学士学位论文

基于知识图谱的智能问答系统研究

学　　号：　20171000724
姓　　名：　杨洋
学 科 专 业：　软件工程
指 导 教 师：　袁国斌 教授
培 养 单 位：　地理与信息工程学院

二〇二一年五月

中国地质大学（武汉）学士学位论文原创性声明

本人郑重声明：本人所呈交的学士学位论文《基于知识图谱的智能问答系统研究》，是本人在指导老师的指导下，在中国地质大学（武汉）攻读学士学位期间独立进行研究工作所取得的成果。论文中除已注明部分外不包含他人已发表或撰写过的研究成果，对论文的完成提供过帮助的有关人员已在文中说明并致以谢意。

本人所呈交的学士学位论文没有违反学术道德和学术规范，没有侵权行为，并愿意承担由此而产生的法律责任和法律后果。

学位论文作者签名：

日 期：　　年　月　日

摘要

随着基于专业领域知识图谱的智能问答系统的发展，面向开放领域的智能问答系统因其能充分利用多个领域的相关知识来有效提高系统的普适性、扩展性受到越来越多专家和学者的青睐。基于知识图谱的智能问答系统的研究主要聚焦于两个方面：高质量的知识来源和准确的问题理解能力。各大专业领域或者开放域知识图谱的出现为智能问答系统提供了高可靠高质量的知识来源，同时不断出现和改进的各类人工智能研究方法为处理问答系统对自然语言的理解奠定了基础。本文基于信息检索的方法实现了基于开放域知识图谱的智能问答系统，为此论文做了以下主要工作：

（1）简述知识图谱有关基础理论和预处理操作。首先，本文介绍了知识图谱和构建知识图谱有关的技术。其次，由于本文使用的是 NLPCC-ICCPOL 2016 KBQA 任务提供的开放域知识图谱，数据量庞大，存在不少噪音，为了提高系统的效率和质量，本文对该知识图谱做了降噪处理，并使用 MySQL 数据库对知识进行存储。

（2）实现了有关的问答算法。本文将问答算法划分为命名实体识别和属性链接两个主要子任务。命名实体识别使用了基于 CRF 模型的算法，并利用特征实体进行评分。该算法在《人民日报》语料上的 F1 值为 0.91，在 NLPCC-ICCPOL 2016 KBQA 任务语料上的 F1 值为 0.75。而属性链接使用基于双层 LSTM 模型的算法，并利用了特征对属性进行评分。该算法在 NLPCC-ICCPOL 2016 KBQA 任务语料上的 F1 值为 0.86，表现良好。

（3）设计并实现了智能问答系统。本文基于 BS 架构和三层架构，使用 Vue 和 Flask 等技术实现了开放领域能够简单交互的问答系统，该系统支持用户问题的输入，并能够及时显示处理的结果。

关键词：知识图谱；问答系统；命名实体识别；属性链接

Abstract

With the development of intelligent question answering systems based on professional domain knowledge graphs, intelligent question answering systems oriented to open fields have attracted more and more experts and experts because they can make full use of relevant knowledge in multiple fields to effectively improve the universality and scalability of the system. Scholar's green squint. The research of intelligent question answering system based on knowledge graph mainly focuses on two aspects: high-quality knowledge sources and accurate problem understanding ability. The emergence of major professional fields or open domain knowledge graphs provides intelligent question answering systems with high-reliability and high-quality knowledge sources. At the same time, various artificial intelligence research methods that continue to appear and improve have laid the foundation for processing question answering systems' understanding of natural language. This paper implements an intelligent question answering system based on the open domain knowledge graph based on the method of information retrieval. For this purpose, the paper has done the following main work:

(1) Briefly describe the basic theories and preprocessing operations of the knowledge graph. First of all, this article introduces the knowledge graph and the technology related to the construction of the knowledge graph. Secondly, because this article uses the open domain knowledge map provided by the NLPCC-ICCPOL 2016 KBQA task, the amount of data is huge and there is a lot of noise. In order to improve the efficiency and quality of the system, this article has done noise reduction processing on the knowledge map and used The MySQL database stores knowledge.

(2) The related question and answer algorithm is implemented. This paper divides the question answering algorithm into two main subtasks: named entity recognition and attribute linking. Named entity recognition uses an algorithm based on the CRF model and uses characteristic entities for scoring. The F1 value of this algorithm on the "People's Daily" corpus is 0.91, and the F1 value on the NLPCC-ICCPOL 2016 KBQA task corpus is 0.75. The attribute link uses an algorithm based on the two-layer LSTM model, and uses features to score attributes. The F1 value of this algorithm on the NLPCC-ICCPOL 2016 KBQA task corpus is 0.86, which performs well.

(3) Designed and implemented an intelligent question answering system. Based on the BS architecture and three-tier architecture, this paper uses Vue and Flask to

implement a simple interactive question and answer system in the open field. The system supports the input of user questions and can display the results of processing in time.
Keywords: knowledge graph; question answering system; named entity recognition; attribute link

目录

第一章 绪论

1.1 研究背景和意义

1950 年，艾伦·麦席森·图灵发表了一篇具有划时代重要意义的学术论文《计算机器与智能》[1]，文章准确预测了创建真正智能机器的可能性。同时，由于当时的科学家很难精确定义“智能”的基本概念，因此他提出了著名的图灵测试以让人们相信“思维机器”是可能的。此后，学者们开始陆续探索自然语言中的语义提取等相关问题。同时，随着互联网技术的不断进步和发展，海量的互联网信息不断出现和增加，通常这些信息结构混乱，并且涉及的范围非常广，存在诸多的干扰因素，从中获取有用的信息变得十分困难。为了解决此问题，智能问答系统成为了一大研究热点。

问答系统（Question Answering system, QA system）是指用来回答人类提出的自然语言问题的系统，它的实现已经涉及到包括数据挖掘、自然语言处理等交叉性领域。传统的智能问答系统使用的信息搜索引擎，信息量大，准确率低，难以进行更加精准的分析和判断。现代化的智能问答系统已经能够根据每个问句中的信息，综合分析，反馈得到更加精确的问题答案，而不是传统问答系统反馈的大量信息。当前出现或者运行的智能问答系统主要应用在一些开放技术领域和特定技术领域，在开放领域，智能问答系统可以回答包括社会、文学、新闻等多方面的的问题[2]；在特殊领域，智能问答系统可以依据相关问题模板回答更加专业的问题。同时，为了实现优秀和高质量的问答系统，特别需要关注两个方面的内容，即高质量的知识来源和准确的问题理解。近年来，大数据的快速发展为这两个技术关键点提供了数据层面的新契机。

知识图谱的不断出现为高质量的智能问答系统带来了高质量的知识数据来源。知识图谱，从根本上说，是一种语义化的网络，目的主要是为了解释实体之间存在的关系，显示知识的发展进程和结构关系。外部客观存在的事实构成了信息，在信息的基础上，为了获得构建知识图谱所需的知识，需要对信息的梳理归纳和分析总结，并构建实体之间的关联。“知识工程”相关的概念最早是由费根鲍姆提出[3]，他认为知识中蕴含很大的力量，并且他通过实验证明，研究和实现智能

行为的主要手段是构建知识特别是指定领域专有的知识。

基于知识图谱的智能问答系统，另一个技术关键点和核心难点在于智能问答系统对自然问句的理解。输入的问题是自然语言形式的，而知识图谱的信息却是结构化存储的，同时输入的问题也可能与结构化存储的信息存表述上面的问题。如输入问句为“黄飞鸿的家乡是哪里？”，而知识图谱中存储的三元组为（“黄飞鸿”，“籍贯”，“广东”）。如何在问句中找到主要的实体“黄飞鸿”，如何找到“家乡”与“籍贯”之间存在的联系，是解决这些问题的关键。

1.2 研究内容

虽然知识图谱渐渐完整，有关的数据也逐渐丰富，但是新一代的智能问答系统如果找不到一种较好的方式从知识图谱和问句中获取正确答案，那就无法发挥和体现知识图谱的作用，也就无法实现更高质量的问答系统设计。为此，本文的主要研究内容如下：

（1）简述知识图谱有关基础理论和预处理操作。本文介绍了知识图谱有关知识，并使用 NLPCC-ICCPOL 2016 KBQA 任务提供的知识库构建了知识图谱，并对该知识图谱做了降噪处理。

（2）实现了基于知识图谱的问答算法。本文使用基于信息检索的方法，主要将算法分为命名实体识别和属性链接。特别说明，本文实现的命名实体识别算法和属性链接算法都是基于中文的。在命名实体识别中，本文通过 CRF 模型提取问句中的实体，查询知识图谱构建候选实体集合，根据特征进行评分；在属性映射步骤中，本文通过双层 LSTM 模型计算问句与不同属性的语义相关度，并根据特征进行评分。在最终确认结果之前，结合实体得分和属性得分，综合排序，得到最高分的三元组的属性值，就是问句的答案。。

（3）设计并实现了问答系统。在知识图谱基础上，实现了开放领域可以简单交互的问答系统，系统主要分为前端展示模块、问答逻辑模块和数据构建模块三个模块，支持问题的输入，并能够及时显示处理的结果。

1.3 研究方法

目前，研究智能问答系统对自然语言问句的理解的基本研究方法可以大致分为四类：基于语义分析（Semantic Parsing）的方法，基于信息抽取（Information Extraction）的方法、基于向量建模（Vector Modeling）的方法[4]和基于信息检索

（Information Retrieval）的方法。

基于语义分析的方法，该方法重点在于将自然语言问句转化成为其他类型的逻辑形式，并且要求这种逻辑形式可以表达问句的语义，然后再基于这种结构化的表达式从已有知识图谱中寻找答案，这种方法通常需要使用一些第三方的自然语言处理工具，所以特别容易造成误差。

基于信息抽取的方法，主要使用到了自然语言处理中常用的依存树，按照一定的依存语法，将自然语言解析成为一颗依存树，再从中提取问题中的关键词，这种方法通常需要大量的人工干预，耗费人力物力。

基于向量建模的方法，主要是将问题和答案都映射到低维空间，得到它们的分布式表达，通过训练之后可以预测问题和答案之间的关联得分，对候选答案排序找出最高分作为答案，这种方法在一对多、多对多等情况下还存在不足。

基于信息检索的方法，首先通过粗略的方式从知识图谱中获取一系列的三元组，通过抽取三元组与问句间的关系等方面的特征，对候选答案进行排序，最终选择得分考前的作为最终答案。

综合来看，基于信息检索的方法设计灵活，能够灵活融入多种深度学习或者机器学习模型，也能更好的提取问句中的语义，所以本文采用了基于信息检索的方法来研究本文基于知识图谱的智能问答系统。

1.4 国内外研究现状

2012 年是知识图谱发展最迅速最重要的年度，Google 花重金收购 Metaweb 公司，并向外界正式发布其知识图谱(knowledge Graph，KG)[5]，并将其应用于自己的搜索引擎当中，自此，知识图谱正式进入了大众的视野。Google 公司将网页上的信息进行刻画，通过将海量的非结构化文本抽取实体和关系，转化为相互关联的大型知识网络，从而让自己的搜索引擎能够获取更加精确的结果。在 21 世纪，知识图谱被看作是人工智能的基础性建设内容，成为了各学术界和各大互联网公司研究的重要内容。在国内，百度公司和搜狗公司率先建立起了自己的知识图谱，即“知心”和“知立方”[6]，并各自服务于自己的搜索引擎中，大大调高了搜索的准确性和高效性。

智能问答系统这个概念最早在上世纪 60 年代被提出，目的主要是让计算机能够理解并回答人类提出的自然语言形式的问题，其中包含了一些建设性的工作：Jaseph W 等人设计了用于对精神病人心理治疗的智能问答系统 ELIZA[7]。1972 年，在斯坦福，诞生了加强版的 ELIZA，名叫 parry[8]，它在说话是能够带有自己的语

气。1980 年左右，使用和研究基于人类自然语言进行提问的智能问答系统成为了当时的研究热点。在上世纪 90 年代，由于信息检索技术和自然语言语义分析技术取得了非常丰厚的研究成果，极大的促进了智能问答系统的技术研究和应用发展。在这段时期内，出现了一些著名的智能问答系统，如 MIT 研发的 Start 智能问答系统[9]，还有微软公司开发的 Encarta 智能问答系统等。

直到现在，智能问答系统的研究已经长达半个世纪。但是早期的研究主要是针对小型封闭或者专有的知识库，主要使用语义分析为主的研究方法，这种方法需要人工标注和干预比较多，往往需要标注为“自然语言-逻辑表达式”，需要研究人员花费大量的时间和精力，可移植性非常差，很难将其移植和运用到其他的平台或者领域。后来，研究人员利用其它形式的语料，采用了基于弱监督的方法进行语义的解析。随之互联网的不断飞速发展，出现了很多的大规模的知识图谱，如 Freebase、DBpedia 和 YAGO 等[10]，也逐渐出现了其它各种形式的网络资源，如百科、社区等等，同时由于机器学习和自然语言处理技术的不断发展和进步，智能问答的方向逐渐向更高和更深入的层次进步和发展，基于知识图谱的智能问答系统逐渐成为了各行各业关注和研究的热点。

基于信息检索的方法，主要研究都在命名实体识别和属性链接上。Wang 等人[11]，将字典树的知识融合于深度学习中，并且通过集合多种特征提取方式，在临床领域的命名实体识别取得了不错的效果。Seung-Hoon 等人[12]基于混合语素表示，提出了 LSTM-ConvNet 方法，运用于韩语命名实体识别的改进。吴茜[13]在农业领域，提出了多种特征的 CRF 农业命名实体识别算法，对一个特征组合以及引入上下文信息的 CRF 模型训练取得了较高的 F1 值。Zhou 等人[14]，通过 CNN 网路研发命名实体识别，并根据 BILSTM 和 CNN 具体实现属性链接。胡婕等人[15]灵活的将序列标准型 Bi-LSTM+CRF 应用于问句的槽位提取，实现了基于深度学习的领域问答系统。王兵等人[16]通过专业领域细粒度的分词和简单的余弦相似度算法，同时结合多种特征，实现了基于 Tr_BiLSTM_CNN 的钻井安全问答系统。鉴于 CRF 和 LSTM 在问答系统领域的良好表现，本文使用了相关的模型进行了命名实体识别和属性链接的研究。

1.5 论文结构安排

本文的内容组织如下：第一章介绍了基于知识图谱的智能问答系统的研究背景、研究方法、国内外研究现状和本文的论文结构；第二章具体介绍了本文使用的知识图谱和图谱的降噪处理；第三章介绍了具体的命名实体识别和属性链接所

采用的算法，并对模型做出了测试；第四章探讨了问答系统的设计与实现，从需求分析、开发环境、系统设计、系统实现和性能分析五个方面介绍整个问答系统；第五章根据前四章的分析进行了总结，对问答系统不足之处做出了思考和展望。

第二章 知识图谱简介和预处理

2.1 知识图谱简介

知识图谱是由一条条可以表示为三元组的知识组成的。三元组主要有两种表现形式，即（实体 1，关系，实体 2）和（实体，属性，属性值）。在图形化的表示中，知识图谱是由节点与边形成的一张知识网络，其中实体或者属性值就是节点，关系或者属性就是节点与节点之间的连线。

知识图谱一般可以分为三类原始数据类型：第一类主要是结构化的数据，如关系数据库等；第二类主要是非结构化的数据，如文本、多媒体、图片等；第三类主要是半结构化的数据，如 XML、JSON、百科等。通用的知识图谱构建过程如图 2.1 所示，就是将这三类原始数据抽取为结构化的数据的过程。

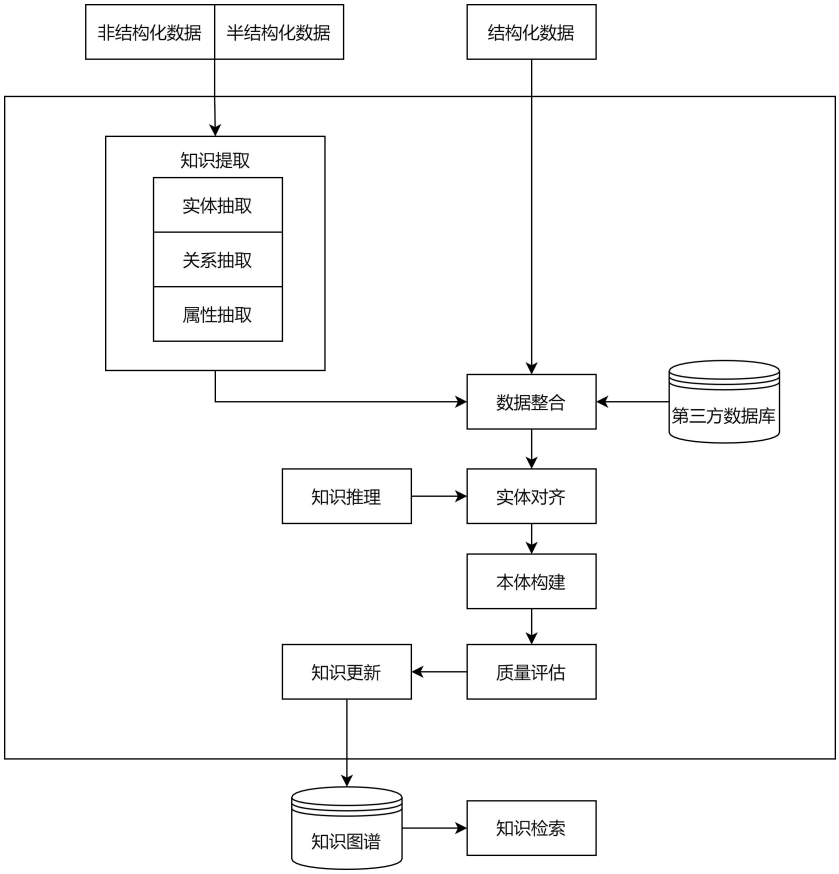

图 2.1 知识图谱构建

同时，知识图谱的构建过程也是一个不断更新迭代的过程，主要涉及知识提取、知识融合[17]和知识加工三个主要步骤。知识提取就是从三类原始数据中提取出实体与实体、实体与属性之间的相互关系，并在此基础上综合形成本体化的知识表达；知识融合就是在获取新知识之后，统一知识的格式，或者消除知识中的部分矛盾和歧义；知识加工是将经过知识融合的知识进行质量优化评估，以确保整个知识库的质量。

2.2 知识图谱预处理

本文采用的知识图谱是由 NLPCC-ICCPOL 2016 KBQA 任务集提供的数据集构建的。该数据集是一个大规模的通用知识库，包含了 6502738 个实体和 43063796 个三元组，涉及到文学、娱乐等各个领域。该库中每一行都是一个事实，也就是每一行都是一个三元组（实体，属性，属性值）或者（实体 1，关系，实体 2），所有的三元组的集合，构成了一张知识网络，本质上这就是一个庞大的知识图谱。该知识图谱中的数据如表 2.1 所示。

表 2.1 知识图谱数据示例

实体	属性	属性值
中华奇石	别名	中华奇石
中华奇石	中文名	中华奇石
中华奇石	外文名	Chinese Stone Arts
中华奇石	创刊时间	2007 年 7 月
中华奇石	出版周期	月刊（每月 5 日出版发行）
中华奇石	语言	中文
中华奇石	国内定价	28 元人民币（336 元/年/12 期）

由于该知识图谱中数据存在不少的噪声，如存在一组相同的三元组，或者属性值出现无用或缺少的字符，都会对实验造成一定的困扰。所以本文首先对该数据进行了降噪处理，具体规则如下表 2.2。

表 2.2 知识图谱去噪规则

规则	去噪前	去噪后
加上未显示全的书名号	《哈姆雷特	《哈姆雷特》
去除中空白的字符	文 章	文章
去除中英文混杂的三元组	存在该三元组	不存在该三元组
去除重复的三元组	存在重复的三元组	不存在重复的三元组

同时，本文将知识图谱存储于 MySQL 数据库中，由于数据量庞大，可能会对后续实验造成影响。所以在数据库中为该知识图谱（实体，属性，属性值）列都建立了普通索引，以加快三元组的查询效率。在建立索引方法的时候，通过比较 BTREE 和 HASH 索引方法对查询速度的影响，最终选择 BTREE 索引方法。

第三章 问答系统算法设计

3.1 算法综述

本文将基于开放领域知识图谱的智能问答拆分为两个主要步骤：命名实体识别和属性链接。其中，命名实体识别的主要目的是找到问句中的实体，而属性链接部分主要目的是找到知识图谱中与问句关系最接近的属性，最终答案选择都基于这两个步骤。这两个步骤的大体框架如图 3.1 所示，用户在问答系统界面输入问句“《高等数学》是哪个出版社出版的？”首先通过命名实体识别步骤找到问句中包含的所有实体作为候选实体，如“高等数学”和“出版社”，并为每个候选实体计算得分；属性链接步骤将候选的实体的全部属性作为候选属性，在本例中为“主编”、“出版社”、“别名”、“外文名”和“隶属”，然后计算每个候选属性与问句的语义相似度，转化为得分，并且将候选实体得分和候选属性得分做加权处理，得到每个三元组的最终分数；最后，通过排序，选取最高分的三元组作为答案，本例中最终加权得分最高的三元组为（高等数学，出版社，武汉大学出版社），即属性值“武汉大学出版社”为系统的答案。

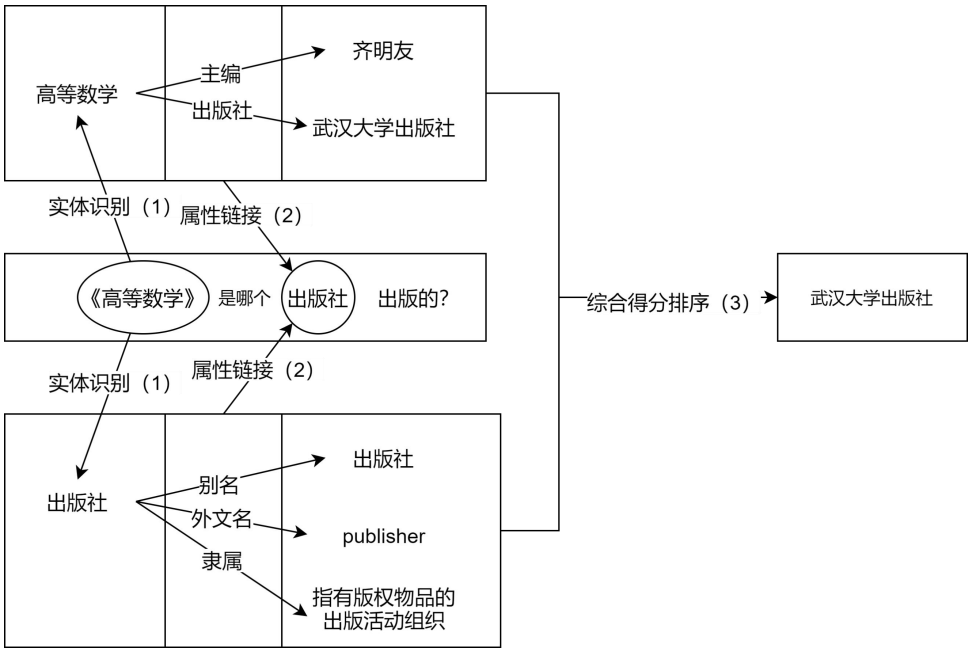

图 3.1 问答系统主要步骤结构图

3.2 命名实体识别

3.2.1 CRF 模型

在自然语言处理中，命名实体识别是一个热门的研究方向，从早期方法，如基于规则的方法[18]、基于字典的方法等，到后来的传统机器学习方法，如 HMM、MEMM 和 CRF 等，再到近期的深度学习方法，如 RNN-CRF、CNN-CRF 等，最后到近期火热的注意力模型、迁移学习模型和半监督学习模型等。前期基于规则等方法处理命名实体识别虽然准确率很高，但需要研究人员拥有大量专业知识，手工编写识别规则，覆盖率不高。后期火热的机器学习和深度学习主要是利用已经提前标注好的语料来进行训练，让模型去学习和计算句子中某个词语或者某个字作为命名实体的概率，接着利用训练好的模型预测一个候选词字作为命名实体的概率，如果这个概率大于一个阈值，那么就认为这个候选词是实体。

CRF（Conditional Random Fields）模型，也称为条件随机场模型。数学界给出的准确的定义为：假设 X 与 Y 是随机变量，P(Y|X)是给定 X 时 Y 的条件概率分布，若随机变量 Y 构成的是一个马尔科夫随机场，则称条件概率分布 P(Y|X)是条件随机场。其中，条件随机场模型是指给定一组输入序列的条件下，另一组输出序列的条件概率分布模型。随机变量的集合称为随机过程，而一个空间变量索引的随机过程，就成为随机场。也就是说，一组随机变量按照某种概率分布随机复制到某一个空间的一组位置上时，这些赋予随机变量的位置就是一个随机场。而条件随机场，就是给定一组观测状态的马尔可夫随机场（Hidden Markov Model，HMM）。马尔可夫随机场可以简单理解为一个位置的值只和它相邻的位置的值有关的随机场，也就是说 CRF 考虑到了观测状态这个先验条件。

在传统机器学习中，CRF 模型虽然训练代价大，复杂度较高，但是 CRF 模型特征设计灵活，与 HMM 模型相比，它不存在苛刻的独立性假设条件，更重要的是它能够结合任何的上下文信息。其次，由于 CRF 模型计算了全局最优输出几点的概率，它还克服了最大熵马尔科夫模型（Maximum Entropy Markov Mode）标记偏执的缺点。最后 CRF 模型是计算整个标记序列的联合概率分布，而不是给定的当前状态下，定义下一个状态的状态分布。本文将易于上手并且热门的 CRF 模型用于命名实体识别，并就此进行了实验和测试。

3.2.2 基于 CRF 模型的命名实体识别

实体识别的过程需要解决两个关键性问题：实体类别判断和实体边界确认。以句子“刘德华出演过很多电影”为例，实体类别判断指必须判断“刘德华”为人名实体，而不是其他类型的实体，如时间实体或者地名实体等；实体边界确认指算法必须将实体词“刘德华”进行正确的标记，而不是在其他位置划分，如划分为“刘德华出”。解决这个问题最简单的办法就是使用 B-I-O 表示法，其中 B 代表实体起始位置，I 表示其他实体字，O 代表非实体字。

模型训练之前需要对语料做数据预处理，数据预处理的目的是提高实体标注的质量，主要步骤包括：字符转换、时间词合并、人名合并和短词合并。在字符串转换步骤中，主要是将一个字符串中存在全角字符，就将它转化为半角字符，针对标点符号；在时间词合并步骤中，主要是将本来连在一起的时间词语拼接在一起，如语料中存在“1 月/t 1 日/t” 的文本，合并之后应为“1 月 1 日”；在人名合并步骤中，主要是将本来连在一起的人名词语拼接在一起，如语料中存在“黄/nr 飞鸿/nr” 的文本，合并之后应为“黄飞鸿”；在短语词合并步骤中，主要是将以“[]”括起来的词进行合并，如语料中存在“[中央/n 人民/n 广播/vn 电台/n]nt” 的文本，合并之后为“中央人民广播电台/ nt”。

在语料标注阶段，需要根据 B-I-O 表示法将实体词进行标注，本文使用的实体标签主要有十三种：O，B_N，I_N， B_NR，I_NR，B_T，I_T，B_NS，I_NS，B_NT，I_NT，B_Z，I_Z，主要区分普通名词、人名、时间、地名、组织机构和专有名词。对于语料中不属于实体的词，使用 O 进行标记，而对于实体词，不仅要标记类型，还要标记实体边界，如表 3.1 所示。

表 3.1 语料标注示例表

语料	标注
袁	B_NR
隆	I_NR
平	I_NR
完	O
成	O
英	B_N
文	I_N
演	B_N
讲	I_N

模型训练之前，还需要定义特征模板，用于在指定数据中提取指定的特征。在命名实体识别步骤中，本文一共用了 5 个模板来制造特征，分别是：当前字、当前字的前一个字、当前字的后一个字、前一个字与当前字的组合，当前字与后一个字的组合，如表 3.2 所示。

表 3.2 特征模板

特征	含义
W	当前字
W-1	当前字的前一个字
W+1	当前字的后一个字
W-1：W	前一个字与当前字的组合
W：W+1	当前字与后一个字的组合

经过上述处理之后，能够更好地运用于模型的训练，整个命名实体识别具体流程如图 3.2。训练好的 CRF 模型是一个庞大的篱笆网络，在该网络中每个节点是每个预测值的不同取值，想要从文本中进行实体识别，就是通过在网络中寻找最大概率的路径来确定输出的实体标记。

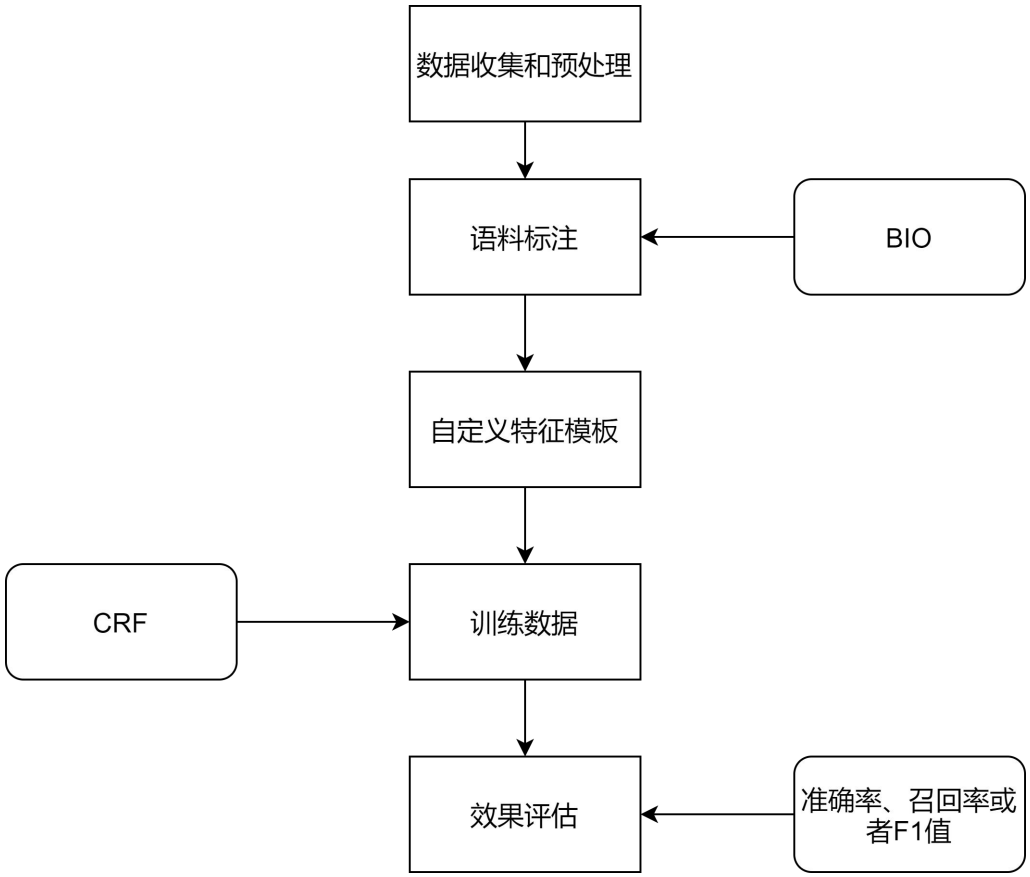

图 3.2 命名实体识别流程图

通过对数据的大量观察，本文抽取两个相关的特征对识别出来的所有候选实

体进行评分，分别为命名实体词长度 L 和词频-逆向文件频率 TF-IDF（为方便书写，公式中简称 TI），通过两个特征的加权得到每个候选实体的得分，公式如 3.1 所示：

$$ENTITY\_SCORE = \lambda * L + (1 - \lambda) * TI \quad (3.1)$$

3.2.3 实验结果与分析

本文采用了已经分词处理的《人民日报》语料库作为实验的训练数据，命名实体识别步骤训练集和测试集的数据集规模如表 3.3 所示。

表 3.3 命名实体识别数据集

训练集	测试集
40158	14609

同时，为了测试本文的 CRF 模型扩展性，本文就《人民日报》语料和处理过后的 NLPCC-ICCPOL 2016 KBQA 问答对语料分别对模型分别进行了测试，具体处理方式是使用 Jieba（Python 的分词工具）对问句分词处理，得到的实验结果如表 3.4 所示。在该表中，Precision 代表精确率，其含义是在被所有预测为正的样本中实际为正样本的概率；Recall 代表召回率，其含义是在实际为正的样本中被预测为正样本的概率；而 F1 同时考虑精确率和召回率，能综合能反映模型好坏的指标。在基于《人民日报》语料的训练中，三个指标的结果都较好，而在 NLPCC-ICCPOL 2016 KBQA 问答对语料中，三个指标的结果表现却较差，分析语料库可以得到两个原因：一是是分词不够准确，分词粒度不够细；二是两个语料的句子特征不一致，也就是说特征模板不太适用于 NLPCC-ICCPOL 2016 KBQA 问答对语料，重要特征在该语料上的分布发生了较大的变化。

表 3.4 命名实体识别实验结果

语料	Precision	Recall	F1
《人民日报》	0.94	0.88	0.91
NLPCC-ICCPOL 2016 KBQA	0.85	0.68	0.75

同时，为了对公式 3.1 中的权值λ进行实验，测试它对正确命名实体的影响，本文对上述 13610 条测试问句进行了处理，具体方法是根据问答对的正确答案，在知识图谱中反向查找相关的三元组，该三元组中就含有正确的实体。所以本文在该数据集能够识别出命名实体的前提下，对该数据集进行实验，同时在实验结果中筛除掉不能够识别出实体的问句和不存在于本文知识图谱的实体有关的问句，得到如图 3.3 的所示的结果，可以看出：λ权值对命名实体的筛选有一定的影响，特别表现在λ权值设置为 0 的时候，命名实体的正确率较低，而设置了λ权值的权

值后命名实体的正确率显著提高；同时随着λ权值的增大，命名实体的正确率变化浮动较小或者没有变化，在 0.1 时去得最大值。综合考虑，本文将λ权值选定为 0.1 以继续进行后面的实验。

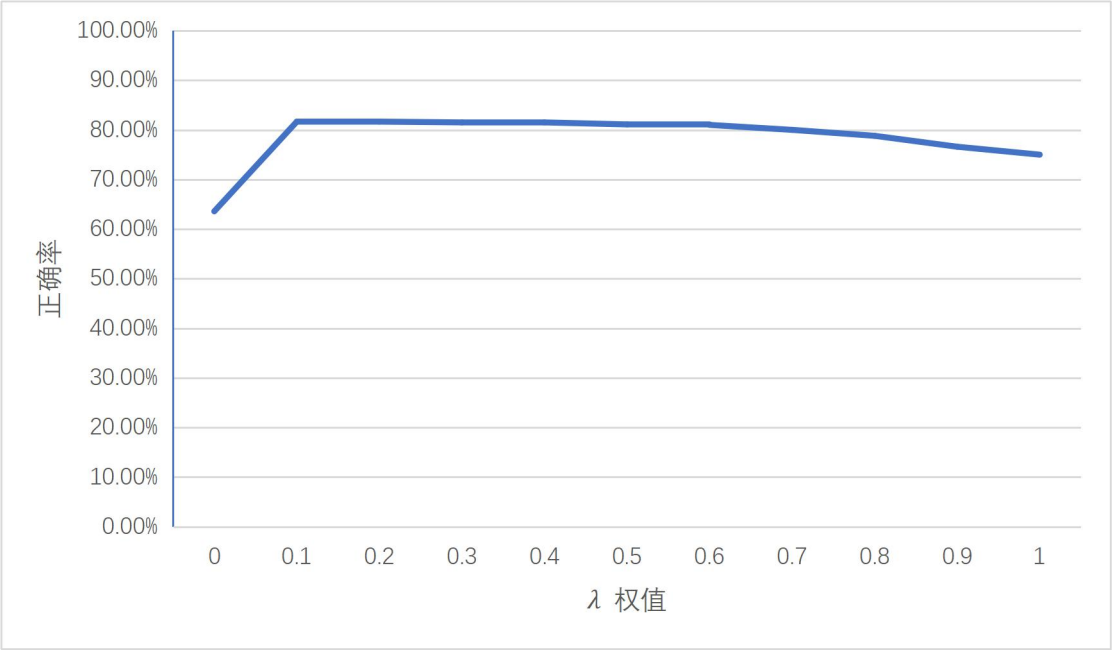

图 3.3 λ权值折线图

3.3 属性链接

3.3.1 LSTM 模型

LSTM 也可以叫做长短时记忆，是一种循环神经网络(RNN)特殊的类型，可以用来学习长期依赖的信息，解决训练过程中远距离传递导致信息丢失问题，如梯度消失和梯度爆炸。与 RNN 相比，在更长的序列，LSTM 明显具有更加好的效果。目前已经证明，LSTM 是解决长序依赖问题的有效技术，并且这种技术的普适性非常高，导致带来的可能性变化非常多。

RNN 和 LSTM 主要输入和输出区别如下图 3.4 所示，h^t代表隐藏状态（hidden state），C^t代表处理器状态（cell state），而x^t和y^t分别代表输入和输出。RNN 只拥有一个传递状态h^t，而 LSTM 拥有h^t和C^t两个传输状态。在 LSTM 中C^t改变的很慢，在通常情况下，输出C^t是上一个状态C^{t-1}增加了一些数值；而h^t改变的很快，在通常情况下，输出h^t和上一个状态h^{t-1}差别非常大。

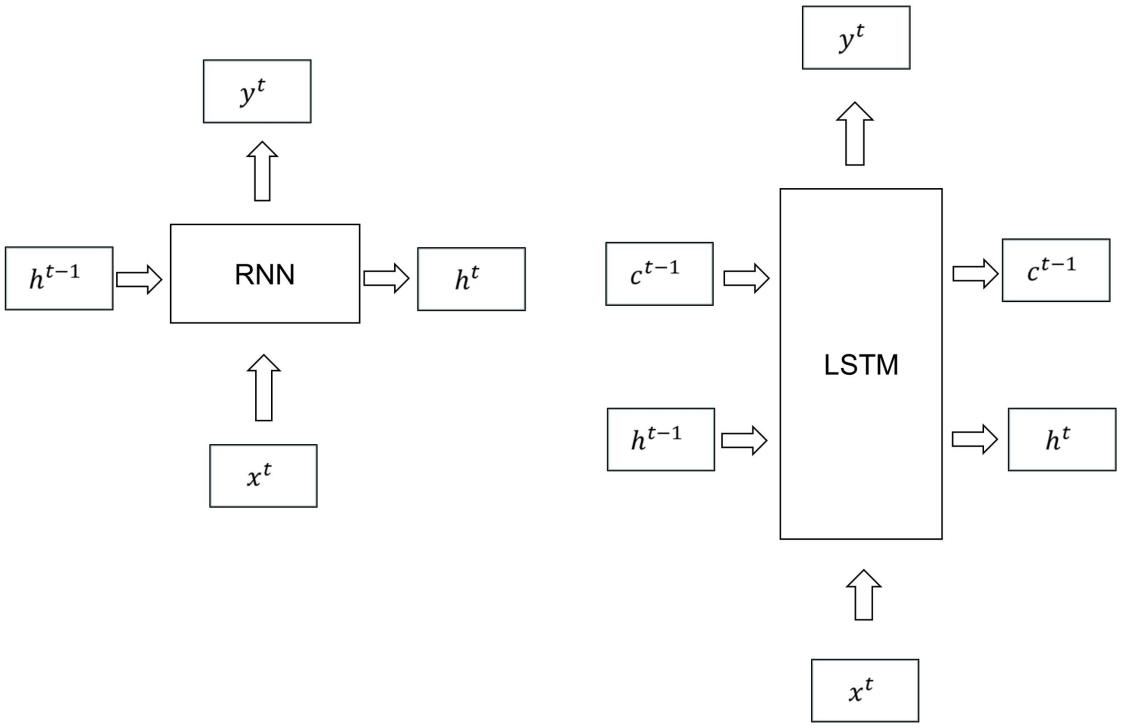

图 3.4 RNN 和 LSTM 输入输出

和 RNN 相比，LSTM 最大的不同主要在于一个被称为 cell 的特殊结构，这个结构是一个处理器，主要功能就是判断信息是否有用。遗忘门、记忆门和输出门，这三类门被放置在每一个 cell 中。一个信息进入 LSTM 的网络当中，可以根据规则来判断是否有用，只会留下符合算法规则的信息，不符合规则的信息在遗忘门中被淘汰。遗忘门的任务就是接受上一个单元模块传过来的输出，也就是传入的状态C^{t-1},并决定要保留和遗忘C^{t-1}的哪一个部分。记忆门的作用是确定是什么样的新信息能够放在细胞状态，也就是单元模块中。输出门的作用就是根据细胞状态，确定输出值。

3.3.2 基于 LSTM 模型的属性链接

经过命名实体识别步骤之后，假设输入的问句为“《高等数学》是哪个出版社出版的？”，则可以得到问句询问的两个命名题“高等数学”和“出版社”，并依据特征公式 3.1 加权评分.知识库中关于这两个候选实体的三元组有 5 个，其属性名分别为“主编”、“出版社”、“别名”、“外文名”和“隶属”。属性链接的目的是在这五个属性中，找到与问句“《高等数学》是哪个出版社出版的？”语义

相似度最为接近的属性，那么这个三元组暂时就是问句询问内容最相关的答案。

属性链接步骤使用基于 LSTM 模型的 Simaese LSTM（孪生网络）模型，本质上是一类特殊的神经网络。与一个学习对其输入进行分类的模型不同，该神经网络是学习在两个输入中进行分类区分，也就是说它学习了两个不同输入之间的相似之处。属性链接步骤总体思想是对问句和属性进行编码，得到对应的语义向量，通过计算这两个向量的曼哈顿距离得到问句和属性的相似度。曼哈顿距离可以定义为在欧几里得空间直角坐标系上两点所形成的对轴产生的投影的距离总和。在多维度空间上，两个 n 维度向量 a（a1;a2…;an）与 b（b1;b2…;bn）之间的曼哈顿距离 d 可以定义为公示 3.2，如下：

$$d = \sum_{k=1}^{n} |a_k - b_k| \tag{3.2}$$

下图 3.5 是整个网络的大致过程，左边输入问句，右边输入属性后，将句子或属性中进行分词，为每个词编码并 embedding 后，那么左右的问句和属性就分别映射成了一个向量矩阵，它们各自经过一个 LSTM 网络提取特征后，使用曼哈顿距离计算这两个向量的差距，最终得到预测结果。

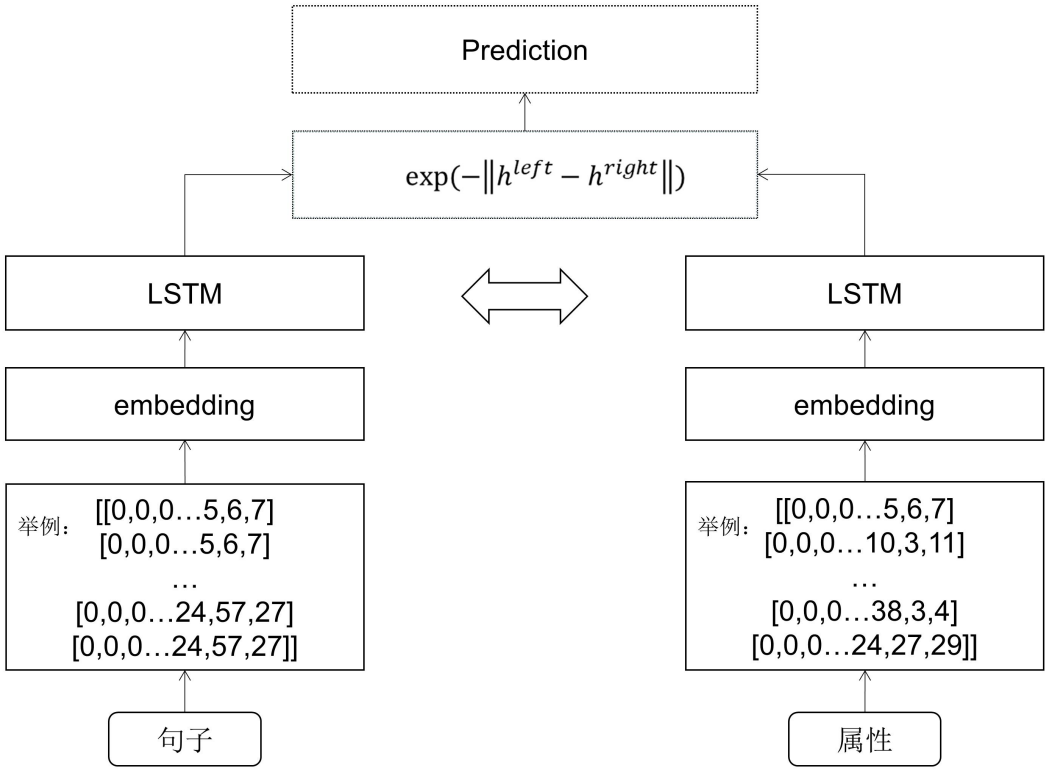

图 3.5 属性链接流程图

在上述每次训练和预测过程中，都需要将句子或属性进行分词，构造字典，

为每个词编码，直接影响了模型训练和预测速度。所以本文，就所有涉及的语料，包括知识图谱，建立了一个字典，字典中存储的是每个词和它对应的序号（索引）。在模型训练和预测之前，只需要将句子分词，在字典中为每个词找到对应词的索引便能更快的将词编码。

同时，通过对数据的大量观察，发现只依据模型预测的语义相似度并不能可靠的判断问句与属性之间的相似度，比如问句“黄飞鸿的电话是多少？”，属性“电影”的语义相似度要属性“电话”的相似度要高。所以本文在属性评分步骤，引入了一个特征 OR（Overlap Ratio）：属性与问句的重叠比例（以字为单位），即属性中的字出现在问句中的总数占问句字总数的比例，在计算相似度的时候将模型得到的相似度 LSTM_SCORE 与该特征进行加权，得到公式如 3.3 所示：

$$ATTRIBUTE\_SCORE = \beta * \mathrm{OR} + (1-\beta) * \mathrm{LSTM\_SCORE} \quad (3.3)$$

3.3.3 实验结果与分析

由于 NLPCC-ICCPOL 2016 KBQA 提供的问答对数据集（包括训练集和测试集）中只提供了问句和答案，本文将该数据集处理之后作为属性链接步骤模型的训练接和测试集，具体处理方法和命名实体识别步骤相同，是根据问答对的正确答案，早知识图谱中反向查找问句对应的三元组，该三元组中包含问句正确的属性。本文在该语料库的基础上添加相同数量的负样例，目的是将数据集中的正样例和负样例比例设定为 1:1，具体方法是在知识图谱中找到该问句正确实体的所有属性，随机筛选剔除了正确属性以外的其他属性就可以用来构造负样例。属性链接部分的数据集使用 label 标注正样例和负样例，其中正样例使用 1 表示，负样例使用 0 表示。如表 3.5 所示，是属性链接步骤的数据集大小。

表 3.5 属性链接数据集

训练集	测试集
19484	14609

在基于 LSTM 模型的属性链接测试中，本文使用的是中文维基百科训练而来的词向量，向量维度为 300 维，得到的实验结果如表 3.6 所示，F1 值达到了 0.86，算法效果良好。

表 3.6 属性链接实验结果

模型	Precision	Recall	F1
LSTM	0.85	0.89	0.86

同时，为了确定公式 3.3 中的β值，本文对表 3.3 中的训练集做了测试。测试

需要从实体识别步骤开始，所以实验结果中筛除掉不能够识别出实体的问句和不存在于本文知识图谱的实体有关的问句，得到如图 3.6 的结果。从图可知，随着β值的增大，属性的正确率也显著逐步提高；同时 β值在 0.7 左右，属性正确率趋于稳定，接近 80%。总体来看公示 3.3 中特征值 ol 对属性的正确率有比较大的影响。综合考虑，本文将β权值选定为 0.7 以继续进行后面的实验。

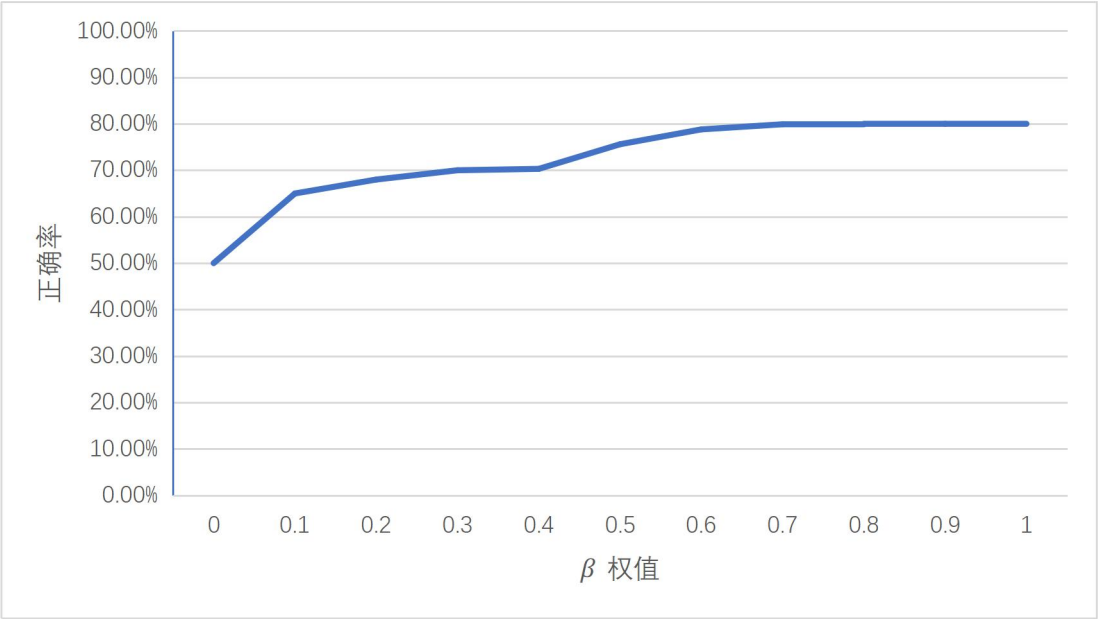

图 3.6 β权值折线图

3.4 答案选择

3.4.1 答案综合选择

在经过命名实体识别和属性链接之后，就到了最后的答案选择步骤。本文最初的选择的策略是：让命名实体步骤得分最高的候选实体经过属性链接步骤之后，选取与问句语义相似度最高的属性，根据该实体和属性确定三元组就能找到三元组的属性值，也就是问句的正确答案。但是，在经过大量试验观察之后，发现这种选择策略存在着一定的误差，正确的命名实体可能不是最高分，最后直接导致了错误的结果，比如问句“城关镇有什么火车站？”，在命名实体识别步骤中实体“火车站”得分略高于实体“城关镇”，但是在属性链接步骤实体“城关镇”的属性“火车站”得分却略高于实体“火车站”的属性“别名”，最后获取到了错误的结果。所以，本文就命名实体识别步骤得分 *ENTITY* _SCORE 和属性链接步骤得分 *ATTRIBUTE* _SCORE 进行加权，对答案综合选择，最后排序选取最高得分，具体公式如 3.4 所示：

$$SCORE = \alpha * ENTITY\_SCORE + (1-\alpha) * ATTRIBUTE\_SCORE \quad (3.4)$$

3.4.2 实验结果与分析

为了选定测试答案选择步骤的公式 3.4 中的α权值，本文将 NLPCC-ICCPOL 2016 KBQA 提供的问答对数据集（包括训练集和测试集合），各随机抽取了部分数据作为整个算法最后答案选择阶段的测试数据，最终该步骤测试数据集为 14609 条问答对。在经过命名实体识别步骤和属性链接步骤处理之后，对公式 3.4 中的α权值进行测试，实验中筛除掉了命名实体识别步骤无法识别出实体的问句和不存在于本文知识图谱的实体有关的问句，结果如图 3.7 所示，实验发现α取值对答案的影响较小，同时发现 α权值取 0.3 时，得到了最好的问答效果，这也侧面反应了正确的实体和正确的属性在实体识别步骤和属性链接步骤基本都取得了最高分，所以导致了综合加权评分对正确结果的影响较小。综合考虑，本文将 α 权值选定为 0.3 以继为问答系统提供良好的算法服务。

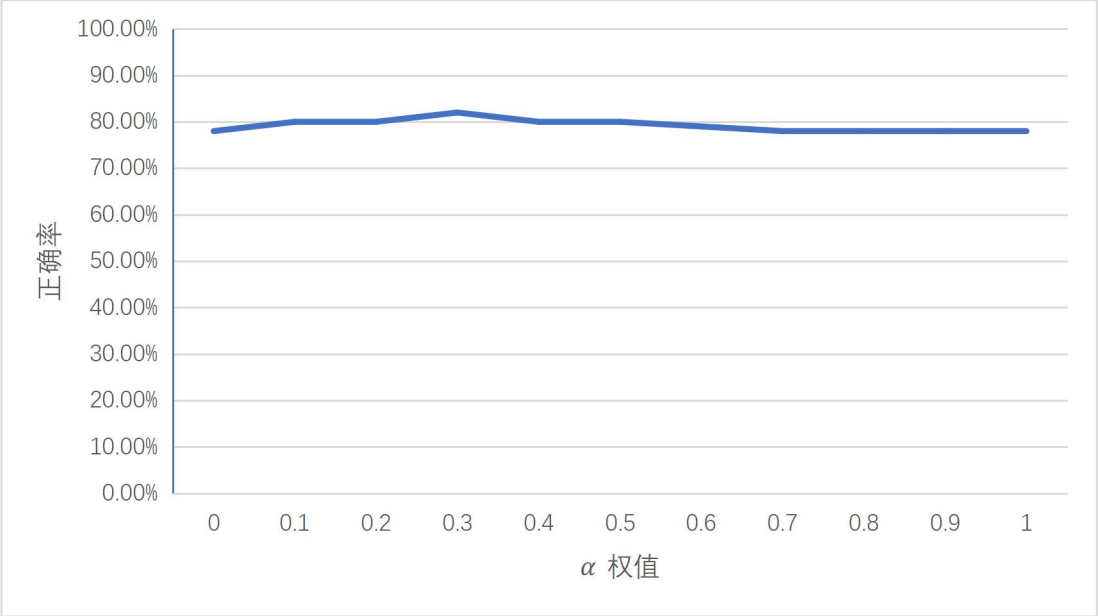

图 3.7 α权值折线图

第四章 基于知识图谱的智能问答系统设计与实现

4.1 需求分析

在系统开发之前，需求分析是非常关键的一步，主要目的是确定系统的功能和应用范围，即知道系统必须要做什么，主要包含了两个方面的需求：功能性需求和非功能性需求。

4.1.1 功能性需求分析

实现本文的智能问答系统需要具备如下的功能性需求：

（1）问题输入：有区域提供用户输入问题。

（2）问题提交：用户输入问句之后可以通过按钮提交。

（3）答案反馈：用户能够看到系统处理之后的答案，以及历史问答记录。

4.1.2 非功能性需求分析

实现本文的智能问答系统需要具备如下的非功能性需求：

（1）安全性要求：系统内数据不会出现丢失错误、恶意修改等，数据传输过程中使用 https。

（2）可靠性要求：系统整体运行稳定，有很强的防错能力，系统管理员能够在短时间内快速恢复系统。

（3）易用性要求：系统不需要登录就可以使用，同时系统符合目前比较流行的界面设计规范，界面清晰、一目了然。

（4）性能要求：综合现有搜索引擎（谷歌、百度等）来看，系统需要秒级的反应速度，操作反应时间应该不超过 10 秒，避免造成用户长时间等待。

（5）可维护性要求：系统每天生成日志，极大简化维护工作。

（6）可移植性要求：系统能够在所有操作系统上使用。

4.2 开发环境

本文所设计的基于知识图谱的智能问答系统主要使用了轻量级 web 框架 Flask 和渐进式框架 Vue。如表 4.1 所示，是实现该系统的开发环境。

表 4.1 开发环境

环境要求	内容
开发语言	Python、HTML、CSS、JavaScript
开发数据库	MySQL
CPU	AMD A12-9700P RADEON R7, 10 COMPUTE CORES 4C+6G 2.50 GHz
内存	8GB
操作系统	Windows10 64 位

Flask 是一个使用 Python 编写的轻量级 Web 应用框架。之所称他为轻量级框架，是因为它的核心非常简单，但同时该框架又具有非常强大的扩展能力。具体来说，Flask 由提供网关接口等功能的 Werkzeug 和提供模板的 Jinja2 这两个重要部分构成。其他的一切，比如数据库处理，文件上传，权限验证等功能都是由第三方库完成。

Vue 是一套构建用户界面的渐进式框架，易与其他项目进行整合。渐进式框架，也就是指 Vue 是逐层设计的，Vue 本身包含了大多数的功能，它的每一层都是灵活的，不是强制性要求的，并且每一层都有自己的解决方法，在使用的时候按需引入官方或者其他三方工具就可以。Vue 的双向绑定是它最大的特性之一，实现双向绑定使用到了 MVVM（ Model-View-ViewModel）的设计模式。该设计模式中，Model 是指模型层，负责业务逻辑的处理和与服务器端的交互；View 是指视图层，负责将数据转化为页面展示出来；而 ViewModel 是用来连接 Model 层和 View 层的中间层，被称为视图模型层。MVVM 是一种数据驱动模式，它尽可能不去改动 dom 树，而是直接去操作数据。View 层和 Model 层并没有直接联系，而是通过 ViewModel 层进行交互。如图 4.1 所示，利用双向数据绑定的原理，View 层和 Model 层被 ViewModel 层连接了起来，最终这两层自动化地同步工作。

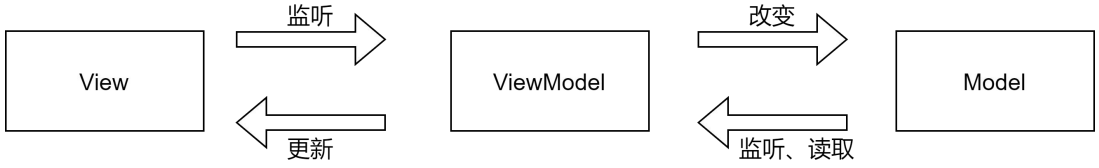

图 4.1 MVVM 模式

4.3 系统设计

本文通过信息检索的方法，提取问句中蕴含的信息，在开放的知识图谱基础上构建智能问答系统来实现问句的解析、知识图谱的检索和答案的生成。本文所构智能问答系统结合了前后端分离的 BS（Browser/Server）架构和三层架构，三层架构主要分为表示层、业务逻辑层和数据访问层，如图 4.2 所示。

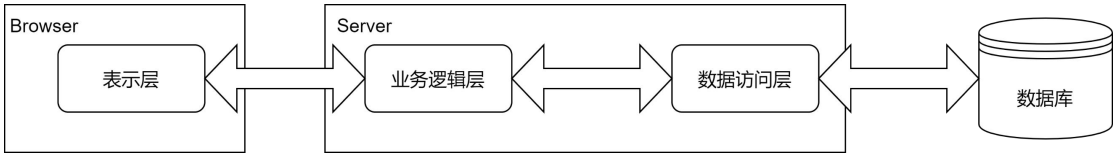

图 4.2 系统整体构架图

浏览器端作为表示层，使用到了 Vue 等技术，主要处理用户的交互，用于接收用户的输入，进行一些简单的分析判断之后，将问题交接到服务器端逻辑处理层，同时显示服务器端处理的结果。服务器端作为业务逻辑层和数据访问层，使用到了 Flask、云 MySQL 数据库等技术，负责处理用户的请求，查询数据并反馈处理的结果。在服务器端中，业务逻辑层是表现层和逻辑层的桥梁，负责一些计算和验证规则，主要使用了基于 CRF 的命名实体识别模型和基于 LSTM 的属性链接模型；数据访问层主要是与 MySQL 数据库交互，实现对数据库的增和查等，比如知识图谱构建和知识查询。

本文将系统主要分为了三大模块：前端展示模块、问答逻辑模块和数据构建模块，如图 4.3 所示。

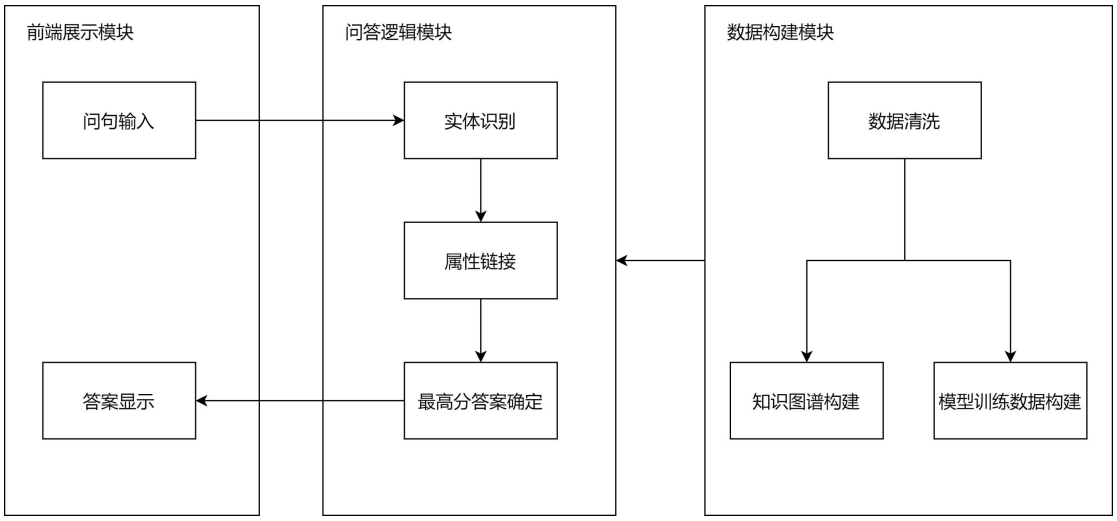

图 4.3 系统模块图

前端展示模块具体使用了 HTML 和 Python 等技术。用户的输入通过

JavaScript 获取，通过一些基础判断之后，通过 Http 提交到后端，后端处理好结果也通过 Http 返回给前端，浏览器通过自己的渲染规则渲染页面。整个页面分为两个部分：第一部分是用户问题输入框和提交按钮，用户在规定区域输入问题，并提交到后端；第二部分是历史问答记录查看框，由于系统不需要登录，用户在使用问答系统且未关闭页面时候，可以查看该次使用系统的历史的问答记录。

问答逻辑模块首先主要利用到的是第三章已经训练好的基于 CRF 的命名实体识别模型，识别来自问句中的实体，然后查找所有有关的候选实体以及它的三元组，再通过特征进行实体得分的计算。等待命名实体识别步骤完毕之后，利用第三章已经训练好得基于 LSTM 的属性链接模型对问句和属性进行语义度计算，然后通过特征对属性进行评分。等到属性链接步骤结束之后，将实体和属性加权得分并将答案排序，最终确定得分最高的答案。

数据构建模块是系统的基础，给系统提供了强大的支持。这一模块主要是为了构建知识图谱和构建模型训练所需要的数据，构建之前需要对数据进行清洗，也就是根据数据的特征，对数据进行一些降噪处理等。

4.4 系统实现

本文实现的问答系统页面如图 4.4 所示，从输入框中输入问题，即可或得该问题的答案。

图 4.4 系统页面图

系统的使用非常简单，比如输入问句“艾玛·沃特森最喜欢的颜色是什么？”，返回结果“紫色与粉色”。点击答案处向下箭头，可以看到其他几个可能的答案，方便用户参考。如果系统未能识别问句中的实体，就会返回“亲，很抱歉，无法

从您的问题中识别出实体!!!”；如果知识图谱中未能查询到有关实体的三元组信息，就会返回“亲，很抱歉，数据库中查找不到您问题相关的答案!!!”。

为了测试本文实现的问答系统的正确率，本文将 NLPCC-ICCPOL 2018 KBQA 提供的问答对数据集（包括训练集和测试集合）用于系统的测试，总共 34093 条问句，结果如表 4.2 所示，系统的正确率只有 60%左右，而错误率达到了大约 40%，综合来看本文实现的问答系统效果较一般。通过对 40%左右没有返回正确答案的问句进行分析发现，其原因主要有三个：一是问句结构不一致，有些中文问句中混入了英文或者问句太短，基于 CRF 模型的命名实体识别算法没有正确识别出实体；二是本文使用的开放领域的知识图谱的规模仍然较小，数据不够完整，无法在知识图谱中查找到命名实体识别步骤识别的实体；三是命名实体识别评分步骤和属性链接评分步骤的特征太少，小部分正确的实体和正确的属性得分依然较低。

表 4.2 问答系统正确率评测

正确率	错误率
59.54%	40.46%

鉴于以上原因，为了提高本文系统的正确率，本文提出三个针对性的解决或者改进方案：一是对系统输入的问句进行过滤，具体是如果用户在问句中加入英文，系统在浏览器给出替换成中文的提示信息，同时在注意事项中提醒用户尽量细化问句；二是扩大本文的知识图谱的规模，加入其它任务或者平台的开放领域知识图谱数据，或者加入不同的专业领域知识图谱数据，并进行去重处理；三是加大对实验数据的观察力度，增加命名实体识别评分步骤和属性链接评分步骤的特征数量，并优化原有的特征。

4.5 性能分析

为了测试本文基于智能图谱的智能问答系统的稳定性，本文使用了 Apache 组织开发的基于 Java 的压力测试工具 JMeter，对问答系统在 1 秒内不同线程数发起问题请求的数据进行收集和分析，共测试 5 组线程数，每组测试 10 次，得到的数据取平均值得到下表 4.3。

表 4.3 问答系统压力测试

线程数/秒	Average/ms	Median/ms	Min/ms	Max/ms	Std.Dev	Error/%
500	1730	1852	68	2996	692.03	0.00
750	1855	2028	75	3252	591.40	37.60
1000	1865	2032	76	2852	578.42	65.00
2000	2259	2132	118	4121	372.68	77.60
3000	2444	2245	404	5778	565.13	85.50

表格中的线程数模拟的是用户数量；Average 指的是系统的平均反应时间；Median 指的是反应时间的中位数；Min 指的是最小的反应时间；Max 指的是最大的反应时间；Std.Dev 指的是响应时间的标准差，越小反应每次请求的反应时间相差越小；Error 代表错误率，即错误的请求占所有请求的比例。从每秒 500 的线程数到每秒 3000 的线程数量，随着用户请求数量的增加，系统的平均反应逐渐增长，同时错误率也在不断增加，在线程数为 500 时，系统的错误率最低，为 0%，说明系统能较好处理 500 左右的低并发量请求，对于高并发量请求性能还不足；表中 Std.Dev 先降后升，不同线程数相差很大，说明本文实现的问答系统的稳定性较差。

第五章 总结与展望

5.1 总结

随着互联网的飞速发展，传统的智能问答系统性能逐渐受到挑战，同时知识图谱的出现和发展，为新一代高效和准确的问答系统奠定了坚实的基础。本文旨在基于开放的知识图谱，实现一个实用、高效的智能问答系统。为此，本文做了以下主要工作：

（1）总结了知识图谱有关的基础理论并介绍了本文使用的知识图谱的预处理。本文简述了知识图谱有关的理论知识，并且使用 NLPCC-ICCPOL 2016 KBQA 任务提供的开放领域知识库构建知识图谱。为了提高系统的效率，本文对该知识图谱做了降噪处理，并使用 MySQL 数据库对知识进行存储。

（2）将问答过程分为命名实体识别和属性链接两个主要步骤进行研究。在命名实体识别步骤，本文基于人民日报语料，使用了 CRF 模型算法，同时通过观察实验，提出了两个特征（命名实体词长度 L 和词频-逆向文件频率 TF-IDF）对实体进行加权评分，最后对该模型的准确率、召回率和 F1 值做出了评测，实验结果表明模型的各项评测数据良好，但实际识别效果一般，分析原因主要是模型的特征模板适用性较差，特别是对短问句的处理能力较差。在属性链接步骤，本文基于 NLPCC-ICCPOL 2016 KBQA 任务提供的问答对语料，使用了双层 LSTM 模型的算法，同时通过对大量的数据的观察实验，提出了一个特征 OR（属性与问句的重叠比例）对属性进行加权评分，最后对该模型的准确率、召回率和 F1 值做出了评测，实验结果表明模型的各项评测数据良好。

（3）设计并实现了基于知识图谱的问答系统。本文实现的系统使用了三层架构，表现层使用了 Vue 等技术，业务逻辑层和数据处理层使用了 Flask 等技术，系统的准确率效果不太理想，为此分析了可能原因和改进方向，最后使用压力测试工具 JMeter 对系统进行了测试。

5.2 展望

基于知识图谱的智能问答系统是一个涉及信息表示、信息检索、自然语言处理、机器学习和人工智能等多个热门领域的研究方向，由于本人学历较低，资历较浅，综合本文所完成的基于知识图谱的智能问答系统来看，还存在许多不足的地方，需要完善和改进：

首先，数据收集不够灵活。相比与其他开放的知识图谱，本文所涉及的NLPCC-ICCPOL 2016 KBQA 任务提供的知识图谱不够全面，所涉及的数据量和知识依然还比较小，不够丰富，知识覆盖率较低。需要进一步丰富数据源，以提高知识图谱的知识广度，进而提高问答系统的解答率。

其次，命名实体识别和属性链接所涉及的算法模型实际效率较低。基于 CRF 的命名实体识别虽然能很好地结合观测条件，自定义特征模板，获得比较高的准确率，但是对于数据的预处理却比较严苛，需要结合其他模型进行改进；基于 LSTM 的属性链接在不同的语料中效果差别较大，还待深入研究。以上两个步骤本文都分别只使用到了一种模型，对于相同语料，相同预处理，在不同模型下的表现无法科学地进行比较。

最后，整个智能问答系统的功能比较单一。本文就主要的问答功能进行了设计，没有考虑好用户其他模块功能，如帮助反馈、知识库或语料捐赠等，还需要不断完善。

致谢

光阴似箭，日月如梭，大学生活一晃而过。在大学的四年里，我积极上进，不断努力提高自己，倍感充实。当我写完这篇论文的时候，我感慨良多。这期间，我的进步离不开各位老师、同学和朋友们的帮助。在此，我由衷的感谢你们。

首先，我要非常诚挚的感谢我的指导老师袁国斌。在论文的撰写过程中，袁老师耐心的解答我的疑惑，为我提供相关技术上和方法上的指导，在忙碌的教学工作中挤出时间来审阅我的论文，给予了我极大的帮助。同时，袁老师严谨细致、一丝不苟的作风更是充分激励我在工作和学习中不断进步。

其次，我要特别感谢在大学四年中所有给予过我帮助的老师、同学、朋友和家人们。是你们教授我丰富的知识，提高了我不断自学的能力。是你们一直支持、鼓励和帮助我，也是你们为我的工作和学习提供了宝贵的建议和意见，我才能不断进步，才能更加快乐充实地度过四年的学习生活。

然后，我还要感谢一路培养我、照顾我、支持我的父母，谢谢你们。

最后，我要特别感谢各位评阅老师的宝贵建议，也要感谢中国地质大学（武汉）和地理与信息工程学院对我四年的培养。

参考文献

[1] 张瀚月, 丁妍, 张毓, 等. 基于深度神经网络的电子商务智能客服系统设计与实现[J]. 软件工程, 2021,24(05):33-37.

[2] 杜泽宇, 杨燕, 贺樑. 基于中文知识图谱的电商领域问答系统[J]. 计算机应用与软件, 2017,34(05):153-159.

[3] 王雪梅. 数字人文领域中知识图谱的研究与应用[J]. 山西科技, 2020,35(06):94-98.

[4] 高顺峰. 基于知识图谱的问答系统设计与实现[D]. 江苏科技大学, 2020.

[5] 俞阳. 基于知识图谱的电力知识平台关键技术研究[D]. 东南大学, 2018.

[6] 马腾, 倪睿康, 李艳茹, 等. 知识图谱在个性化教学中的应用研究[J]. 中阿科技论坛(中英文), 2021(02):177-180.

[7] Bishnu P P, Rupak T, Anil U. Chemical composition, antioxidant and antibacterial activities of essential oil and methanol extract of Artemisia vulgaris and Gaultheria fragrantissima collected from Nepal[J]. Asian Pacific Journal of Tropical Medicine, 2017,10(10):952-959.

[8] Colby K M. Clinical implications of a simulation model of paranoid processes[J]. Arch Gen Psychiatry, 1976,33(7):854-857.

[9] 安超. 面向问答系统的问题相似性研究[D]. 国防科学技术大学, 2016.

[10] 谭晓, 张志强. 知识图谱研究进展及其前沿主题分析[J]. 图书与情报, 2020(02):50-63.

[11] Wang Q, Zhou Y, Ruan T, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition[J]. J Biomed Inform, 2019,92:103133.

[12] Seung-Hoon N, Hyun K, Jinwoo M, et al. Improving LSTM CRFs Using Character-based Compositions for Korean Named Entity Recognition[J]. Computer Speech & Language, 2018,54.

[13] 吴茜. 基于知识图谱的农业智能问答系统设计与实现[D]. 厦门大学, 2019.

[14] Zhou G, Xie Z, Yu Z, et al. DFM: A parameter-shared deep fused model for knowledge base question answering[J]. Information Sciences, 2021,547.

[15] 胡婕, 陶宏才. 基于深度学习的领域问答系统的设计与实现[J]. 成都信息工程大学学报, 2019,34(03):232-237.

[16] 王兵, 郑亚梅, 陈茂柯, 等. 基于Tri-BiLSTM-CNN的钻井安全问答系统[J]. 西南石油大学学报(自然科学版), 2020,42(06):157-164.

[17] Hoang L N, Dang T V, Jason J J. Knowledge graph fusion for smart systems: A Survey[J]. Information Fusion, 2020,61.

[18] 张俊盛, 陈舜德, 郑萦, 刘显仲, 柯淑津. 多语料库作法之中文姓名辨识[J]. 中文信息学报, 1992(03):7-15.