亮;.我国基层医疗卫生机构医疗服务能力发展与现状[J].中国医院管理, 2018(06).

[8] 王飞;陈立;易绵竹;谭新;张兴华;. 新技术驱动的自然语言处理进展[J].武汉大学学报, 2018(08).

[9]Berant J,Chou A，Frostig R，et a1．Semantic parsing on freebase from question-answer pairs[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.2013：1533-1544．

[10] Yao X， Durme B V. Information Extraction over Structured Data: Question Answering with

Freebase[C]//Meeting of the Association for Computational. Linguistics．2014:956—966．

[11]Dong L, Wei F, Zhou M，et a1．Question Answering over Freebase with Multi— Column Convolutional Neural Networks[C]// Meeting of the Association for Computational Linguistics and the，International Joint Conference on Natural Language Processing．2015: 260-269．

[12]Suldabaatar S，Weston J'Fergus R．End-to—end memory networks[C]// Advances in neural information processing systems．2015：2440—2448．

[13]Zhang Y Liu K He S, et a1．Question Answering over Knowledge Base with Neural ARenfion Combining Global Knowledge Information[J]．2016．

[14]Feng G F, Du Z K, Wu X.A Chinese Question Answering System in Medical Domain[J].Shanghai Jiao Tong Univ.(SCI), 2018, 23(5):678-683.

二、项目研究内容（项目主要研究内容；拟解决的关键问题、重点和难点）

1. 主要研究内容

(1)设计并实现常见病知识图谱构建系统，用于实体和实体间关系的抽取以及语句的分析;
(2)设计并实现知识图谱的管理系统，用于知识图谱的存储和可视化;
(3)设计并实现基于知识图谱的医疗语义搜索系统，用于根据用户的输入在知识图谱中进行搜索并返回答案;
(4)融合以上系统，基于知识图谱为用户提供常见病医疗知识图谱可视化、医疗语义搜索服务。

2. 拟解决的关键问题

技术层面:

(1) 如何构建一个实体识别模型和关系抽取模型，以实现从数据库中提取高质量的实体和关系，并且对用户的输入进行语义解析;

(2) 医疗保健数据非常模糊和嘈杂，疾病的名称可能因不同的网站而异，现有的系统难以处理患者的问题，特别是对于同义词等情况;

(3) 如何存储知识图谱。存储知识图谱既要考虑到大量数据存储的空间压力，也要保证数据检索时的速度;

(4) 如何基于存储结构可视化知识图谱;

(5) 如何对用户的输入进行语义分析;

(6) 如何根据提取出的语义在知识图谱中进行搜索，返回答案。

社会层面:

(1)如何获得广大患者对搜索结果的信任

(2)如何在欠发达地区群众中推广

三、项目特色及创新点

1.项目特色

①利用知识图谱存储常见病症信息并可视化展现抽象知识；

②通过语音交互板和语义识别进行语义搜索，可提供常见病医疗咨询的服务功能，成为家中的专业医生；

③聚焦基层医疗的医疗资源分配不均衡问题，以基于知识图谱的语义搜索形式提供常见病症咨询,使得智慧医疗推动基层医疗发展。

2.创新点

①知识图谱应用于语义搜索的实现。传统的医疗信息检索总是基于搜索引擎和领域专家，用户必须搜索大量的在线网站以找到有用的信息。语义搜索利用知识图谱可以准确捕捉用户搜索意图，借助于项目所需医学领域的知识图谱，将获取的输入与知识图谱映射，根据知识图谱结构返回符合搜索意图的答案，而不是包含关键词的相关网页链接，从而大幅提升搜索结果质量。

②聚焦基层医疗层面，智慧医疗新思路。提供常见病症医疗语义搜索功能，以期成为家庭中的专业医生，将在日常医疗保健和自身疾病的诊断中发挥重要作用。

四、申请理由（1、团队条件——自身/团队具备的知识、素质、能力、特长、兴趣；2、前期准备基础等）

1.团队条件

●李超：计算机科学与技术学院大二学生，学习成绩年级前 15%，曾获得校二等奖学金。参加过开放性创新实验并获得优异成绩，已经掌握 Java、python 、C、C++、等知识，编程基础扎实。对语义分析、深度学习具有兴趣。参加过数学建模国赛、美赛，具备阅读及理解文献能力，有中英文学术论文写作经验。具有快速学习能力，思维活跃，踏实认真，努力肯钻，团队合作能力强。

●唐思源：软件学院大二学生，学习成绩年级前 15%，曾获得校二等奖学金。已经掌握 Java、数据库、数据结构、C 语言、C++等知识，编程基础扎实。参加过数学建模国赛、美赛，有良好的阅读及理解文献能力，有中英文学术论文写作经验。乐于探索新事物，做事认真负责，具有良好的团队协作能力。

●邓新柯：计算机学院网络信息安全专业大二学生，能熟练使用 Java、python、HTML、JS、CSS、PHP、C、C++等语言，对 Oracel、MsSQL、MySQL 等数据库的操作都熟练掌握，对 IIS、Apache 等服务器也能熟练使用，此外对网络协议，网络通信等也比较熟悉。目前对机器学习，知识图谱等人工智能方面也有所涉及。个人性格而言，自我学习能力强，喜欢思考，乐于和别人交流，热衷并且善于团队合作。

●高明龙：数学学院统计学专业大二学生，专业排名前 10%，曾获得校二等奖学金、院优秀学生。对数学分析、高等代数等数学专业课程有较好的理解，了解 python 编程知识，对机器学习相关知识具有兴趣，并有一定的了解。参加过数学建模国赛、美赛，具有研究文献能力，具备中英文学术论文写作经验。学习认真努力，态度端正，能够独立自主地思考，团队合作能力强。

2.前期准备：

(1)小组成员已经完成了微积分、线性代数、概率论等必备数学知识的学习，数学知识储备较为充足并对深度学习基础知识和自然语言处理相关知识具有一定了解。

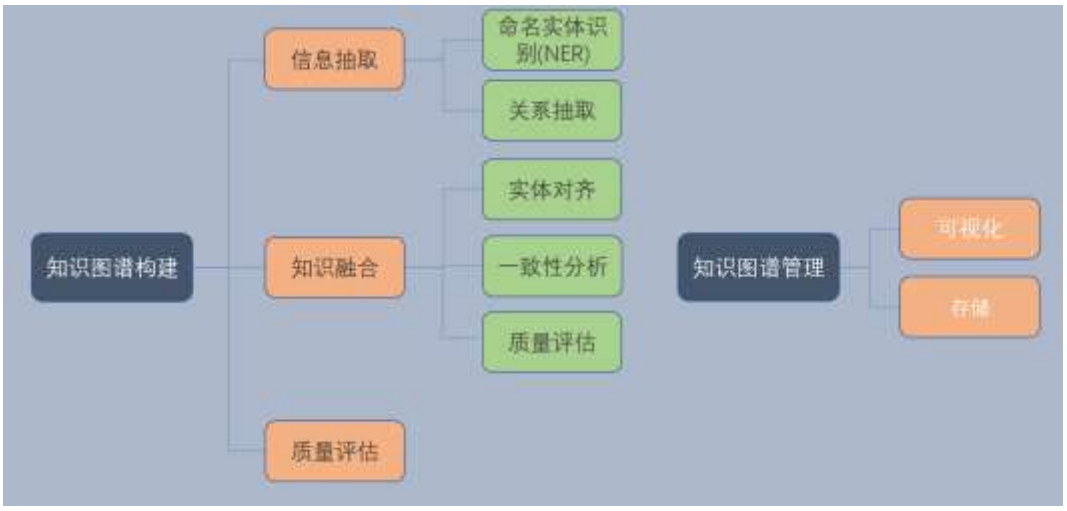
(2)我们团队已经具备了不错的编程能力，熟练掌握了 python、Java、PHP、C、C++、HTML 等编程语言，学会使用 python 爬取所需数据，也掌握了数据库的基本使用，并已经搭建了自己的网站。

(3)通过寒假的准备我们对知识图谱构建、存储和可视化以及基于知识图谱的语义搜索原理有了基本的了解。并能够快速学习后续所需专业知识。

五、项目实施方案（研究思路和方法，实施计划、技术路线、人员分工等）

1.研究思路及实施计划

(1)知识图谱构建与管理



知识图谱构建与管理

知识图谱构建流程：

对医疗数据库数据利用 Bootstrapping 思想，通过模式挖掘实体的方法，构造命名实体识别训练集；利用命名实体识别训练集，训练 Bi—LSTM+CRF 命名实体识别模型；对医疗电子病历数据利用 Bootstrapping 思想，通过模式挖掘关系的方法，构造关系抽取训练集；利用关系抽取训练集，训练 CNN 关系抽取模型；医疗电子病历数据经过实体识别和关系抽取，得到大量实体关系，经过知识融合，形成知识图谱；将知识图谱存储在 Neo4j 数据库中。

知识图谱的存储与可视化：

①在知识图谱的存储方面，使用图数据库，综合比较了各大常用图数据库后，选择了速度较快，扩展性好，稳定性好的 Neo4j 数据库；

②在知识图谱的可视化方面，综合比较各大主流可视化软件，采用了开源的，性能较好的 Gephi 可视化工具

(2)基于知识图谱的语义搜索系统

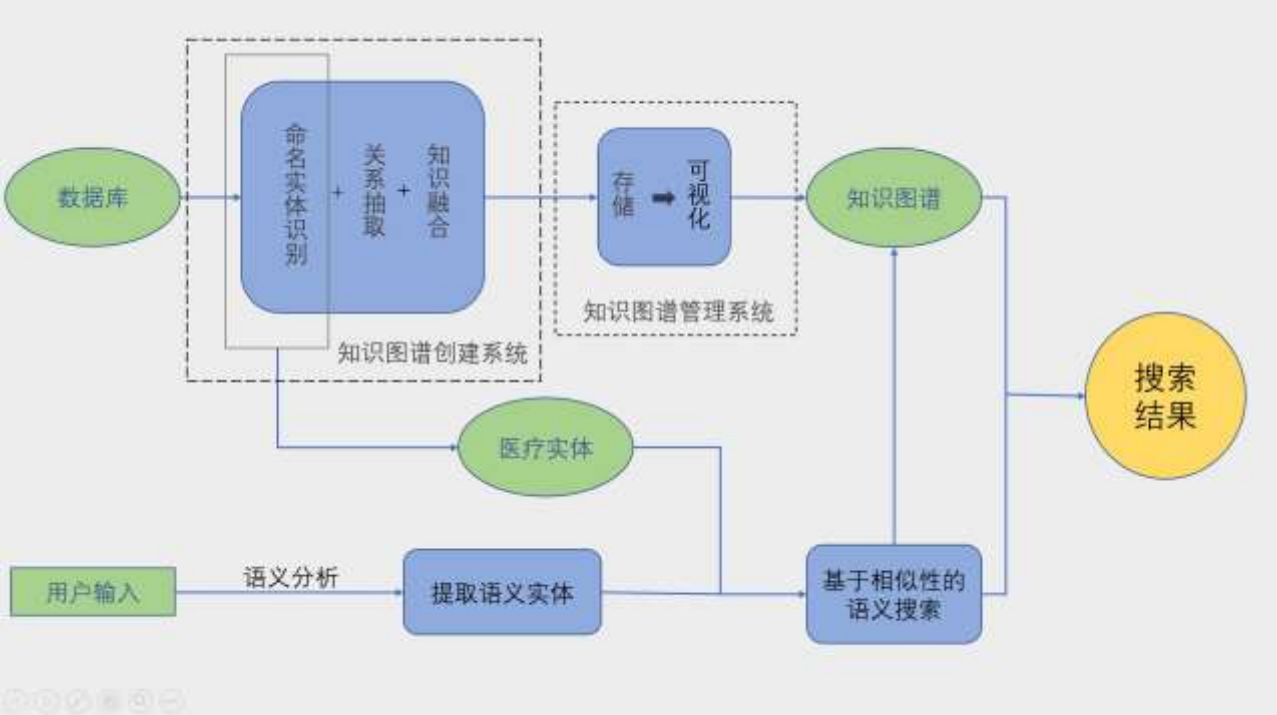
实现语义搜索的功能的具体流程如下；

① 利用知识图谱构建模型的算法对用户输入语句进行分析，提取出医疗实体；

② 对提取出的每个实体进行权重赋值，如果某个实体与越多的其他实体相关联，则这个实体越的权重越高；

③ 在知识图谱中检索与提取出的全部实体的权重距离最近的实体，并将其作为结果返回。

2. 技术路线



技术路线图

3. 人员分工

①高明龙负责算法推导与数据爬取；

②李超负责知识图谱构建、存储和可视化；

③唐思源负责匹配语义搜索和咨询；

④邓新柯负责服务器的搭建与数据库的构建和维护；

⑤全体共同撰写研究报告与学术论文。

注：以上任务分工有主负责人，但具体工作均由四人共同完成

六、项目进度安排（文献查阅、社会调查、方案设计、开题报告、实验研究、数据处理与分析、研制开发、填写结题表、撰写论文和研究报告、结题答辩和成果推广等时间安排）

1.2019 年 5 月——2019 年 7 月 查阅文献，学习知识图谱构建与管理算法；

2.2019 年 7 月——2019 年 8 月 爬取数据，完成所需知识图谱的构建与管理；

3.2019 年 9 月——2019 年 12 月 实现对常见病的语义搜索功能；

4.2020 年 1 月——2020 年 3 月 获得软件著作权，撰写学术论文；

5.2020 年 4 月——2020 年 5 月 撰写研究报告，学术论文投稿，成果推广。

七、**项目研究所需资源**（实验室、仪器设备、实验材料、资料等）

1.深度学习、知识图谱、语义识别、服务器数据库搭建相关文献资料；

2.高性能计算机；

3.计算机楼实验室。

八、**项目经费预算与用途**（购置实验消耗材料、低值品、资料、加工测试、打字复印、调研、市内公交、论文发表、专利申请等经费开支）

序号	名称	总计
1	设备仪器购置费	1000
2	论文版面费	4000
3	软件著作权申请费	500
4	打印复印费	1000
5	实验材料费	1500
	总计	10000

九、**项目完成预期成果**（成果形式：研究论文、专利、设计、产品、软件、研究或调研报告等）

1.学术论文一篇

2.软件著作权一项

3.研究报告一份

十、项目诚信承诺

本项目负责人和全体成员郑重承诺：该项目研究不抄袭他人成果，不弄虚作假，按项目研究进度保质保量完成各项研究任务。

项目负责人签名：　　年　月　日

项目组成员签名：

年　月　日

十一、指导教师意见（从项目科学性、前沿性、可行性、研究性、可操作性和成效性进行评价，是否同意立项）

签 名：

年　月　日

十二、学院评审意见（学术价值、预期效果、研究方案可行性、是否同意立项）

工作组组长签名（公章）：

年　月　日

十三、学校意见

年　月　日