

基于知识图谱的中国近代史知识 问答系统构建研究^{*}

赵浩宇 陈登建 曾桢 张虹雨
(贵州财经大学信息学院, 贵阳 550025)

摘要: 历史文献内容通常以线性文本的形式供用户浏览阅读, 但当用户意在获取非线性的碎片化信息时, 由于普通搜索引擎缺乏理解用户检索意图, 且传统的信息组织方式也无法对历史知识元素进行细粒度的语义关联, 因而无法完成意图理解式的智能搜索, 故本文拟搭建基于知识图谱的中国近代史知识问答系统, 以打破传统信息获取方式的局限性。首先从多个维度搭建中国近代史本体模型, 并通过知识获取、知识抽取、知识融合、知识存储等方式完成知识图谱的实例构建, 然后在此基础上遵循MVC模式三层架构设计实现了一个基于知识图谱的前后端分离的中国近代史知识问答系统, 最后优化了知识问答系统中的语句解析任务, 并对问答系统的性能进行测试, 测试结果取得较好效果, 表明该系统能够较为准确地回复自然语言提问。

关键词: 问答系统; 知识图谱; 领域本体; 中国近代史

中图分类号: G250 **DOI:** 10.3772/j.issn.1673-2286.2022.06.005

引文格式: 赵浩宇, 陈登建, 曾桢, 等. 基于知识图谱的中国近代史知识问答系统构建研究[J]. 数字图书馆论坛, 2022 (6): 31-38.

历史文献资源中蕴含着巨大的价值, 具体体现在历史文献资源既服务于当代社会经济建设, 又能促进人们的精神文化建设。当前, 因得益于信息技术的大规模应用, 文献馆藏机构中浩如烟海的纸质资源也已基本完成数字化转型升级, 但由于大量经过数字化的历史文献资源是以非结构化的形式记载, 以平面化的方式供用户线性浏览阅读, 而计算机无法读懂文本格式的非结构化数据, 也无法以更细粒度的知识元进行语义关联, 检索结果难以展示实体之间的隐含关系, 一定程度上阻碍了历史文献资源的深层次开发与利用。可见, 传统知识获取方式有其固有弊端, 既不利于用户对其感兴趣的历史信息进行语义检索, 也不利于对历史知识脉络进行宏观把握和深层了解。此外, 研究中国近代史需要厘清历史人物之间错综复杂的关系, 发现人物之间的隐性关系, 探讨历史事件的发展脉络和其中的历史缘由, 深入挖掘潜在的历史意义和学术价值。因此, 从

历史文献资源的开发利用入手, 以本体和知识图谱等语义网技术为基础, 以知识问答系统平台为载体, 以实现用户意图检索为导向, 构建基于知识图谱的中国近代史知识问答系统, 用户可以用自然语言进行提问直接获取所需答案, 无须把大量时间和精力浪费在历史文献的检索、浏览和筛选中, 为用户提供便利的中国近代史知识查询服务, 同时还可以通过已构建的知识图谱可视化展示前端平台, 实现历史知识要素的可视化浏览, 帮助用户以直观形式厘清历史知识的关联。

1 相关研究回顾

知识图谱是通过存储实体和关系来实现语义检索的图数据库, 其本质是一种大规模的语义网络图。知识图谱和语义技术为事物的分类、属性和关系的描述提供了新方法^[1], 使得搜索引擎可以直接对事物进行索引

^{*} 本研究得到国家自然科学基金项目“基于知识图谱的农产品价值链信息融合研究”(编号: 2020XSXM)资助。

和搜索,因此基于知识图谱的智能问答系统是一种新型的信息检索方式,本质上是搜索引擎的高级形态^[2],现在已经广泛应用于各个领域的学术研究中,并在各领域大放异彩。智能问答系统的构想最早可追溯到1950年Turing^[3]在其论文中提出的关于机器思维的问题,即机器在多大程度上能够模仿人类并且能够与人类展开互相问答。20世纪60年代,麻省理工学院人工智能实验室的Weizenbaum^[4]设计了聊天机器人ELIZA,通过模式匹配和替换实现人机对话,并且可以模拟特定角色通过聊天的方式与人类展开互动。20世纪70年代,随着知识表示和自然语言处理技术的发展,以及知识库构建技术的成熟,促进了问答系统相关研究的进一步发展,如耶鲁大学在1975年开发的SAM系统^[5],它能够使用脚本来理解所提的问题。在随后的几十年里,相关研究成果仍停留在机器与人类的简单交流上,并未实现真正意义上的智能问答。近年来,随着芯片技术的高速发展,计算机的算力得到空前提升,加之自然语言处理技术的成熟,智能问答技术从特定领域小规模数据库转向基于大规模文本数据的互联网开放领域,知识问答系统走入大数据时代。2011年*Nature*杂志上刊登了一篇关于问答系统的文章中指出:“以直接而准确的方式回答用户自然语言提问的自动问答系统将构成下一代搜索引擎的基本形态”^[6]。由此可知,下一代搜索引擎的发展方向是向用户返回精准、有效信息的问答系统,并且通过自然语言完成人机交互将成为知识获取的新趋势。

近年来,随着信息技术的高速发展,数字人文领域的研究方法和思路不断拓宽,基于知识图谱的知识组织方法被广泛应用于知识发现、智能问答和个性化推荐中,已成为知识服务的重要手段。目前,基于知识图谱的知识问答系统主要有基于模板匹配、基于语义解析、基于深度学习3种实现方法。①Tunstall-Pedoe^[7]早在2010年就率先提出了基于模板匹配的方法,其核心思想是将用户提问先转换为人为预定义的规则或模板,再转换为可执行的查询操作。其优点是匹配响应速度较快、精确度较高,缺点是大量模板维护需要人工完成。例如:丁斌^[8]采用了模板库的方式实现了汽车领域问答系统的构建;李贺等^[9]结合AC多模式匹配算法实现对用户问题的匹配,构建了疾病知识图谱的自动问答系统。②基于语义解析的方法,关键在于对提问语句成分进行解析,并将查询转化成固定的逻辑表达式,再利用知识图谱的语义信息将逻辑表达式转换成知识图谱查询,最后以自然语言形式返回查询结果。其优点是可

解释性较强,但缺乏一定的通用性。例如:高劲松等^[10]在构建馆藏文物资源关联数据知识模型的基础上,提出基于多粒度语义查询的智能问答服务框架;单良等^[11]通过解析自然语言语义信息,构建了中国历史人物知识的智能问答系统。③基于深度学习的方法,关键是将用户的问题投射到一个高维向量空间,获得相应词向量,通过深度学习模型对向量进行相似度计算,再通过相应打分机制获得候选项排序,最后将相似度最大的候选项作为答案返回给用户。其优点是对复杂问题回答的正确率较高,缺点是模型训练成本较高,可解释性较差。如朱建楠等^[12]利用深度学习算法构建了机械智能制造知识问答系统,姜雨娇^[13]探讨了基于深度学习的苹果生产知识自动问答方法。

当前,基于知识图谱的智能问答系统相关研究和实践应用日益成为学界关注的焦点,并且已有学者在中国历史领域开展了知识组织服务和智能问答系统的构建与应用。例如:肖大军^[14]在改进表示学习中翻译模型的基础上,设计了基于知识图谱的中国历史人物亲属关系自动问答系统;周亦等^[15]依托知识图谱等可视化技术,实现了中国古代历史人物之间的复杂关联并对其进行可视化呈现;张云中等^[16]以红色历史人物数字资源为核心,以知识图谱和KBQA为框架构建了红色历史人物知识问答模型,并通过实证研究证明了问答系统的可行性;王颖等^[17]依托国史本体知识库构建了国史知识检索平台,实现了知识检索、智能问答等深度检索服务;Liu等^[18]探讨了基于问答数据库与知识图谱结合的方法,构建辽代历史文化领域智能问答系统。以上相关学者的研究从多方面阐述了基于知识图谱的知识问答系统构建方法与思路,对本文所构建的问答系统具有借鉴意义。

目前还没有针对中国近代这一特定历史时期全领域范畴的知识问答系统,部分研究只聚焦于某一特定主题,或是某一较窄时间段,而且现有的垂直领域问答项目系统框架可移植性较差。此外,大部分研究实现了语义检索和知识图谱可视化,但由于模式层本体构建不完善,未能有效规范数据层中的实例对象,并且缺乏多维细粒度的语义信息,因此导致历史领域知识解释性较差、系统性较弱、关联性不足等问题,不利于对中国近代史内容的关联挖掘和细粒度知识元的语义检索。本文以中国近代时期为经、以历史文献内容为纬,结合数字人文技术,构建基于知识图谱的知识问答系统,其优点在于可以实现历史文献内容中细粒度知识重组与语义化关联等知识组织服务,即从历史人物、事件、机

构、时间地点等多个维度出发,搭建细粒度的知识模型并使其进行语义化关联。

2 知识问答系统的总体设计

中国近代史知识问答系统采用MVC (Model View Controller)^[19]模式进行开发,其整体架构(见图1)分为展示层、逻辑层和数据层,其中数据层包括数据来源、数据处理及知识图谱构建,逻辑层是知识问答系统设计模块,展示层是对问答系统的具体应用。系统分层

的目的是将系统中各部分分离,以降低不同部分之间的耦合度,以提高系统模型的可移植性。

数据层是问答系统的前提和基础,采用自顶向下的方法构建本体模式层,当有新知识需要加入时,则采用自底向上的方法更新知识图谱,并选择Neo4j作为知识图谱数据仓储,将多源异构的数据以图结构的方式进行存储,既能直观展现各种信息之间的关系,又能提升知识检索效率。知识图谱中存在大量具有语义关联的数据使得该系统理解用户提问成为可能。

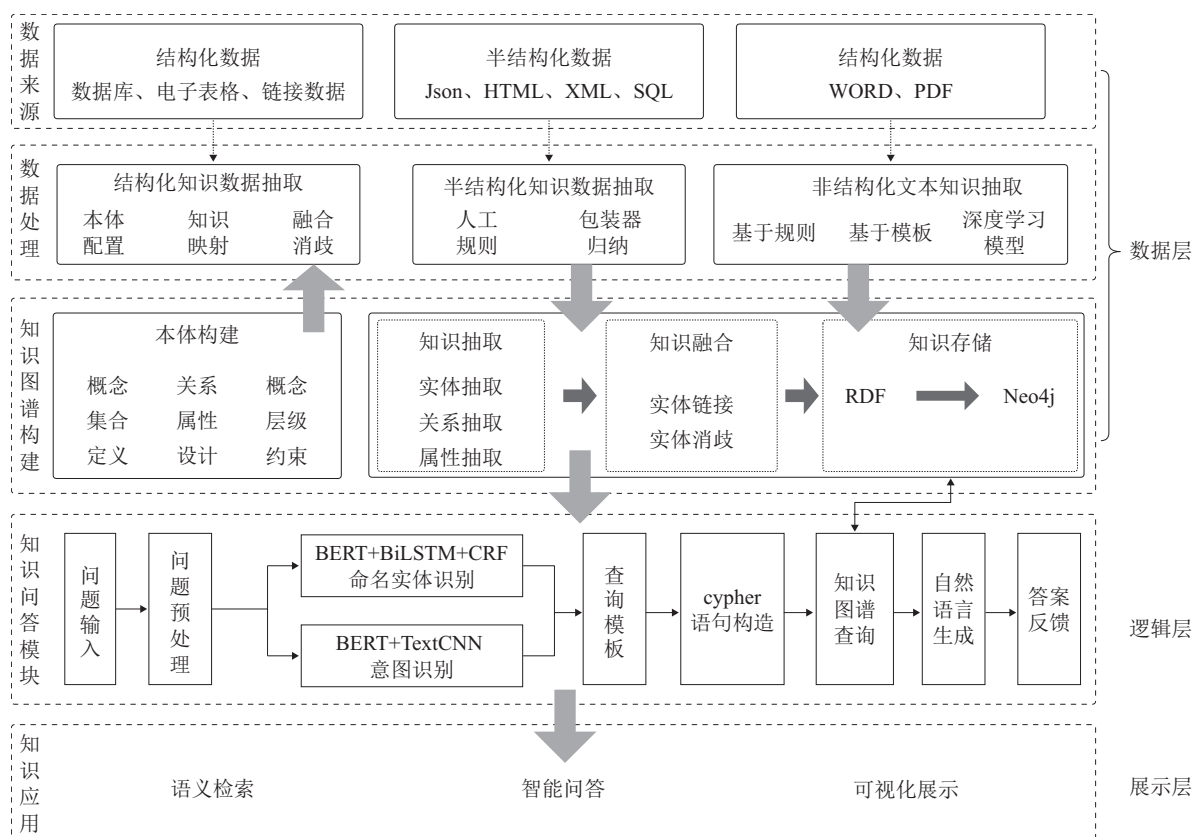


图1 中国近代史知识问答系统的整体架构

逻辑层是整个问答系统的核心,该层主要负责问题解析和问题生成两部分。问题解析是对用户的自然语言提问进行处理和识别,首先是对用户的输入进行相关预处理;其次根据BERT+BiLSTM+CRF算法对问句中的实体进行识别,借助BERT+TextCNN算法完成用户意图识别工作;最后根据cypher语法生成相对应的查询句式,再对Neo4j数据库进行查询并以通俗易懂的语句将答案返回给用户。

展示层的工作原理是基于前端技术生成的网页服

务项目,用户交互是其主要功能。本文构建的基于知识图谱前后端分离的知识问答系统,利用Flask框架构建前端平台,Neo4j作为后端数据仓储,使用Python语言完成对图数据库的连接及操作,最终完成与用户的问答交互服务。

综上,基于知识图谱的中国近代史知识问答系统的实现由三大模块和两大流程构成,三大模块即上述MVC三层开发模式,两大流程包括中国近代史知识图谱的构建流程以及知识问答系统的实现流程。

3 中国近代史知识图谱构建

3.1 本体层构建

知识图谱在逻辑上可分为模式层和实例层,模式层是知识图谱的“骨骼框架”,是对实例层的约束和规范,也是完成知识图谱构建的中心内容。本文为提高本体模型的质量和后期知识问答实例数据的准确性,通过文献调研深入剖析文本内容特征,并结合领域专家知识设计相关概念、关系及属性,运用工具Protégé构建中国近代史本体,完成知识建模。具体而言,首先确定

中国近代史核心概念集,其次确定概念间的层次关系,再次定义类的对象属性和数据属性以及属性约束。

中国近代史本体层包括“历史人物”“历史事件”“历史文献”“地点”“时间实体”“历史时期”“思想理念”“领域”“行为主体”“组织机构”10个核心概念。中国近代史本体模型共有10个一级类、53个二级类、88个三级类,以及包括数据属性和对象属性在内的95个属性约束,并根据近代史历史知识定义了一些属性约束和推理规则,从而确定了知识实体的分类以及实体关系类型。中国近代史本体核心概念关系模型部分展示,如图2所示。

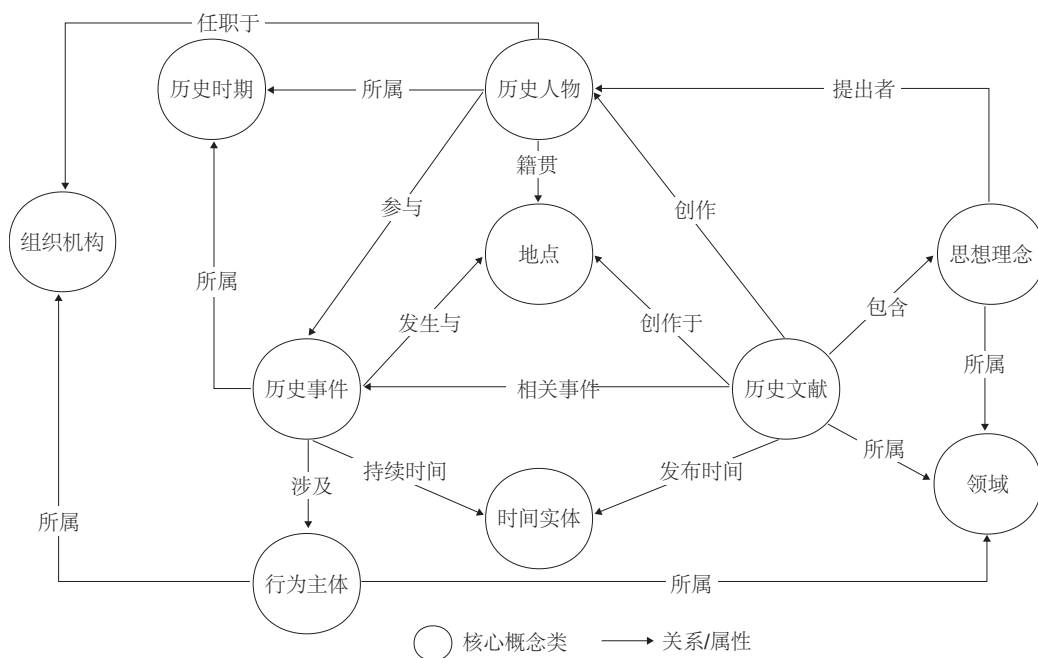


图2 中国近代史本体核心概念关系模型展示 (部分)

3.2 知识获取

系统所需的数据来自开放数据集、网络论坛、垂直站点、国家图书馆人物专题数据库、电子化的文本资料、维基百科以及百度百科等,通过网络爬虫、人工筛选、自然语言处理等方式从多种数据源中获取相关知识,并将其转换为结构化的数据,作为构建知识图谱实例层所需要的资源。

3.3 知识抽取

知识抽取是根据已定义好的本体模型,完成所需

相关实体的抽取工作。本研究根据数据来源的不同,通过网络爬虫、模式匹配、包装器适配、机器学习等方法完成实体识别、信息抽取等任务,抽取所需要的实体、关系和属性,并将其转换成实体关系三元组。知识抽取具体方法和步骤见参考文献[20]。

3.4 知识融合

抽取后的知识存在大量重复、同名同义的实体,需要进行实体消歧和实体对齐等知识融合的方法来对数据进行整合。实体消歧的主要任务是把有歧义的命名实体映射到实际所指的实体概念上。实体对齐是指对

具有相同指称的不同实体进行相互链接的操作。本研究通过计算相似度的方式消除歧义,如实体“辛亥革命”可能指的是作品类别中的《辛亥革命》著作,也可能指“辛亥革命”这一历史事件。解决此类问题的方法是,先找到“辛亥革命”所有解释的描述,由于上文通过知识获取从各种信息源中爬取了实体的解释内容,因而方便转换成向量表示,再将文本中包含“辛亥革命”的句子转成向量,将这个向量和所有解释的向量进行相似度计算,选择相似度最高的,以此完成消歧工作。历史人物的属性如“出生年月”“出生日期”“出生时间”等,这些属性说法不同但都表示相同含义,则采用基于词典匹配的属性对齐方法。历史文献中某一地名的称呼可能随着时间的不同而发生改变,如“奉天府”是今“辽宁省沈阳市”的旧称。为了使历史文献数据中的地名和当代的行政区划名称统一,本文利用基于Python的CPCA (Chinese Province City Area) 开源工具包和中国国家省市行政区划数据库作为参照标准,然后将已抽取的中国近代史相关地名与开放数据集进行实体链接,完成实体对齐工作。

3.5 知识存储

图数据库主要用于存储更多相互关联的数据,图结构相比其他数据结构而言,能保存更多数据间的关系,能高效率地处理非结构化等复杂数据,而且图数据库的维护相对容易,还可以即时呈现出图谱效果。

这一环节将上文抽取得到的实例知识进行汇总处理,选择Neo4j图数据库完成知识存储。首先使用cypher命令语句LOAD CSV将实体和属性导入图数据库Neo4j中,然后再将对应的关系导入其中。由于本文获取的人际关系数量较多,所以选择采用Python工具包py2neo将其导入Neo4j中。知识存储完毕后,便可进行知识图谱的可视化呈现,由于Neo4j自带的展示效果可自定义选项较少,故选择使用Neovis.js连接Neo4j数据库,将JavaScript可视化和Neo4j无缝对接,可以灵活地为标签、属性、节点和关系进行样式和色彩的自定义设计。本研究总共构建了包括历史人物、事件、机构等在内的11 768个实体节点以及16 592个三元组关系,中国近代史知识图谱可视化(部分)展示效果见图3。

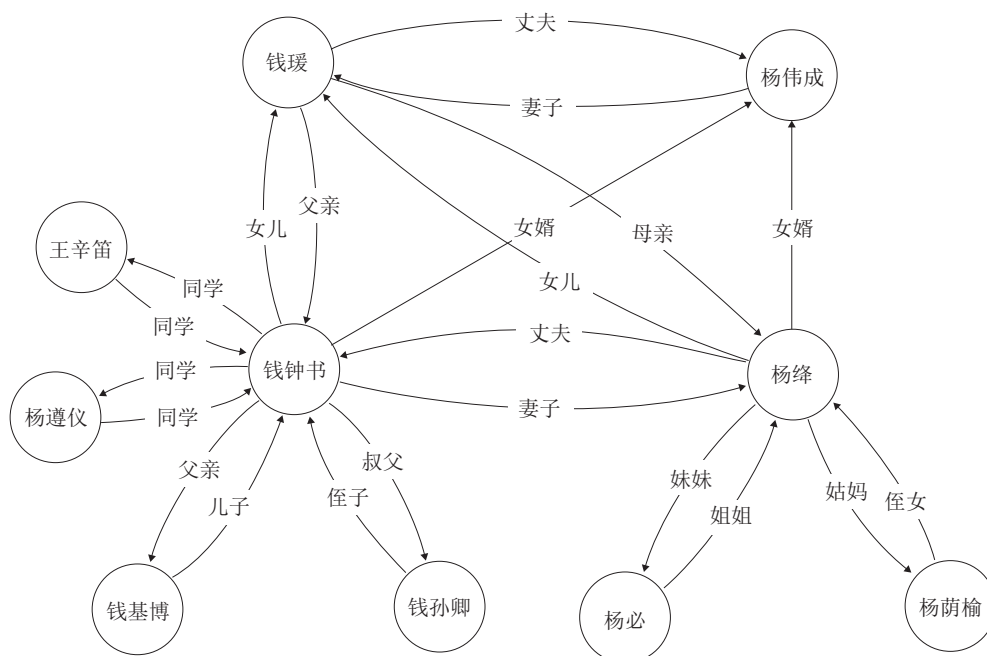


图3 中国近代史知识图谱可视化(部分)展示

4 基于知识图谱的知识问答系统构建

4.1 知识问答系统架构

中国近代史知识问答系统主要完成的功能有接受

用户的问题、对用户的问题识别分析、检索出最合适的答案、将问题的答案反馈给用户,知识问答系统的架构如图4所示。

用户通过系统接口输入相应的问题语句,首先借助分词工具进行语句的分词处理、去除停用词处理等

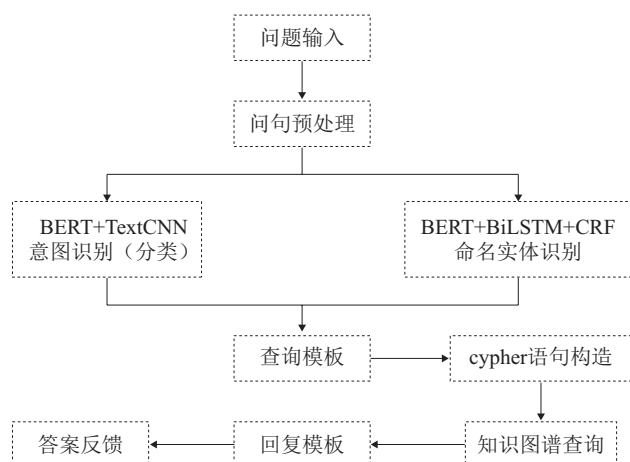


图4 中国近代史知识问答系统架构图

操作。本文预先训练相应的BERT+TextCNN模型，对用户意图类别进行分类编码，将上文预处理得到的问题语句输入模型，获取用户的意图类别标签，借助BERT+BiLSTM+CRF模型抽取出文本中的实体数据，将用户意图类型与实体数据传入预先定义的查询模板，通过模板匹配，构建出符合用户查询要求的cypher语法并借助py2neo库连接到知识图谱获取返回结果，最后将返回数据与回复模板进行匹配，输出答案反馈。预先构建的回复模板是根据人类自然语言使用习惯设计而来，意在提升用户的整体使用体验。

4.2 意图识别

本研究将用户问题划分为人物信息类、人物关系类、事件信息类、著作信息类、人物经历类和地理信息类6个类别。其中，人物信息类中主要定义人物基本属性信息、头衔、职位、信仰、毕业院校等，人物关系类将人物实体之间的关系分为父子关系、夫妻关系、朋友关系、师生关系、同学关系等，事件信息类主要包含事件的基本属性信息与事件简介等问题，著作信息类主要包含著作的属性信息与相关人物等，人物经历类主要包含相关人物实体的历史经历，地理信息类主要包含人物、事件等涉及的地理方位信息。

传统基于模板的问题分类模型存在许多固有弊端，如用户提问方式的词汇不在模板之中，就无法正确识别用户的意图。本文首先采用BERT+TextCNN文本分类模型实现对于用户意图的识别，借助BERT文本预训练模型，将用户提问数据进行动态向量化的表达，将学习到的词向量特征，输入到TextCNN模型之中。

TextCNN是借助不同的卷积核实现对于不同位置序列的句子特征信息的提取；然后通过最大值的池化操作提取文本特征；最后将获取到的不同卷积核下的特征进行拼接，接上Softmax层，完成对用户意图的识别，输入用户意图的类别标签。

为验证模型的分类效果，笔者将互联网收集的中学生近代史问答竞赛题数据与自行构建的近代史问题数据进行汇总，将不同的问题进行分类编码，汇总为训练集、验证集和测试集。借助训练集训练模型性能，验证集用于调整模型参数，测试集用于评估模型的性能。经验证，准确率为92.75%，召回率为93.05%，F1值为92.89%，模型整体效果良好，可以实现对于用户的意图识别。

4.3 实体识别

由于开源项目Jiagu工具包对本文所需的历史实体抽取精度尚不理想，于是本文借助主流的BERT+BiLSTM+CRF训练近代史领域的命名实体识别模型，完成对于输入问句中的实体数据进行识别。该模型是由BERT嵌入层、BiLSTM双向语义特征提取层、CRF实体序列信息标注层搭建而成。首先将预处理完成的问题语句传入BERT层，在该层中将问句中的字符转化为相应的词向量表达。由于BERT模型中内嵌Attention机制，可以完成对于字符级别下的权重调整，因此能更好地捕捉字符的信息特征。然后将相应的字符向量传入BiLSTM层，该层是将正向的LSTM与反向的LSTM层堆叠而成，既可以获取正向的字符序列信息，又可以获取反向的字符序列信息。借助BiLSTM获取句子中字符所对应的各个标签的得分矩阵。最后，将结果输入到CRF层中，根据相邻标签中的相关性，调节相应的字符标签序列，保证最终的预测结果符合预先定义的标注方案。

中国近代史领域命名实体识别模型的训练，首先借助前期Jiagu库对非结构化数据中的人物、历史事件、地理位置和机构实体打上标签，结合人工筛选、矫正实体标注，为所有的文本打上相应的序列标注信息。其中本文选用BIO（B-begin, I-inside, O-outside）序列标注方案对于人物、历史事件、地理名称和机构等非结构化数据实体，分别采用不同的BI标签进行区分，非实体数据用O表示，实体开头用B表示，实体后续用I表示，完成初期标注工作。其次，按照句号作为分隔符号，对非结构化数据进行划分，借助BERT+BiLSTM+CRF模

型对非结构化的命名实体进行识别。最后,对经过训练后模型进行评估后,发现BERT+BiLSTM+CRF模型的准确率为93.1%,召回率为92.35%,F1值为92.72%,实体识别效果精度良好。

4.4 查询语句构建

借助上述的模型识别出用户输入的核心信息,程序将该信息与查询模板进行匹配判断,通过Python语言中的占位符传入字符参数,生成预先定义好的cypher查询语句。在6类不同的用户意图类别下分别搭建相应的查询语句模板。如“我想问一下李鸿章和张爱玲是什么关系?”,系统识别出相应的用户意图与实体数据,构建出相应的cypher查询语句, MATCH (n:历史人物{name:"李鸿章"}) -[r]-> (c:历史人物{name:"张爱玲"})。

4.5 构建回复语句

现阶段对于回复模板的构建方式主要有借助模板与深度学习两种方法,其中深度学习的方法主要借助的是Seq2Seq以及其变种的Attention等方法,通过Encode和Decode的编码方法,实现多样化的回答,但是该方法生成的回答语句还不够严谨,综合考虑各方因素,本文选择借助模板生成自然语言的回复语句,就是借助py2neo实现脚本程序与Neo4j的数据库连接,实现数据的交互。首先将生成的cypher查询语句借助py2neo传入图数据库,在知识图谱中实现属性查询、关系查询和实体查询等功能,再将图数据库返回的数据传入预先定义的答案回复模板,最后把通俗易懂的答案反馈给用户。

5 知识问答系统测试及结果分析

用户与问答系统的交互是该系统开发的主要功能。该系统不仅能回答如蒋介石有关的事件、蔡元培的生平简介、浙江籍贯的历史名人和《定军山》历史作品的演员等直接信息,而且支持复杂推理信息查询以及多问句问题等复杂查询,如李鸿章和张爱玲的关系,历史事件发生的时间、空间以及影响等信息的查询。

为检验中国近代史知识问答系统的准确性,本研究选择7位在校大学生作为志愿者,让每个人针对不同问题类别提出10条问题,一共得到420条问题数据。借

鉴以往学者的经验,本文选取ACC作为整体的问答系统性能的评价指标^[21]。ACC是问答系统回答的准确率,表示为系统回答正确问题数量与总测试语句数量的比值。经测试统计可知,问答系统准确率的均值为94%,大多数问题可以被模型精准地识别并返回有效的答案语句,但回答人物信息类、事件信息类与人物经历类问题的准确度稍低。经回溯分析发现,系统对于人物信息类与人物经历类问题会出现类别判断错误的现象,是因为用户提问这两种类型的信息时所用的自然语言表述语句十分相似,因此问答系统容易错误归类,导致判断错误,这些问题将在后续的研究中进一步完善。

6 结语

本研究构建了基于知识图谱的智能问答系统,不仅能解决用户的语义检索问题,而且能够通过知识图谱可视化方式对历史领域碎片化的知识进行直观呈现,帮助用户节省大量时间和精力,直接获取所需信息,使其从线性阅读文档的体力劳动中获得解放。该项目的主要特点如下。①通过网页平台,实现智能交互。依托知识图谱技术,实现了从网页文档链接向实体链接转变,支持用户按主题和意图,而不是按字符串检索,从而实现真正意义上的语义检索。②以知识图谱为数据仓储,展现历史各要素之间的复杂关联。本系统实现了历史人物、历史事件各实体属性的细粒度知识关联,为历史爱好者提供便利。③丰富人文科学研究手段,扩展研究思路。本文以多学科交叉视角,从历史文献资源的开发利用入手,以知识问答系统构建方案为主要内容,以具体的实例样本进行验证和实现,构建了较为完备的跨学科研究思路,能够为相关的学科交叉研究提供借鉴和参考。

当前,知识问答系统正在朝着基于深度学习的方法靠拢,解决多条推理问题以及提高模型的可解释性仍是当前知识问答系统发展的趋势和挑战。未来还需提高知识图谱中数据的质量,扩充知识图谱的范围以提高问答知识的覆盖率,提高对复杂问题的理解能力,打造更智能的中国近代史知识问答系统。

参考文献

- [1] 王昊奋,漆桂林,陈华钧. 知识图谱:方法,实践与应用[J]. 自动化博览,2020,37(1):7.

- [2] 王智悦, 于清, 王楠, 等. 基于知识图谱的智能问答研究综述 [J]. 计算机工程与应用, 2020, 56 (23): 1-11.
- [3] TURING A M. Computing machinery and intelligence [J]. Mind, 1950, 59 (236): 433-460.
- [4] WEIZENBAUM J. ELIZA-a computer program for the study of natural language communication between man and machine [J]. Communications of the ACM, 1966, 9 (1): 36-45.
- [5] SCHANK R C, ABELSON R P. Scripts, plans, and knowledge [C] // Proceedings of the 4th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc., 1975. 151-157.
- [6] ETZIONI O. Search needs a shake-up [J]. Nature, 2011, 476 (7358): 25-26.
- [7] TUNSTALL-PEDOE W. True knowledge: Open-domain question answering using structured knowledge and inference [J]. AI Magazine, 2010, 31 (3): 80-92.
- [8] 丁斌. 汽车领域智能问答系统中模板库自动生成方法的研究 [D]. 上海: 上海交通大学, 2017.
- [9] 李贺, 刘嘉宇, 李世钰, 等. 基于疾病知识图谱的自动问答系统优化研究 [J]. 数据分析与知识发现, 2021, 5 (5): 115-126.
- [10] 高劲松, 方晓印, 刘思洋, 等. 基于关联数据的馆藏文物资源知识关联与智能问答研究 [J]. 情报科学, 2021, 39 (5): 12-20.
- [11] 单良, 刘欣. 基于中国历史人物知识的智能问答系统构建 [J]. 情报探索, 2019 (6): 101-105.
- [12] 朱建楠, 梁玉琦, 顾复, 等. 基于深度学习的机械智能制造知识问答系统设计 [J]. 计算机集成制造系统, 2019, 25 (5): 1161-1168.
- [13] 姜雨娇. 基于深度学习的苹果生产知识自动问答研究 [D]. 咸阳: 西北农林科技大学, 2021.
- [14] 肖大军. 中文历史人物亲属关系自动问答研究 [D]. 武汉: 华中师范大学, 2018.
- [15] 周亦, 周明全, 王学松, 等. 大数据环境下历史人物知识图谱构建与实现 [J]. 系统仿真学报, 2016, 28 (10): 2560-2566.
- [16] 张云中, 郭冬, 王亚鸽, 等. 基于知识图谱的红色历史人物知识问答服务框架研究 [J]. 图书情报工作, 2021, 65 (16): 108-117.
- [17] 王颖, 张智雄, 孙辉, 等. 基于本体的国史知识检索平台构建研究 [J]. 图书情报工作, 2015, 59 (16): 119-128.
- [18] LIU S, TAN N, YANG H, et al. An Intelligent Question Answering System of the Liao Dynasty based on knowledge graph [J]. International Journal of Computational Intelligence Systems, 2021, 14 (1): 1-12.
- [19] 王曼维, 杨获, 李岩, 等. 基于SpringBoot框架的智慧医疗问诊系统设计与实现 [J]. 中国医学装备, 2022, 19 (3): 133-136.
- [20] 曾桢, 赵浩宇. 基于文献的中国近代史知识图谱构建与实证研究 [J]. 数字图书馆论坛, 2022 (4): 35-42.
- [21] 陈璟浩, 曾桢, 李纲. 基于知识图谱的“一带一路”投资问答系统构建 [J]. 图书情报工作, 2020, 64 (12): 95-105.

作者简介

赵浩宇, 男, 1996年生, 硕士研究生, 研究方向: 信息资源管理、知识图谱, E-mail: galaxyeric582@163.com.

陈登建, 男, 1995年生, 硕士研究生, 研究方向: 数据挖掘、机器学习。

曾桢, 男, 1982年生, 博士, 教授, 研究方向: 信息资源管理、知识图谱。

张虹雨, 女, 1998年生, 硕士研究生, 研究方向: 信息资源管理。

Research on the Construction of Knowledge Q&A System of Modern Chinese History Based on Knowledge Graph

ZHAO HaoYu CHEN DengJian ZENG Zhen ZHANG HongYu

(Information School, Financial and Economics of Guizhou University, Guiyang 550025, P. R. China)

Abstract: The content of historical documents is usually in the form of linear text for users to browse and read. However, when users intend to obtain non-linear fragmented information, the general search engine lacks an understanding of users' retrieval intention, and the traditional information organization method cannot make fine-grained semantic associations with historical knowledge elements, so it is impossible to complete the intelligent search with intention understanding. Therefore, this paper plans to build a knowledge question-and-answer system of modern Chinese history based on a knowledge graph to break the limitations of traditional information acquisition methods. In this paper, we build an ontology model of modern Chinese history from multiple dimensions and complete the instance construction of a knowledge graph through knowledge acquisition, knowledge extraction, knowledge fusion, knowledge storage, etc. On this basis, we follow the three-layer architecture design of the MVC pattern to realize a knowledge question and answer system based on a knowledge graph with front and back-end separation. Finally, the project optimized the utterance parsing task in the knowledge Q&A system and conducted experimental tests on the performance of the Q&A system, which achieved good results and showed that the system can respond to natural language questions more accurately.

Keywords: Q&A System; Knowledge Graph; Domain Ontology; Modern Chinese History

(收稿日期: 2022-06-04)