学校代码：10730
分类号：TP311　　密级：公开

兰州大学

硕士学位论文

（专业学位）

论文题目（中文）	基于医学文献的抑郁症知识图谱构建及应用研究
论文题目（外文）	**Construction and Application of Depression Knowledge Graph Based on Medical Literature**
作者姓名	张玉峰
类型领域	工程硕士·计算机技术
研究方向	知识图谱
教育类型	学历教育
指导教师	李泽鹏　副教授
合作导师	
论文工作时段	2021 年 6 月至 2022 年 3 月
论文答辩日期	2022 年 5 月 28 日

校址：甘肃省兰州市城关区天水南路 222 号

基于医学文献的抑郁症知识图谱构建及应用研究

中文摘要

抑郁症是一种常见的精神疾病，具有潜伏周期长、致残率高、死亡率高、复发率高等特点，在世界范围内构成了严重的健康问题。因此，越来越多的研究人员将目光转移到了抑郁症的研究上，并希望借助开放领域的医学知识来促进抑郁症的研究。而开放领域的医学知识最主要、最直观的来源就是生物医学文献。关于抑郁症的生物医学文献数量庞大且杂乱无章，这无疑会增加生物医学研究人员和医务工作者获取抑郁症相关知识的负担，从而阻碍了抑郁症的研究。

知识图谱技术是实现机器认知智能和推动各行业智能化发展的关键技术，作为下一代人工智能的基石，知识图谱技术吸引了来自学术界和工业界的广泛关注。近年来，知识图谱领域涌现出了大量的理论与技术研究成果,将知识图谱技术与各行业进行深度融合已经成为了一个重要趋势。因此，本文将抑郁症与知识图谱相结合开展了研究工作，以期能够为抑郁症的病理研究及治疗方法的研究提供辅助作用。

本文的研究内容主要分为三个部分，第一部分是抑郁症知识图谱的构建部分，该部分的工作是以医学网站 PubMed 上的抑郁症相关的生物医学文献为主要数据源，采用自底向上的方式进行了知识图谱的构建，并针对构建过程中涉及的知识获取、知识表示、知识融合、知识存储等关键技术展开了研究。最终，构建出了一个质量较高的带权重的抑郁症知识图谱——DKG，该知识图谱中包含 136364 个三元组，其中实体数量为 37112，关系种类为 30。第二部分为基于知识图谱嵌入的药物发现研究，这一部分在 DKG 的基础上，以知识图谱的结构信息和语义信息为出发点，将知识图谱嵌入方法应用到了药物发现领域的药物重定向任务中，并通过实验验证了该方法的有效性。第三部分为基于抑郁症知识图谱的智能问答系统设计与实现，该部分以 DKG 为数据支撑，采用基于模板的 KBQA 方法，然后进行了系统需求分析和设计，最终使用 Java 语言和 SpringBoot 框架实现了一个面向抑郁症的智能问答系统 Depression automatic Q&A System，经过测试发现该问答系统能够有效地回答 70%以上的问题。

关键词：知识图谱，抑郁症，药物发现，知识图谱嵌入，问答系统

CONSTRUCTION AND APPLICATION OF DEPRESSION KNOWLEDGE GRAPH BASED ON MEDICAL LITERATURE

Abstract

Depression is a common mental disease, which has the characteristics of long latent cycle, high disability rate, high mortality and high recurrence rate. It constitutes a serious health problem all over the world. Therefore, more and more researchers turn their attention to the research of depression and hope to promote the research of depression with the help of open field medical knowledge. The most important and intuitive source of medical knowledge in the open field is biomedical literature. The biomedical literature on depression is large and disorderly, which will undoubtedly increase the burden of biomedical researchers and medical workers to obtain depression related knowledge, thus hindering the research of depression.

Knowledge graph is the key technology to realize machine cognitive intelligence and promote the intelligent development of various industries. As the cornerstone of the next generation of artificial intelligence, knowledge graph has attracted extensive attention from academia and industry. In recent years, a large number of theoretical and technical research results have emerged in the field of knowledge graph. The deep integration of knowledge graph with various industries has become an important trend. Therefore, this paper combines depression with knowledge graph to carry out research work, in order to provide assistance for the pathological research and treatment of depression.

The research content of this paper is mainly divided into three parts. The first part is the construction of the knowledge graph of depression. This part takes the biomedical literature related to depression on the medical website PubMed as the main data source, constructs the knowledge graph in a bottom-up way, and focuses on the key technologies such as knowledge acquisition, knowledge representation, knowledge fusion and knowledge storage are studied. Finally, a high-quality weighted depression

knowledge graph named DKG is constructed. The knowledge graph contains 136364 triples, of which the number of entities is 37112 and the type of relationship is 30. -- The second part is the research on drug discovery based on knowledge graph embedding. Based on DKG, taking the structural information and semantic information of knowledge graph as the starting point, this part applies the knowledge graph embedding method to the drug redirection task in the field of drug discovery, and verifies the effectiveness of this method through experiments. The third part is the design and implementation of the Intelligent Question Answering System Based on the knowledge graph of depression. This part takes DKG as the data support, adopts the KBQA method based on template, and then analyzes and designs the system requirements. Finally, an intelligent question answering system called Depression automatic Q&A System for depression is realized by using Java language and SpringBoot framework. After testing, it is found that the question answering system can effectively answer more than 70% of the questions.

Keywords: Knowledge Graph, Depression, Knowledge Graph Embedding, Drug Discovery, Question and Answer System

目　录

第1章 绪论

1.1 研究背景及意义

抑郁症（Depressive Disorder）是一种常见的精神疾病，它具有高发病率、高致残率、高死亡率和高复发率等特点，已经在世界范围内构成了严重的健康问题。抑郁症的主要症状为情绪低落、悲观厌世、思维迟钝、认知障碍等[1]。抑郁症的分类有很多，按照患病程度可分为轻度抑郁症、中度抑郁症和重度抑郁症。临床上常见的抑郁症分类有内源性抑郁症、反应性抑郁症、隐匿性抑郁症、药物引起的继发性抑郁症、躯体疾病引起的继发性抑郁症等。抑郁症每次发作的持续时间为 2 周至数年不等，并且大多数病例都有复发的倾向。长期患有抑郁症的人甚至会出现自残行为，重度抑郁症患者还可能出现自杀行为[2]。因此抑郁症不仅会给患者的家庭带来巨大的伤痛，还会带来巨大的经济负担。虽然已知抑郁症的发生跟生理、心理都有关系，但是抑郁症的发病机理目前依然不明确，故抑郁症的研究具十分重大的现实意义。

世卫组织（World Health Organization，WHO）官方网站的最新数据显示，全球抑郁症患者约有 3.5 亿，每年因抑郁症导致的自杀人数高达 100 万[3]，现如今抑郁症已成为世界第二大疾病。Huang 等人在 2019 年的一项调查报告中指出截至 2019 年 2 月，中国的抑郁症患病人数达到了 9500 万[4]。《中国国民心理健康发展报告》（2019~2020）显示，我国青少年的抑郁症检出率高达 24.6%，其中小学阶段的重度抑郁症检出率为 1.3%-3.3%，初中阶段的重度抑郁症检出率为 7.6%-8.6%，高中阶段的重度抑郁症检出率为 10.9%-12.5%，这是十分可怕的数字[5]。但我国在抑郁症的医疗防治方面还处于较为落后的地位，人们长此以往的认为抑郁症只是单纯的心情不好，并不觉得这是一种精神疾病，因此绝大多数的抑郁症患者不会去医院进行医治。因此，抑郁症已经成为了生命科学研究的热点，迫切需要利用开放领域的医学知识来促进抑郁症的研究。目前，生物医学信息的主要来源是生物医学数据库，这些数据库大多存在以下问题。首先，大多数生物医学数据库都是由人类专家从医学文献中手动提取的，这一过程耗时、费力且效率低下。其次，由于生物医学文献数量在不断增加，生物医学数据库中的信息无法及时更新。此外，生物医学数据库中的生物医学知识系统太大，对医生和研究人员来说不够直观。

因此，现在迫切的需要一个工具来对抑郁症的领域知识进行有效的整合，而

知识图谱（Knowledge Graph）就是一个不错的选择。自从谷歌 2012 年发布谷歌知识图谱（Google Knowledge Graph）[6]以来，知识图谱技术发展迅速，其理论体系日趋完善，应用效果也日益明显，已经成为大数据时代知识工程的代表性进展。此外，知识图谱也被认为是认知智能的基石，机器可以基于知识图谱进行知识推理，从而具备“理解”和“解释”的能力。故本文拟选择使用知识图谱技术来对抑郁症领域的医学知识进行整合，并对构建的知识图谱的应用进行了探究，以期能够辅助医学研究人员对抑郁症的发病机制和治疗方法的研究。

1.2 国内外研究现状

知识图谱主要用于将互联网上零零散散的知识进行整合并建立联系，从而成为知识的载体。尽管“知识图谱”一词早在 1972 年就已经出现[7]，但直到 2012 年谷歌发布谷歌知识图谱后，这一概念才再次出现在人们的视野中。传统的搜索引擎通常只能根据关键词去互联网上查找包含该关键词的 web 页面，并按照与关键词的相关程度以及网页的重要性对这些网页进行排序，然后由用户自行进行筛选，这种方式不仅效率低下，而且用户体验也不好。此外，大多数时候用户不仅仅想知道“是什么（what）”，还想知道“为什么（why）”“什么时候（when）”以及“怎么样（how）”，因此谷歌引入了知识图谱来更好地表示互联网上的非结构化、半结构化和结构化信息，使搜索引擎能够理解自然语言并作出回答，从而提升用户体验。此后，越来越多的知识图谱开始出现，比较著名的通用知识图谱有 Cyc[8]，ConceptNet[9]，Freebase[10]，DBpedia[11]，YAGO[12]，WikiData[13]，搜狗知立方，百度知心，CN-DBpedia[14]，Zhishi.me[15]、Probase[16]，Bigcilin，CN-Probase[17]、Google Knowledge Graph[6]；医学知识图谱有 Knowlife[18]，BioGrakn[19]，CMeKG[20]，新冠知识图谱 COVID-KG[21]等；多模态知识图谱有 IMGpedia[22]，Richpedia[23]等。

1.2.1 通用知识图谱研究现状

传统的知识图谱可以分为通用知识图谱和领域知识图谱。一般来说，常识类知识图谱、概念类知识图谱和百科类知识图谱属于通用知识图谱，这些知识图谱覆盖的知识面广但却不精，多用于问答系统和百科网站的数据支撑。而领域知识图谱常常是专门针对某一领域的知识图谱，里面包含的知识仅限于这一领域，但是精度很高且具有一定的深度，多用于专家系统和决策支持。

常识知识图谱中最具代表性的有 Cyc[8]和 ConceptNet[9]。Cyc 源于 MCC

（Microelectronics and Computer Technology Corporation）公司 1984 年的 Cyc 项目，它试图将人类的全部常识进行编码并建成知识库，并用一阶逻辑来表示知识库中的知识，然后在这些知识的基础上进行高效的推理。Cyc 知识库中的常识知识一般形如“大象是素食动物”、“狮子是肉食动物”。Cyc 目前包含了 700 万条断言，其中的实体概念约有 63 万个，关系有 38000 条；ConceptNet[9]源于 2004 年麻省理工的 OMCS 项目，它是一个大型的多语言常识知识库，其中包含着人们日常生活中经常使用的词语和短语以及它们之间的常识关系。目前 ConceptNet 的最新版本为 ConceptNet5.5[24]，其中包含 800 万个实体，2100 万条关系。

百科知识图谱中最具代表性的有 Freebase[10]，YAGO[12]。Freebase 始于 2005 年，由 MetaWeb 公司所研发。Freebase 的主要数据来源为 Wikipedia、NNDB、MusicBrains，然后通过众包编辑的方式由社区志愿者进行数据标注，最后以 RDF 三元组的形式进行表示。Freebase 中 24 亿个事实，涉及到的实体概念有 4400 万；YAGO 是一个多语言的百科知识图谱，由德国马克思•普朗克计算机科学研究所于 2007 年所研发，YAGO 中的知识都是经过人工评估的，准确率高达 95%，此外，YAGO 中的很多事实还加入了时间和空间信息，这是之前的知识图谱所没有的。目前 YAGO 中有超过 1000 万个实体，1.2 亿条关系。国内比较知名的百科知识图谱有搜狗知立方，百度知心，CN-DBpedia[14]，Zhishi.me[15]。搜狗知立方是由搜狗公司于 2012 所研发的一个娱乐领域的百科知识图谱，其主要数据来源为搜狗百科，并且在其中加入了推理功能，可以回答一些需要进行推理的问题；百度知心是百度于 2013 年推出的一个百科知识图谱，并将其加入到了自己的搜索引擎中；CN-DBpedia 是复旦大学知识工场实验室于 2015 年所研发的大规模中文百科知识图谱，其主要数据来源为中文百科类网站的半结构化网页，经过一系列的处理之后形成了高质量的结构化数据，前包含 1600 万个实体，2.2 亿条关系；Zhishi.me 是一个百科知识图谱，由上海交通大学所研发，目前东南大学在进行维护。Zhishi.me 致力于打造中文开放链接知识库（Chinese Linked Open Data），其数据来源为百度百科、互动百科和中文维基百科。

概念知识图谱中最具代表性的是 Probase[16]。Probase 是一个大规模的概念图谱，由微软亚洲研究院于 2012 年所研发。它的数据来源为微软搜索引擎 Bing 上的网页，从 16.8 亿个网页的语料库中自动获取了 5401933 个概念，12551613 个实例，87603947 条 IsA 三元组。与传统的知识库不同，传统的知识库认为知识非黑即白，而 Probase 的目标是实现知识计算，它支持对其包含的信息进行概率解释。Probase 目前已更名为微软概念图谱（MicroSoft Concept Graph）。国内较为知名的概念知识图谱为 Bigcilin，CN-Probase[17]。Bigcilin 是一个大规模开放域中

文概念图谱，由哈工大社会计算与信息检索研究中心所研发。2019 年，Bigcilin 升级到了 2.0 版本，目前 Bigcilin2.0 中的实体数量超过一千万，关系数量超过十六万；CN-Probase 是由复旦大学知识工场实验室所研发的大规模中文概念图谱，基本涵盖了常见实体和概念，包含约 1700 万实体、27 万概念和 3300 万 isA 关系。

综合知识图谱中最具代表性的为 Google Knowledge Graph[6]，Google Knowledge Graph 于 2012 年发布，是目前规模最大的知识图谱。谷歌公司将其应用于搜索引擎领域引起了极大地反响，被认为是搜索引擎的一次重大革新。传统的搜索引擎并不能真正理解用户意图，而有了知识图谱的加持后，搜索引擎就可以理解用户意图，从而得到准确的结果。以前，搜索“姚明的身高”，搜索引擎则会返回一些与姚明相关的网页，需要用户自己去筛选答案。现在，搜索“姚明的身高”，搜索引擎会直接返回姚明的身高为 2.26 米，并且还会返回其他与姚明相关的信息，例如姚明的生日、妻子，教育背景，职业等。

1.2.2 医疗领域知识图谱研究现状

医疗领域知识图谱是针对医学领域的知识而构建的图谱，主要有两种，一种是医学百科知识图谱，另一种是针对特定疾病的知识图谱，属于领域知识图谱。目前已有的医学知识图谱并不多，主要是因为医学知识较为复杂，并且大都存在于医学文献中，想要获取大量的高质量医学知识是十分困难的。具有代表性医学知识图谱有 Knowlife[18]，BioGrakn[19]，CMeKG[20]，新冠知识图谱 COVID-KG[21]等。

KnowLife 是一个大型的健康和生命科学知识库，由不同的网络资源通过远距离监督的方式自动构建。KnowLife 能够从不同的文本类型中获取知识，例如科学出版物、健康门户和在线社区等。KnowLife 包含超过 5 万个事实，准确率高达 93%，涉及到的关系有 13 种；BioGrakn 是一个基于知识图谱的生物医学语义数据库，存储了化合物，基因，疾病、蛋白质以及它们之间的复杂关系，并且引入了机器推理的功能。BioGrakn 将不同来源的生物医学知识进行建模、表示、聚合和集成，建立了这个药物发现知识图谱。它为医学研究人员提供了一个集成的智能数据库，并且提供了查询和推理功能用以进行药物发现；COVID-KG 是一个关于新冠肺炎的多媒体知识图谱，它的数据是从科学文献中提取的细粒度多媒体知识元素（实体及其可视化化学结构、关系和事件等）。该知识图谱能够进行问答和报告生成，并且提供了详细的上下文句子、子图和知识子图作为证据。研究人员和临床医生能够通过 COVID-KG 了解疾病机制和相关生物功能，从而促进新冠肺炎的研究。

国内较为著名的医学知识图谱主要有 CMeKG，医药卫生知识服务系统（http://med.ckcest.cn），中医药知识图谱系统（http://www.tcmkb.cn）。CMeKG 是由北京大学计算语言学研究所牵头研发的中文医学知识图谱，于 2019 年发布了 CMeKG1.0，其中包含 6310 种疾病，19853 种药物，1237 种诊疗技术及设备，涉及到的医学实体达 20 余万。目前，CMeKG2.0 也已经发布，CMeKG2.0 包含的疾病种类超过 1 万种，症状数量超过 1 万个，药物近 2 万种；医药卫生知识服务系统由中国医学科学院医学信息研究所承建，于 2013 年启动建设，2014 年正式对外服务。医药卫生知识服务系统已经发布了数个疾病领域知识图谱和药品领域知识图谱；中医药知识图谱系统是一个面向中医药领域的知识图谱系统，由中国中医科学院中医药信息研究所研发，该系统目前包含中医养生知识图谱、中医医案知识图谱、中药知识图谱等 9 个子领域知识图谱。

百科类医学知识图谱的优点在于其包含的知识范围很广，缺点是对于某种单一疾病的知识的深度有限。而领域类医学知识图谱则正相反，它所包含的知识粒度更细，且具有一定的深度。此外，医学类的领域知识图谱较为缺乏，目前还没有针对抑郁症的知识图谱，故本文尝试构建了抑郁症知识图谱。

1.2.3 多模态知识图谱研究现状

最近几年，随着知识图谱技术、自然语言处理技术和图数据库技术的不断发展，多模态知识图谱（multi-modal knowledge graph）开始出现。所谓多模态知识图谱，是指知识图谱中的知识包括多个不同的模态，传统的知识图谱一般只有一种模态（通常是文字），而多模态知识图谱中的模态可以是文字，也可以是视频、音频、图像、时间、链接等。例如，我们想知道关于姚明的一些信息，传统的知识图谱只会以文字的形式来展示姚明的一些基本信息，而多模态知识图谱不仅可以展示文字信息，还可以展示姚明的图像、说话的声音，甚至在 NBA 打球的视频等其它模态的信息。多模态知识图谱中所包含的不同模态知识可以相互补充、相互增强，在实现机器认知智能、智能问答和推荐系统领域具有极大的研究价值和应用前景。目前，已有的多模态知识图谱有 IMGpedia[22]，Richpedia[23]等。

IMGpedia 首次将语义知识图谱与多模态数据相结合，是最早的多模态知识图谱。IMGpedia 从维基百科中搜集了大量的术语及对应的图片并为每张图片构建了一个视觉实体，最终生成了约 1500 万个相应的图像描述符。同时，将这些视觉实体与维基百科中对应的文章进行关联，还与 DBpedia 中对应的实体进行了关联。此外，IMGpedia 计算了图片之间的相似度，在图片之间建立了相似性链接，其中包含 4.5 亿个视觉相似关系；Richpedia 是由东南大学漆桂林教授牵头构

建的一个大规模的多模态知识图谱。Richpedia 定义了一个更全面的视觉关系本体，其中包含了图谱实体、文本实体、图像实体及其之间的关系。Richpedia 中包含 30,638 个关于城市、景点和名人的实体，并基于这些实体从维基百科、谷歌、必应和雅虎四大搜索引擎中获取相应的图像实体从而构建出了一个多模态知识图谱。

1.3 本文的主要内容

本文的主要工作是以抑郁症领域的医学文献为主要数据来源，从零开始构建了一个质量较高的面向抑郁症的领域知识图谱，并对领域知识图谱的构建方法进行了改进。然后，基于该知识图谱和知识图谱嵌入方法进行了药物发现研究。最终，基于构建的抑郁症知识图谱和药物发现方法实现了一个面向抑郁症的智能问答系统。详细内容如下：

（1）生物医学文献数量庞大，其中蕴含着大量的知识，但却没有得到充分的利用。因此，为了充分利用医学文献中的知识，这部分工作基于生物医学文献构建了一个抑郁症的领域知识图谱 DKG。以 PubMed 上抑郁症领域的生物医学文献摘要作为数据源，借助自然语言处理工具从中抽取知识，然后采用基于规则的方法进行数据精炼和数据融合，最后对知识图谱中的数据进行了质量评估并将这些数据存入图数据库以可视化知识图谱。

（2）药物发现是目前生物医药领域最为热门的研究方之一，因此本文尝试将知识图谱应用于药物发现研究。这部分工作主要针对的是药物发现中的药物重定向任务，也即“老药新用”，以期能够辅助抗抑郁药物的研发。这一部分工作选取了知识图谱嵌入方法中最为经典的四种嵌入模型：TransE，DistMult，ComplEx 和 RotatE，以知识图谱的结构信息和语义信息为出发点，结合 DKG 进行了药物重定向实验，并通过设置实验证明了该类方法的可行性。

（3）构建知识图谱的最终目的在于应用，智能问答系统是知识图谱的一个十分重要的下游应用，因此这部分工作构建了一个面向抑郁症的智能问答系统 Depression automatic Q&A System。该问答系统以本文构建的抑郁症知识图谱为数据支撑，采用基于模板的 KBQA 方法进行构建，并取得了不错的效果。此外，该问答系统还加入了药物发现功能，即将提出的基于知识图谱嵌入的药物发现方法应用到了该问答系统中。

1.4 本文的组织结构

本文一共分为六个章节，具体组织结构如下：

第一章为绪论部分，该部分介绍了本文研究内容的背景和研究意义，并对国内外的知识图谱的研究现状进行了介绍，最后介绍了本文的主要工作和组织结构。

第二章为知识图谱构建技术概述，首先介绍了知识图谱的概念及知识图谱构建方法的分类，然后介绍了知识图谱构建过程中常用的本体构建、命名实体识别和关系抽取等关键技术，分析并对比了这些方法和工具优缺点以及存在的挑战。

第三章为基于医学文献的抑郁症知识图谱的构建部分，该部分将知识图谱构建过程分为六个步骤，分别是本体构建、数据提取、数据精炼、数据融合、质量评估和数据存储，并详细介绍了每一个步骤所遇到的问题及解决方法，最后介绍了知识图谱的构建结果。

第四章为基于知识图谱嵌入的药物发现研究部分，该部分首先介绍了 AI 在药物发现领域的应用及研究进展，以及知识图谱在药物发现领域的应用和研究进展，然后介绍了 TransE，DistMult，ComplEx 和 RotatE 四种经典的知识图谱嵌入模型，最后提出了基于知识图谱嵌入的药物发现方法，并通过实验验证了该方法的有效性。

第五章为基于抑郁症知识图谱的智能问答系统设计与实现，该部分对智能问答系统和智能问答系统构建技术进行了介绍，选取了基于模板的 KBQA 方法来构造问答系统，并且在系统中加入了药物发现功能，然后进行了系统需求分析和设计，最终实现了一个面向抑郁症的智能问答系统。

第六章为总结与展望部分，该部分首先对本文的三部分工作进行了总结，阐述了本文的主要内容、创新点以及存在的不足之处，然后对未来的研究方向提出了展望。

第2章 知识图谱构建技术概述

本章主要介绍了知识图谱构建的方法及用到的技术和工具等方面的内容。首先介绍了知识图谱的基本概念，然后介绍了知识图谱构建方法的分类，最后介绍了知识图谱构建过程中用到的技术和工具。

2.1 知识图谱概念

知识图谱的本质是一种大规模语义网络，由实体、概念、属性及其之间的关系构成，主要作用是对现存的知识进行有效整合和表示。知识图谱通常可以分为通用知识图谱和领域知识图谱，通用知识图谱一般是百科类或者常识类图谱，这类图谱覆盖的知识面很广但是不深，而领域知识图谱则是针对于某一领域构建的图谱，覆盖的知识面不广，但是具有一定的深度。知识图谱属于知识工程的范畴，从上世纪 70 年代开始主要的知识工程是知识库和专家系统，直到 2012 年谷歌正式推出谷歌知识图谱意味着知识工程进入了大数据时代。知识图谱以图的形式来对知识进行表示，并且加入了语义信息，旨在让机器“理解”知识并可以基于已有的知识进行推理，从而实现机器认知智能。因此，知识图谱一度被认为是下一代人工智能的基石。与传统的知识工程相比，知识图谱中知识的质量更高，规模也更大，还加入了语义信息，并且可以通过自动化的手段来进行构建。特别是当今时代信息爆炸，互联网上的资源十分庞杂，知识图谱自然而然地成为了最理想的知识表示方式。从工业层面来看，知识图谱现已经应用到了搜索引擎、推荐系统、智能问答和决策支持等领域以提升用户的体验。

2.2 知识图谱构建方法

知识图谱构建是从非结构化、半结构化和结构化数据出发，借助一系列自然语言处理技术通过自动化或者半自动化的方式来获取知识，并将这些知识存储到知识库以得到知识图谱的过程。知识图谱的构建是一个复杂的系统工程，会涉及本体、知识表示、自然语言处理、图数据库等技术。目前，知识图谱的构建方法大致分为自顶向下（top-down）和自底向上（bottom-up）两种。自顶向下的构建方法通常会以百科类网站等结构化数据作为数据源，然后从中提取本体和模式信

息，最后加入到知识库中，从而构建出图谱；而自底向上的构建方法则是先构建本体，然后再借助一些自然语言处理技术，从结构化、半结构化和非结构化数据中获取知识，通过人工的方式进行筛选并加入到知识库中，从而构建成图谱。由此可见，自顶向下的构建方法更适合构建百科图谱，而自底向上的构建方法更适合构建领域图谱，故本文主要围绕自底向上的构建技术进行介绍。

2.3 本体构建

采用自底向上的方式构建知识图谱需要先确定本体（Ontology），本体是描述概念及概念之间关系的概念模型[25]。本体可以清晰地描述领域中的概念以及概念间的复杂关系，并以直观的方式展现出来。本体构建的过程就是为知识图谱打地基的过程，只有明确了知识图谱中的实体概念、实体类别、实体属性和关系等信息，才能基于这些规则或者约束进行后面的构建工作。常见的本体构建工具有Protégé[26]，Ontolingua[27]，WebOnto[28]，OntoEdit[29]等，但是目前使用最多的还是Protégé，下面对 Protégé 进行一个简要介绍。

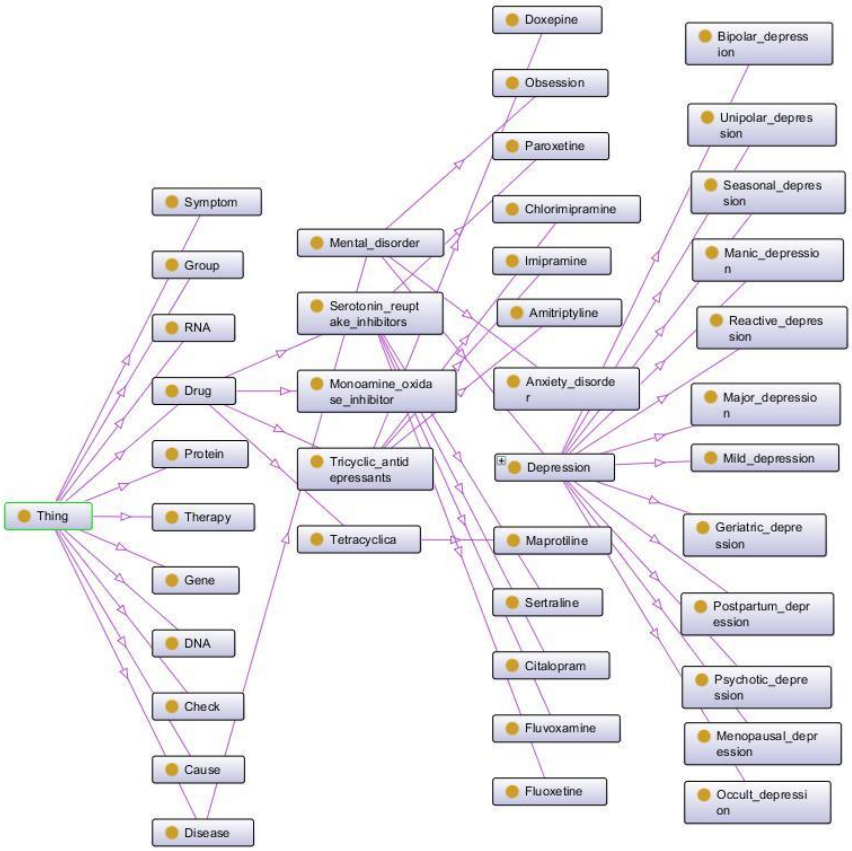

图 2-1 使用 Protégé 构建本体示例

Protégé 是一款开源的本体构建工具，由斯坦福大学医学院生物信息研究中心于 1999 年所研发。Protégé 提供本体概念类、关系、属性和实例的构建，操作简单且易于扩展。此外，Protégé 还支持中文，并且推出了网页版的 webprotege 以方便用户在线进行本体构建。图 2-1 为使用 Protégé5.0 进行抑郁症知识图谱本体构建的示例。

2.4 知识获取

知识获取的核心任务就是从不同的数据源中获取知识。数据源一般有文本、网页、视频、音频和图像等，当前绝大多数的知识图谱构建都依赖于文本和网页，只有极少量的多模态知识图谱会融入视频、音频和图像等数据。由于多模态知识图谱的研究尚处于起步阶段，故本文的研究工作仅针对传统的知识图谱。传统的知识图谱数据来源主要有结构化、半结构化和非结构化数据。通常来说，通用领域知识图谱多以半结构化的网页数据作为主要的数据来源，例如 CN-DBpedia，其主要数据来源就是中文百科类网站的半结构化网页。而垂直领域知识图谱使用的数据源会尽可能地多，以确保知识的深度和准确性，例如医学知识图谱，其主要数据来源就包括生物医学文献（非结构化）、电子病历（结构化）以及医疗网站上信息（半结构化）等。

不同的数据源往往需要用不同的方法来进行处理。一般来说，结构化数据都是由前人的工作得来的，属于质量较高的数据，无需过多的处理；半结构化数据大多来源于网页，数据质量不如结构化数据，通常使用正则表达式从网页中抽取知识；非结构化数据大多为文本数据，通常使用自然语言处理技术从中抽取知识。在现实世界中非结构化数据的数量要远多于另外两种，而本文的知识图谱也是以非结构化的生物医学文献为主要的数据源，故下面主要介绍从非结构化数据中抽取知识的方法。从文本中抽取的知识包括实体、关系和实体属性三个要素，涉及到的自然语言处理技术主要有：命名实体识别，关系抽取，属性抽取。

2.4.1 命名实体识别

命名实体识别（Named Entity Recognition）在第六届信息理解会议 MUC-6 中首次被提出，主要用来识别文本中的组织名、人名、地理位置、货币、时间和百分比表达式等[30]。命名实体识别的主要任务是从给定的文本中定位命名实体的边界，并将其分类到预定义类型的集合中。早期的方法主要依靠隐马尔科夫（HMM）模型[31]和条件随机场（CRF）模型[32][33]。随着深度学习的不断发展，

基于深度学习的命名实体识别方法得到了极大地提升，逐渐成为命名实体识别的主流方法。目前基于深度学习的命名实体识别方法中最常见的架构方式是BiLSTM-CRF[34]。

在自然语言处理中，命名实体识别被建模为序列标注问题，对于一个一维的线性输入序列 X:

$$X=x_1,x_2,\dots,x_i,\dots,x_n \tag{2–1}$$

给线性序列 X 中的每个元素打上标签集合 Y 中的某个标签：

$$Y=\{y_1,y_2,\dots,y_j,\dots,y_m\} \tag{2–2}$$

常用的数据标注方式主要有两种：BIO 和 BIOES，这里面的 B 为 Begin 的缩写，是一个命名实体开头的标识；I 为 Intermediate 的缩写，表示当前元素依然处于某个命名实体的中间；E 为 End 的缩写，表示命名实体的结尾；S 为 Single 的缩写，表示单个字符；O 为 Other 的缩写，用于标记无关字符。例如，给定句子为“张艺谋担任北京冬奥会总导演”，对这句话的标注结果为[B-PER,I-PER,E-PER,O,O,B-LOC,E-LOC,B-ORG,I-ORG,E-ORG,B-PER,I-PER,E-PER]，其中 PER 为 PERSON 的缩写，LOC 为 LOCATION 的缩写，ORG 为 ORGANIZATION 的缩写。

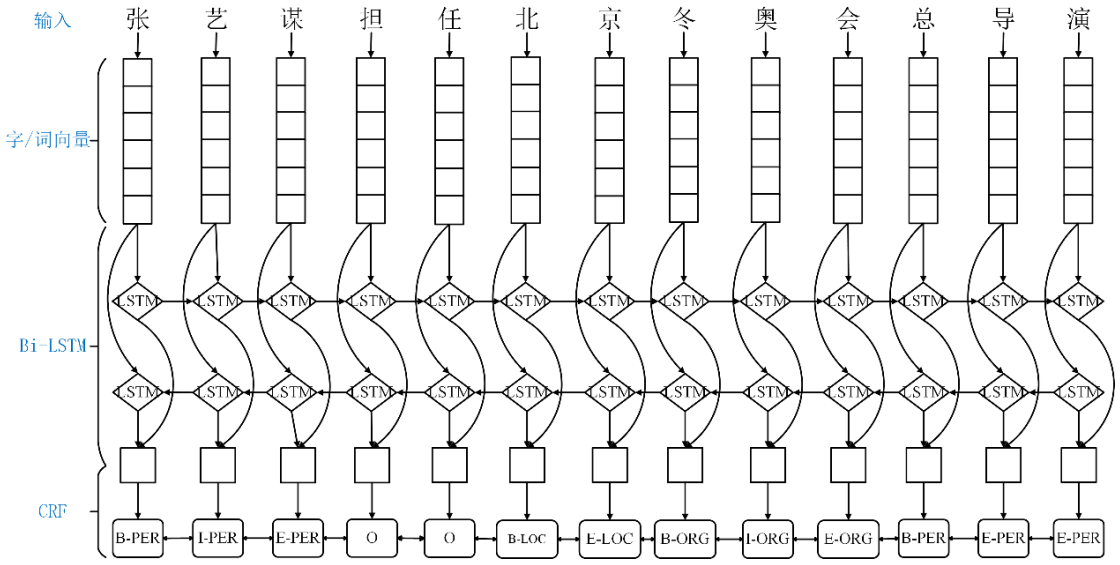

图 2-2 BiLSTM-CRF 的模型结构图

图 2-2 展示了基于 BiLSTM-CRF 的命名实体识别模型，通过 Word2Vec 等技术将输入的句子序列编码为字/词向量，然后输入到 BiLSTM 进行特征提取（编码），最后使用 CRF 进行解码得到标注序列。

此外，谷歌于 2018 年提出了大规模预训练模型 BERT[35]，是最近几年自然语言处理领域最具有突破性的一项技术。一般来说，预训练好的 BERT 模型中蕴

含着大量的通用知识，只需要借助少量的标注数据进行微调（FINE TUNE）就可以将其用于特定任务。鉴于 BERT 强大的功能，一些研究人员开始将 BERT 与传统的命名实体识别方法相结合并取得了不错的效果。BERT 与 BiLSTM-CRF 的组合模型是目前最为常见的，相比于 BiLSTM 与 CRF 的组合模型来说，BERT 与 BiLSTM-CRF 的组合模型使用 BERT 进行词嵌入，即先用 BERT 来进行语义编码，再用 BiLSTM-CRF 进行解码。图 2-3 展示了基于 BERT 与 BiLSTM-CRF 的命名实体识别模型。

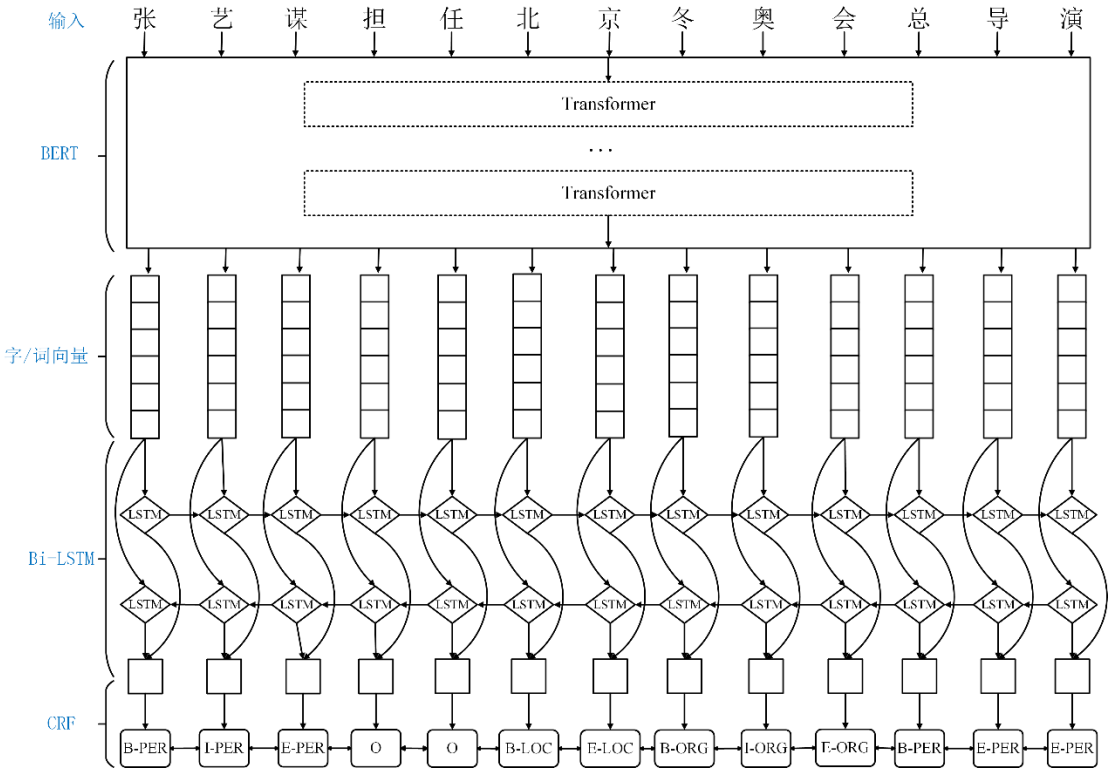

图 2-3 BERT+BiLSTM+CRF 的模型结构图

2.4.2 关系抽取

关系抽取（Relation Extraction）是信息抽取领域的重要任务之一，需要对抽取的实体对进行语义分析以解决实体间语义链接的问题，并判断两个实体之间是否存在某种关系[36]。关系抽取通常可以分为两步，第一步是判断实体对之间是否存在关系，第二步则是判断实体对之间的关系具体属于哪种关系类型。目前，关系抽取的方法主要有基于模板的方法和监督学习方法。

基于模板的方法[37]主要有两种，基于触发词/字符串的方法和基于依存句法的方法。基于触发词/字符串的方法需要人为制定一些关系模板，然后根据模板从句子中抽取关系，同时还可以限定实体对的实体类型来帮助关系抽取。基于依存

句法的方法通常需要手工构建规则，然后对句子进行分词、词性标注、命名实体识别和依存分析等处理，根据句子的依存语法树结构与手工构建的规则进行匹配，将符合规则的子树抽取出来就可以生成 SPO 三元组，再根据扩展规则对抽取的三元组进行扩展，最后根据扩展后的三元组实体和触发词进一步抽取出关系。基于模板的方法优势在于准确率高，在小规模数据集上更容易实现。缺点也很明显，需要领域专家来制定模板，而这一过程会耗费大量的人力、物力和时间精力，此外人为制定模板不能覆盖所有的关系模式，无法对开放域关系进行抽取。

基于监督学习的关系抽取方法通常分为机器学习方法和深度学习方法。机器学习方法将关系抽取建模为分类问题，根据输入的句子判断该句子中的关系属于哪种关系类型。一般来说这种方法会训练两个分类器，第一个分类器用于判断实体对间是否存在关系，如果有关系再送到第二个分类器，判断具体属于哪种关系，这样做的好处是可以排除大多数不存在关系的实体对，从而加快模型的训练过程。基于深度学习的方法分为 Pipeline 和 Joint Model 两类[38]，Pipeline 模型将实体抽取与关系抽取分为两个独立的过程，由于关系抽取依赖于实体抽取的结果，因此这种方法容易造成误差累积。而 Joint Model 会同时进行实体抽取与关系抽取，通常用模型参数共享的方法来实现，可以将实体识别和关系分类的过程共同优化从而避免误差累计。监督学习的方法面临的最大问题是，需要耗费大量的人工进行数据标注，无法对开放域关系进行抽取，模型泛化能力有限。

2.4.3 属性抽取

属性抽取（Attribute Extraction）的目标是从不同的数据源中采集特定实体的属性信息，例如实体为“姚明”，属性抽取就需要去获取姚明的身高、体重、年龄、国籍、职业、配偶等属性信息。大多数方法将属性抽取转化为关系抽取，将属性看作实体与属性值之间的一种名词性关系。最常用的属性抽取方法还是基于规则的方法和启发式算法，因为属性抽取不仅要识别属性名还要识别属性值，而实体属性十分丰富且属性值结构也不确定，并不能事先定义好属性与属性值的类型，导致监督学习方法并不能很好地发挥其作用。属性抽取目前仍然是一个巨大的挑战，虽然学术界引入了诸多深度学习的方法，但是表现依然不佳，主要还是因为实体属性太过丰富，就算是手工进行数据标注，也无法覆盖多种多样的属性。

2.5 知识表示

获取了实体、关系和属性之后，需要通过适当的方式对知识进行表示（建模），

知识图谱领域常用的知识表示方式主要有两种：基于三元组的表示和基于图的表示。

知识图谱在语义网络领域通常用 W3C 提出的 RDF[39]（Resource Description Framework）来进行表示。将知识图谱中的每一个实体对及其之间的关系表示为一个三元组，即（主体（Subject），谓词（Predicate），客体（Object））的形式，实体与实体属性也可以表示为三元组，即（主体（Subject），属性（Attribution），属性值（Value）），这样一来每一个三元组就构成了一个事实或一条知识。利用这些关系和属性就可以将大量的知识连接起来，形成一个大规模的知识库。

在实际应用中，知识图谱通常是以图的形式进行展现的，这个图可以是有向图，也可以是属性图。将三元组的头实体和尾实体作为有向图的节点，根据头尾实体间的关系构造有向边，三元组中的谓词作为有向边的标签。基于有向图的表示方式更有利于展示通过语义关联建立起来的知识图谱的全局结构。除了基本的有向图之外，还有带权重的有向图和带概率的有向图。带权重的有向图会统计每个三元组出现的次数，并将统计的数值作为有向图的权重，权重越大表明该三元组的可信度越高；带概率的有向图则是事先计算每个三元组的置信度，然后将置信度的值作为权重赋给有向边，置信度的值就是该事实三元组成立的概率。属性图也是一种常用的知识图谱表示模型，属性图的节点和边都可以添加属性，并且每个节点和边都具有唯一的标识符，因此属性图能够在表达知识图谱全局结构的同时表达更加丰富的信息。

2.6 知识存储

获取了大量的三元组后，还需要进行有效的存储才能将知识图谱更好地投入到实际应用中。目前，知识图谱的存储方式主要有三种：基于关系型数据库的存储、基于 RDF 的存储和基于图数据库的存储。

基于关系型数据库的存储方式只能存储小规模的知识图谱，通常以三元组表、属性表和垂直表的方式进行存储。在用关系型数据库存储知识图谱时，需要先确定所有的概念类型和关系，即先要定义好本体才可以方便建模。此外，如果出现了新的知识，就需要修改表结构，而这一过程将会十分麻烦，由此可见基于关系型数据库的存储方式扩展性极差。

基于 RDF 的存储会利用国际化资源标识符（Internationalized Resource Identifier，IRI）来对知识图谱中的实体、关系和属性进行标识。基于 RDF 的存储方式通常使用 Jena 和 Virtuoso 作为存储库，然后通过 RDF/XML 的方式进行序

列化以存储和传输。RDF 最大的优点在于其语义表达能力强，大部分开放的知识图谱，都是以 RDF 形式对外开放的。

基于图数据库的存储方式一般会以属性图为基本的表示形式来存储知识图谱。图数据库是一种非关系型数据库，它可以弥补传统的关系型数据库在存储知识图谱时查询复杂且缓慢的缺陷。图数据库本身提供了完善的图查询语言，而且还支持各种图挖掘算法。目前常用的图数据库软件主要有 Neo4j、Titan、OrientDB、JanusGraph、Dgraph 和 AllegroGrap 等。

对于一般规模的知识图谱，使用 RDF 和图数据库都可以进行存储，但是对于节点数高达几十亿甚至上百亿的大规模知识图谱，RDF 和图数据库也都无能为力。因此，大规模知识图谱的存储还是一个巨大的挑战。

2.7 本章小结

本章主要介绍了知识图谱的概念和构建方法，并对用到的方法和技术进行了详细的介绍。本文的知识图谱是采用自底向上的方法构建的，故首先从本体构建开始介绍，然后分别介绍了知识获取的方法、知识表示的方法和知识存储的方法，分析并对比了各种方法的优缺点以及存在的不足和挑战。

第3章 基于医学文献的抑郁症知识图谱构建

本章介绍了一种领域知识图谱的构建流程并基于该构建流程构建出了一个名为 DKG 的带权重的抑郁症知识图谱。本文采用自底向上的方法构建该抑郁症知识图谱，这个过程可分为六个部分：构建本体和关系、数据提取、数据精炼、数据融合、质量评估、数据存储。图 3-1 展示了构建抑郁症知识图谱的整个过程。首先，获取知识图谱的本体和关系。其次，使用信息抽取工具从非结构化的文本数据中提取抑郁症相关的实体、关系和实体属性。第三，使用另一种信息抽取工具对已抽取的数据进行修正。第四，进行实体对齐和实体消歧，并过滤掉噪声数据。第五，对处理后的数据进行质量评估。最后，使用 Neo4j 图数据库来存储知识图谱中的数据并可视化。

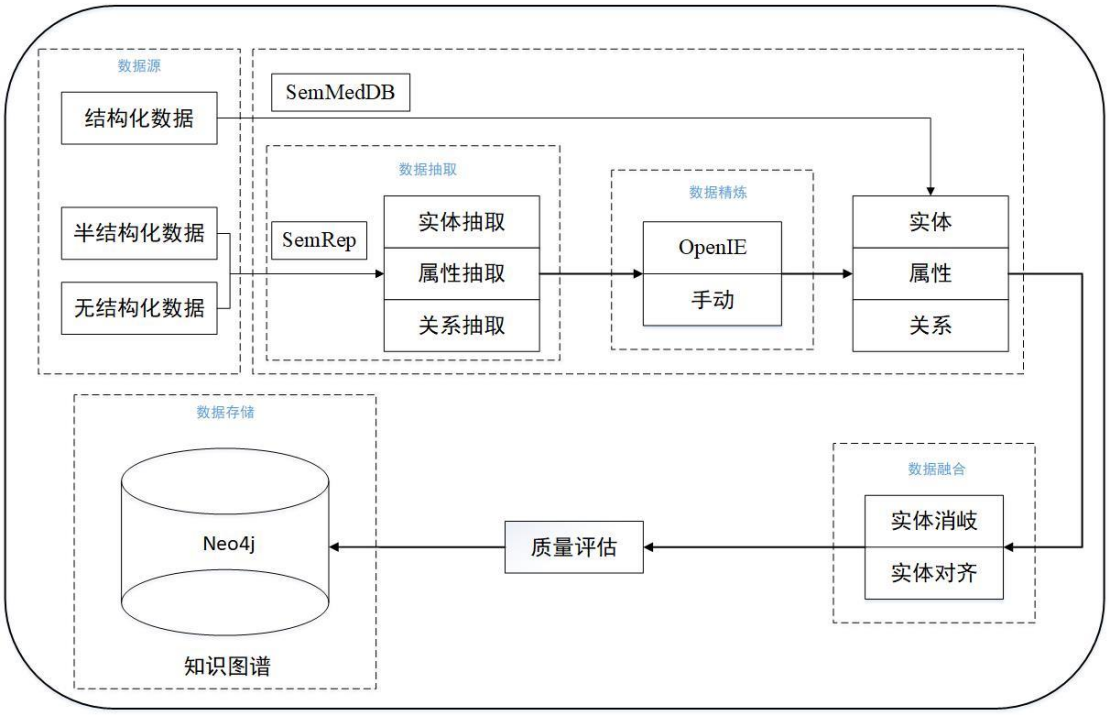

图 3-1 基于医学文献的抑郁症知识图谱构建流程图

3.1 本体构建

由第二章的分析可知，自底向上的构建方法更适用于领域知识图谱，故本文使用自底向上的方法来构建该抑郁症知识图谱。因此，构建的第一步工作就是明确知识图谱中的特定概念和内容，也即知识图谱中的实体类型、实体属性，以及

实体之间的关系。这一步工作也称为本体构建，本体一词来源于英文单词Ontology。

本文的本体构建工作主要依赖于统一医学语言系统UMLS（Unified Medical Language System）[40]。UMLS是美国国家医学图书馆（NLM）研制的一套医学语言系统，是相对早期的生物医学知识库，主要由超级叙词表（Metathesaurus），语义网络（Semantic Network）,专家词典（Specialist Lexicon）及与之相关的工具软件组成。UMLS是最大的生物医学词典集合，其中包含290万个实体和1140万个实体名称及同义词。它还包含复杂的分类，包括物理对象、事件、甚至医疗设备。然而，本文的知识图谱是关于抑郁症的，所以只需要从UMLS中选择与抑郁症相关的实体概念和关系。

通过对UMLS现有的分类结果进行筛选，总结出了本文知识图谱中的11个概念，这些实体概念见表 3-1。然后根据这些实体概念，从UMLS的语义网络中筛选关系，最终筛选出了30种关系，表 3-2中列出了这些关系类型。

表 3-1 基于医学文献的抑郁症知识图谱的实体概念表

实体类别	英文名称	含义
病因	Cause	发病的原因
检查	Check	检查某种疾病时经历的程序
疾病	Disease	疾病的名称
DNA	DNA	脱氧核糖核酸，遗传信息的携带者
药物	Drug	治疗疾病所用的药物
基因	Gene	基因，遗传因子
群体	Group	不同的患者群体
蛋白质	Protein	由氨基酸组成的高分子化合物
RNA	RNA	核糖核酸，引导蛋白质的合成
症状	Symptom	发病时的症状
治疗	Therapy	治疗疾病的医学技术

表 3-2 基于医学文献的抑郁症知识图谱的关系类型表

关系名称	含义（A→relationship→B）
ADMINISTERED_TO	A 可用于治疗 B
AFFECTS	A 对 B 会产生直接影响
ASSOCIATED_WITH	A 与 B 相关
AUGMENTS	A 可以刺激（或强化）B
CAUSES	A 会引发（导致）B
COEXISTS_WITH	A 是 B 的并发症
COMPLICATES	A 会导致 B 变得更加严重或复杂（导致不利影响）
CONVERTS_TO	A 可以转化为 B
DIAGNOSES	A 可用于诊断 B
DISRUPTS	A 会改变或影响 B 的状况、状态
HIGHER_THAN	A 高于（优于）B
INHIBITS	A 会抑制、减少、限制或阻止 B 的作用或功能
INTERACTS_WITH	A 和 B 之间存在相互作用
ISA	A 是 B 的一种，A 属于 B
LOCATION_OF	A 位于 B（某个受体位于身体的某个器官）
LOWER_THAN	A 低于（不如）B
MANIFESTATION_OF	A 是 B 的表现形式
MEASURES	A 可用于测量（或评估）B
METHOD_OF	A 是 B 的一种方法
OCCURS_IN	A 发生于 B（在一个群体中发生）
PART_OF	A 是 B 的一部分
PRECEDES	A 早于 B（发生）
PREDISPOSES	A 易患（对…有风险）B
PREVENTS	A 可以防止（或预防）B
PROCESS_OF	A 发生在 B（紊乱（disorder）发生在人体内）
PRODUCES	A 可以产生 B
SAME_AS	A 与 B 效果相同
STIMULATES	A 可以刺激，增加或促进 B 的作用或功能
TREATS	A 可以治疗 B
USES	A 使用 B（检查会使用某种医疗器械）

3.2 数据提取

本文的知识图谱是基于生物医学文献摘要构建的，这些医学文献的摘要主要来源于 PubMed[41]医学网站，是本文的主要数据来源。此外，还从 SemMedDB[42]中获取了一部分结构化的数据作为补充。针对非结构化的文本数据，本文使用信息抽取工具 SemRep[43]来从中提取实体、实体属性和关系，并将这些信息表示为初始的三元组。对于结构化的数据，可以直接使用基于规则的方法从中提取事实三元组。详细过程如下。

3.2.1 从非结构化数据中提取数据

1） 数据源

医学文献是医学信息的主要来源之一，最新的生物医学信息都出现在医学文献中。因此，本文选择使用 PubMed 网站上的生物医学文献作为非结构化数据的主要来源，从而保证本文的知识图谱能够保持最新。

PubMed 是美国国家医学图书馆（NLM）国家生物技术信息中心（NCBI）于 2000 年 4 月开发的基于网页的生物医学信息检索系统。其数据库来源为 MEDLINE[44]，该数据库收录了世界范围内的 9075 种期刊，1100 万条记录，其中 75%是英文文献。MEDLINE 中的文献涉及生物医学的各个领域，包括药学（pharmacy）、临床医学（clinical medicine）、基础医学（basic medicine）、预防医学（preventive medicine）、法医学（forensic medicine）、生物医学工程（biomedical engineering）等。

首先，需要从 PubMed 网站上下载与抑郁症相关的医学文献摘要。具体做法是以“depression”、“depressive disorder”为主要关键词在 PubMed 上进行搜索，并辅以“symptom”、“DNA”、“RNA”、“therapy”、“cause”、“gene”、“check”等词进行搜索，最终从 PubMed 网站下载了 94735 篇抑郁症相关的摘要。从 PubMed 网站下载的文献摘要包含了许多文献自身的信息，例如文献在 PubMed 网站上的文章编号 PMID，文献的出版日期 DP，文献来源 STAT 等信息。

2） 信息抽取

在获得了这些非结构化数据之后，下一步的工作就是从这些文献摘要中抽取实体、实体属性和关系。本文使用的信息抽取工具是 SemRep。

SemRep 是一个基于 UMLS 的程序，它可以从生物医学文本中中提取三部分命题（three-part propositions），称为语义谓词（semantic predications）。语义谓词

由一个主语（subject）、一个宾语（object）以及它们之间的关系（relation）组成。将这三个部分放在一起就构成了三元组。以“*Psychological therapies are effective in the treatment of depression in primary care, have longer lasting effects than drugs, are preferred by the majority of patients.*”这句话为例，SemRep 从这个句子中提取了语义谓词为 *TREATS(Psychological therapies, Depressive disorder)*，其中 *Psychological therapies* 是主语，*Depressive disorder* 是宾语，*TREATS* 是主语与宾语之间的关系。将这个语义谓词稍加改变就得到了初始的三元组（*Psychological therapies, TREATS, Depressive disorder*）。同时，为了方便后面的数据精炼步骤，将这一句话也保存在三元组中，就得到了包含句子的四元组，也即（*Psychological therapies, TREATS, Depressive disorder*，“*Psychological therapies are effective in the treatment of depression in primary care, have longer lasting effects than drugs, are preferred by the majority of patients.*”）。

图 3-2 和图 3-3 展示了 SemRep 进行信息抽取的过程。在将文本送入到 SemRep 之前，先通过预处理将摘要中的一些无关的信息删除掉，只保留 PMID、STAT 等关键信息，最后输入的内容如图 3-2 所示。

```
PMID- 28797375
OWN - NLM
STAT- MEDLINE
DCOM- 20180508
LR  - 20180508
IS  - 1558-299X (Electronic)
IS  - 0095-4543 (Linking)
VI  - 44
IP  - 3
DP  - 2017 Sep
TI  - Depression in Older Adults: A Treatable Medical Condition.
PG  - 499-510
LID - S0095-4543(17)30057-X [pii]
LID - 10.1016/j.pop.2017.04.007 [doi]
AB  - Depression is not a normal part of the aging process. Depression in older adults is
      a treatable medical condition; a variety of psychotherapeutic options are available.
      Electroconvulsive therapy is a useful treatment. Older patients must be viewed in
      their medical, functional, and social context for effective management. Cognition
      must be assessed along with mood in the older patient with depression.
```

图 3-2 SemRep 的输入示例

SemRep 处理的结果如图 3-3 所示，可以看出 SemRep 在进行信息抽取时是一句一句抽取的。它在抽取过程中，会抽取句子中的所有实体，实体所属类型以及每个实体在这段摘要中的起始位置信息。以图中第一行至第六行为例，这六行是对一句话处理的结果。第一行的第二列（以“|”分隔）为 PMID 值，第四列表明这个句子属于文章的标题部分，第五列的数字表明了当前的句子在所属部分是第几句话，第六列的 text 表明第七列是当前正在抽取的句子。第二行的前五列含义与第一行相同，第六列的 entity 表明这一行是从句子中抽取的实体，第七列是

PubMed 中该实体的编号，第八列是该实体的标准名称，第九列是该实体的类型，第十二列是该实体在文中的名称，最后两列分别是实体在这段摘要中出现的起始位置和结束位置。第三、四、五行同上。第六行的第六列的字段为 relation，表明这一行是关系抽取的结果，第二十三列就是抽取的关系。

```
SE|28797375||ti|1|text|Depression in Older Adults: A Treatable Medical Condition.
SE|28797375||ti|1|entity|C0011581|Depressive disorder|mobd|||Depression||||1000|162|172
SE|28797375||ti|1|entity|C0001792|Elderly (population group)|popg|||Older Adults||||983|176|188
SE|28797375||ti|1|entity|C0205476|Medical|ftcn|||Medical||||790|202|209
SE|28797375||ti|1|entity|C0348080|Condition|qlco|||Condition||||790|210|219
SE|28797375||ti|1|relation|1|1|C0011581|Depressive
disorder|mobd|mobd|||Depression||||1000|162|172|PREP|PROCESS_OF||173|175|2|1|C0001792|Elderly (population
group)|popg,humn|humn|||Older Adults||||983|176|188

SE|28797375||ab|1|text|Depression is not a normal part of the aging process.
SE|28797375||ab|1|entity|C0011581|Depressive disorder|mobd|||Depression||||1000|313|323
SE|28797375||ab|1|entity|C1518422|Negation|ftcn|||not||||1000|327|330
SE|28797375||ab|1|entity|C0205307|Normal|qlco|||normal||||888|333|339
SE|28797375||ab|1|entity|C0449719|Part|spco|||part||||888|340|344
SE|28797375||ab|1|entity|C1510835|Aging-Related Process|orgf|||aging process||||1000|352|365

SE|28797375||ab|2|text|Depression in older adults is        a treatable medical condition; a variety of psychotherapeutic options are
available.
SE|28797375||ab|2|entity|C0011581|Depressive disorder|mobd|||Depression||||1000|367|377
SE|28797375||ab|2|entity|C0001792|Elderly (population group)|popg|||older adults||||983|381|393
SE|28797375||ab|2|entity|C0205476|Medical|ftcn|||medical||||790|416|423
SE|28797375||ab|2|entity|C3864998|Condition:Find:Pt:^Patient:Nom|clna|||condition||||790|424|433
SE|28797375||ab|2|entity|C2346866|Assortment|cnce|||variety||||1000|437|444
SE|28797375||ab|2|entity|C0033978|Psychotropic Drugs|phsu|||psychotherapeutic||||888|448|465
SE|28797375||ab|2|entity|C1518601|Options|ftcn|||options||||888|466|473
SE|28797375||ab|2|relation|1|1|C0011581|Depressive
disorder|mobd|mobd|||Depression||||1000|367|377|PREP|PROCESS_OF||378|380|5|1|C0001792|Elderly (population
group)|popg,humn|humn|||older adults||||983|381|393

SE|28797375||ab|3|text|Electroconvulsive therapy is a useful treatment.
SE|28797375||ab|3|entity|C0013806|Electroconvulsive Therapy|topp|||Electroconvulsive therapy||||1000|496|521
SE|28797375||ab|3|entity|C3827682|Useful|qlco|||useful||||888|527|533
SE|28797375||ab|3|entity|C0001554|Administration occupational activities|ocac|||treatment||||888|534|543

SE|28797375||ab|4|text|Older patients must be viewed in       their medical, functional, and social context for effective management.
SE|28797375||ab|4|entity|C0580836|Old|tmco|||Older||||872|545|550
SE|28797375||ab|4|entity|C0030705|Patients|podg|||patients||||872|551|559
SE|28797375||ab|4|entity|C0205476|Medical|ftcn|||medical||||888|591|598
SE|28797375||ab|4|entity|C0205245|Functional|ftcn|||functional||||888|600|610
SE|28797375||ab|4|entity|C0037414|Social Environment|idcn|||social context||||1000|616|630
SE|28797375||ab|4|entity|C1280519|Effectiveness|qlco|||effective||||888|635|644
SE|28797375||ab|4|entity|C0001554|Administration occupational activities|ocac|||management||||888|645|655

SE|28797375||ab|5|text|Cognition       must be assessed along with mood in the older patient with depression.
SE|28797375||ab|5|entity|C0009240|Cognition|menp|||Cognition||||1000|657|666
SE|28797375||ab|5|entity|C2713234|Mood:-:Point in time:^Patient:-|clna|||mood||||1000|702|706
SE|28797375||ab|5|entity|C0580836|Old|tmco|||older||||872|714|719
SE|28797375||ab|5|entity|C0030705|Patients|podg|||patient||||872|720|727
SE|28797375||ab|5|entity|C0011581|Depressive disorder|mobd|||depression||||1000|733|743
SE|28797375||ab|5|relation|1|1|C0011581|Depressive
disorder|mobd|mobd|||depression||||1000|733|743|PREP|PROCESS_OF||728|732|3|1|C0030705|Patients|podg,humn|humn|||patient||||872|720|727
```

图 3-3 SemRep 的输出示例

可以看出 SemRep 在提取实体和关系的同时，还抽取了包括实体 ID、文本名称、实体类型、PMID、起始索引、结束索引、句子和数据源在内的属性信息，具体的属性信息及含义见表 3-3。

表 3-3 实体属性表

序号	属性名称	含义
1	Entity_ID	实体的编号
2	Text Name	实体在文本中的名称
3	Entity Type	实体所属类型
4	PMID	实体所属文章的 PMID
5	Sentence	实体所属的句子
6	Start_index	实体的起始索引
7	End_index	实体的结束索引
8	Source	实体所属文献的来源

3.2.2 从结构化数据中提取数据

除了非结构化的文本数据外，本文的另一数据源就是 Semantic MEDLINE Database（SemMedDB）中的结构化数据。SemMedDB 是 SemRep 抽取的语义谓词的存储库，截至 2017 年 12 月 31 日，SemMedDB 约有 9400 万条语义谓词[45]。以“depression”为关键词去 SemMedDB 数据库中进行查询，并将查询结果下载下来，然后根据之前构建的本体和关系从这些结构化的数据中筛选出所需的三元组。

3.3 数据精炼

在使用 SemRep 进行信息抽取的过程中存在一些问题，导致获取的三元组质量不高，主要有以下几个表现。第一，SemRep 在进行实体抽取时可能会丢失一些信息，导致抽取的结果不准确。例如，*pregnant women with depression* 会被抽取为 *women*，*women* 与 *pregnant women with depression* 有着巨大的差别，这种修饰词丢失的问题在 SemRep 抽取过程中普遍存在；第二，SemRep 会将一个句子中所有可能的实体及关系都进行抽取，其中有一些实体和关系与本文的知识图谱并没有关系，例如“*Objective To quantify and compare the level of depression among the students undertaking undergraduate and graduate level nursing education in Kathmandu University School of Medical Sciences, Nepal.*”，SemRep 抽取的三元组为（Kathmandu Universities，ISA，School）和（Depressive disorder，PROCESS_OF，student），显然第一个三元组是正确的，但是其中的内容与本文的知识图谱并没

有关系；第三，SemRep 抽取的一些三元组可能会不准确，因为有些句子中会包含一些不确定的词语，例如 may、maybe、probably、seems 等词语，从包含这些词语的句子中抽取的三元组很可能不是事实三元组，这就会对整体的数据造成干扰；第四，有一些医学文献的摘要中存在疑问句，如果一些三元组是从疑问句中抽取出来的，那么这些三元组可能不是事实三元组。上述这些问题将会影响知识图谱的质量，因此需要进行数据精炼。

第一个问题出现的主要原因是 SemRep 在抽取实体时会丢失一些重要信息，为了解决这个问题，本文引入了另一种信息抽取工具，名为 Stanford OpenIE[46]（以下简称为 OpenIE）。OpenIE 是由斯坦福大学自然语言处理小组所研发的一个功能强大的开放域信息抽取工具，它可以从纯文本中提取关系元组，例如（Mark Zuckerberg，founded，Facebook）。OpenIE 与其他信息抽取工具的主要区别在于，它不需要预先指定需要抽取的关系的模式。它的工作原理是首先将每个句子拆分为一组子句，然后最大限度地缩短每个子句的长度，产生一组更短的句子片段。然后将这些字句片段分割成 OpenIE 三元组，并由系统输出，图 3-4 清晰地展示了 OpenIE 工作的过程。从图中可以看出，OpenIE 只能抽取开放域的关系，也就是说它所抽取的关系都是基于句子本身的，无法按照本文定义的关系进行抽取，但是在实体抽取方面它的表现要好于 SemRep。

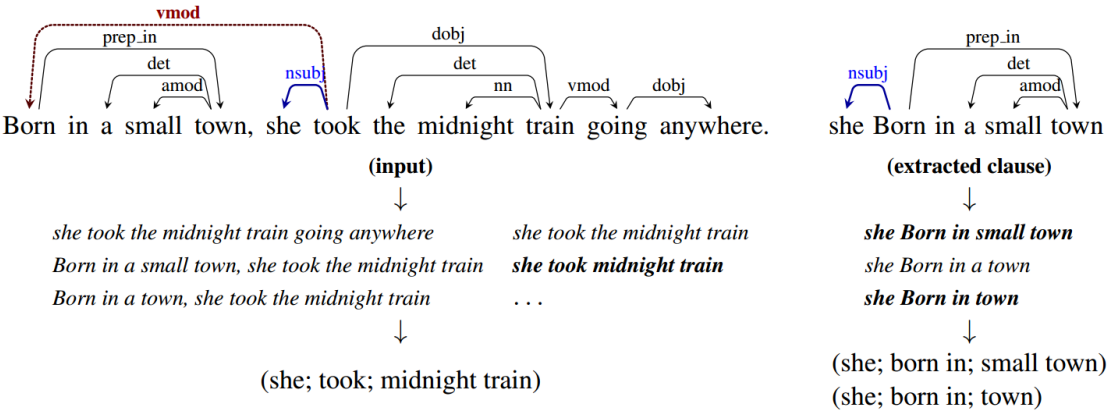

图 3-4　OpenIE 抽取示例[46]

对于同一段文本，分别使用 SemRep 和 OpenIE 进行信息抽取，以 SemRep 提取的结果为主，然后使用 OpenIE 抽取的结果对 SemRep 抽取的结果进行修正（主要是对抽取的实体进行修正）。具体过程如下：对于 SemRep 抽取的实体，如果该实体只包含一个单词，则使用 OpenIE 从原始的句子中重新抽取该实体，OpenIE 将提供许多包含该单词的候选短语，因此只需要从候选短语中筛选出最好的那一个。根据 OpenIE 实际的抽取效果，制定了以下的筛选规则：①如果候选短语以该单词开头，则必须包含介词“with”或“without”，且不能包含其他的

介词，短语中的单词数应小于或等于 4。②如果候选短语以该单词结尾，则短语中的单词数必须为 2，并且第一个单词不能是介词、量词、人称代词或着符号。③如果有多个候选短语都符合要求，则从候选短语中随机选择一个并对 SemRep 抽取的结果进行替换。④如果没有筛选出合格的短语，则不进行替换。为了验证筛选规则的有效性，对修改前后的数据进行了人工标注，用以计算修正率。在人工标注时，如果修改后的三元组中包含的信息比修改前的三元组包含的信息多，则将此条数据标注为 TRUE，否则标注为 FALSE。修正率的具体计算方法为，随机选取 1000 对修正前后的三元组，统计这 1000 条数据中标注为 TRUE 的数量，这个统计结果与 1000 的比值就是修正率。通过计算得出修正率为 84.1%，说明这一套筛选规则是有效的。

对于第二个问题，只需要将 SemRep 抽取的结果中与本文知识图谱的主题相关的部分筛选出来即可。具体来说，对于每一个三元组，如果头实体或者尾实体不属于本体构建的 11 个概念中的任何一个，则将该三元组删除。对于第三个问题，只需要设置一个简单的规则就可以将无效的三元组给剔除掉。如果一个句子中有类似于 may、maybe、probably、seems 等模棱两可的词语存在，则将从该句子中抽取的三元组全部删除。至于第四个问题，也只需制定一个简单的规则，如果一个句子是以问号结尾的，则将从该句子中抽取的三元组全部删除。

3.4 数据融合

在分别从非结构化数据和结构化数据中提取了大量的三元组后，接下来的工作是将这些三元组进行融合。由于这些三元组是从不同的文献摘要中提取出来的，故这些数据之间可能存在冗余，甚至在某些值上存在冲突，因此这些三元组不能直接进行融合。数据融合的主要问题是实体的多样性和模糊性。实体的多样性是指同一实体在文本中有不同的名称，例如 soccer 和 football，它们都可以表示足球。而实体的模糊性是指同一个实体名称在不同的文本中可以表示不同的实体，例如 Apple，它既可以表示苹果也可以表示苹果手机。为了保证知识图谱的质量，将这两部分不同数据源的数据分为三个步骤进行融合：实体对齐、实体消歧和数据去重。

3.4.1 实体对齐

这些三元组是从不同的文献摘要中抽取出来的，由于每个人的表达习惯不同，对于同一个实体可能有不同的实体名称，例如 depression 和 depressive disorder 都

表示精神疾病抑郁症，interferon-alpha 和 IFN’都表示α干扰素。SemRep 在进行抽取时会从文本中抽取实体的标准名称和文本名称（文本名称是句子中实体的名称，实体的标准名称是实体的首选名称）。本文将实体的文本名称作为实体的属性，使用实体的标准名称作为实体名称，如此一来就解决了同一实体有多个名称的问题。

3.4.2 实体消歧

在这些三元组中还存在着大量的同一个实体名称有多种不同含义的问题，特别是有很多缩写的实体名称，更容易造成歧义，例如：MD 即可以表示重度抑郁症（Major Depression），也可以表示医学博士（Doctor of Medicine）。因此，本文提出了一种基于规则的实体消岐方法。对于具有相同名称的实体，按照顺序一一比较它们的属性。如果对应的属性值一致，则认为它们代表同一实体。如果它们的属性值不一致，则将这些三元组单独保存，后续通过人工的方式检查它们是否是同一个实体。

3.4.3 数据去重

经过了实体对齐和实体消歧之后，数据质量已经得到了很大的提升了，但是还存在一个问题，那就是存在很多重复的三元组。由于本文的数据绝大多数来源于文本语料，在进行信息抽取时必然会有很多重复的三元组。如果直接将重复的三元组都删除的话，会浪费一些信息。本文的做法是先统计每个三元组出现的次数再删除重复的三元组，并将统计的数值作为权重放在边上，从而构建了一个带权重的知识图谱。类似于 Probase，在本文的知识图谱中，边上的权重越大表明该事实的可信度越高。

经过了这三个步骤之后数据中的冗余部分就被去除掉了，并且统一了存在冲突的值，至此就得到了本文知识图谱中的全部三元组。该知识图谱的三元组数目为 136364，实体数目为 37112，关系种类为 30。表 3-4 展示了知识图谱中各种实体类型数量的分布，表 3-5 展示了知识图谱中各种关系数量的分布情况。

表 3-4 知识图谱中各种实体类型的数量

实体类型	数量	实体类型	数量
Cause	35460	Group	52287
Check	5533	Protein	3696
Disease	65533	RNA	692
DNA	685	Symptom	36963
Drug	17142	Therapy	24746
Gene	29918	—	—

表 3-5 知识图谱中各种关系的数量

关系名称	数量	关系名称	数量
ADMINISTERED_TO	2700	LOWER_THAN	54
AFFECTS	13186	MANIFESTATION_OF	256
ASSOCIATED_WITH	7520	MEASURES	948
AUGMENTS	1640	METHOD_OF	541
CAUSES	4272	OCCURS_IN	955
COEXISTS_WITH	12082	PART_OF	1890
COMPLICATES	140	PRECEDES	825
CONVERTS_TO	52	PREDISPOSES	4374
DIAGNOSES	2180	PREVENTS	1507
DISRUPTS	1273	PROCESS_OF	36447
HIGHER_THAN	449	PRODUCES	377
INHIBITS	2344	SAME_AS	84
INTERACTS_WITH	5331	STIMULATES	2766
ISA	6266	TREATS	18830
LOCATION_OF	4807	USES	2268

3.5 质量评估

质量评估是知识图谱构建过程中必不可少的环节，质量评估的主要对象是知识图谱中知识的质量。从知识图谱的构建层面来看，高质量的知识图谱是知识图谱构建任务的最终目标。从知识图谱的应用层面来看，知识图谱中知识的质量高

低很大程度上决定了其在实际应用中的效果。目前，不论是通用领域知识图谱还是垂直领域知识图谱，其构建都力求做到自动化。通过自动化的方式构建知识图谱可以提升构建的效率，降低时间成本和人力成本，但也不可避免地会带来一些质量问题。

3.5.1 质量评估的维度

对于知识图谱的质量评估会涉及到知识图谱的方方面面，一般需要从准确性、一致性、完整性和实效性这四个维度去进行评估。准确性主要考察的是知识图谱中知识的准确程度，即实体，实体属性和实体之间关系的准确性，只有准确性得到了保证知识图谱才能得以有效的应用。一致性主要考察的是知识图谱中的知识表达是否一致，也即知识图谱中是否存在互相矛盾的知识。例如，知识图谱中有A、B、C三个节点，节点A为“Male”，节点B为“Female”，节点C是某一个人并且节点C同时与节点A和节点B相连，则此处必然存在错误，因为一个人不可能同时为男性和女性。完整性主要考察的是知识图谱对某一领域知识的覆盖程度，而绝对完整的知识图谱是不存在的，因为知识是在日益增长的，几乎做不到全覆盖。实效性主要考察的是知识图谱中的知识是否是最新的，知识会随着时间的变化而变化。就拿全世界抑郁症患者人数来说，这个数字每年都在发生变化，如果没有及时更新就会导致人们的认知错误，因此这种过期的知识也是一种错误，特别是一些对实效性有较高要求的知识图谱。

3.5.2 质量评估的方法

知识图谱的质量评估旨在对知识图谱中知识的质量进行量化，根据评估的结果，保留置信度较高的知识，舍弃置信度较低的知识，从而有效保证知识图谱中知识的可靠性。常见的质量评估方法有以下几种：人工抽样检测法、一致性检测法和基于外部知识的对比评估法。具体来说，人工抽样检测法是由领域专家对知识图谱中的知识进行抽样检测和评估。虽然人工检测和评估的效果更好，但是需要耗费大量的人力和时间，代价较大。在进行抽样检测时，也有多种抽样的方法，既可以使用随机抽样的方法，也可以使用不均匀随机抽样的方法。一致性检测法需要预先制定一致性检测规则，然后基于制定的规则去检测知识图谱中的知识。一致性检测法相较于人工抽样检测法的成本更低，缺陷也很明显，只能检测到规则所定义的类型的问题且实际效果取决于规则制定的好坏。基于外部知识的对比评估法的主要思想是使用与知识图谱有较高重合度的高质量外部知识源作为基

准数据来对知识图谱进行质量检测。这种方法的优点在于它可以利用已经经过人为校对过的高质量的基准数据来对知识图谱进行准确且高效的质量检测。缺点在于绝大多数的知识图谱都很难找到与其高度吻合的高质量外部知识源。

3.5.3 质量评估的结果

人工抽样检测法仍是目前使用的最多的方式，本文采用的方式也是人工抽样检测法。具体做法是随机从所有的三元组中选择 1000 个三元组，然后将这 1000 个三元组分别交给两个标注者进行标注，最后根据标注的结果计算 Jaccard 相似度[47]以评估知识图谱的质量。Jaccard 相似度是信息检索或者搜索引擎领域中常见的方法，主要用于衡量两个词库的相似性和差异性，Jaccard 相似度的值越大则两个词库的相似度越高。Jaccard 相似度的计算方法见公式(3–1)：

$$\mathrm{JACCARD}(A,B)=\frac{|A\cap B|}{|A\cup B|},(A\neq\emptyset\ \vee\ B\neq\emptyset) \tag{3–1}$$

显然，$\mathrm{JACCARD}(A,A)=1$，$\mathrm{JACCARD}(B,B)=1$；当$A\cap B=\emptyset$时，$\mathrm{JACCARD}(A,B)=0$；当$A\cap B\neq\emptyset$时，$0\leq\mathrm{JACCARD}(A,B)\leq 1$。例如，有两个词库，词库 A 的内容为“My dream is to be a doctor.”，词库 B 的内容为“She is a teacher.”，图 3-5 展示了这两个词库的交集与并集情况。若词库 A 中的单词构成的集合为$Set_A=(My,dream,is,to,be,a,doctor)$，词库 B 中的单词构成的集合为$Set_B=(She,is,a,teacher)$，则词库 A 与词库 B 的交集为$A\cap B=(is,a)$，词库 A 与词库 B 的并集为$A\cup B=(My,dream,is,to,be,a,doctor,She,teacher)$，由此可以计算出词库 A 与词库 B 的 Jaccard 相似度为 2/9。

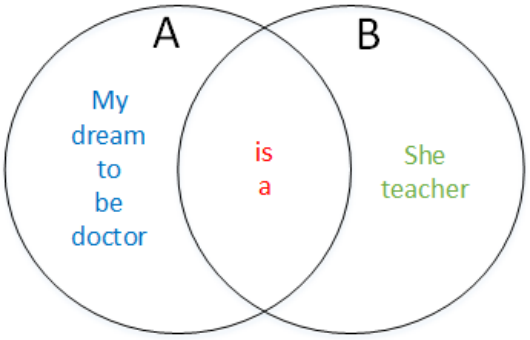

图 3-5 词库 A 与词库 B 的交集与并集图示

将抽取好的三元组交给两个标注者进行标注，对于同一个三元组，标注者 A 如果认为该三元组是正确的则将其标注为 TRUE，标注者 B 如果认为该三元组是错误的则将其标注为 FALSE。根据两个标注者的标注结果，再结合公式(3–1)就

可以计算出 Jaccard 相似度了，最终计算出本文知识图谱中知识的准确率为 72.8%。

3.6 数据存储

知识图谱得以应用的前提是知识的有效表示与存储，因此需要将知识图谱数据有效地存储在计算机中。本文选择基于图模型的存储方式进行存储，这类存储方式通常使用基于邻接表的存储方式或者基于领接矩阵的存储方式。知识图谱的实质是一个语义网络，因此使用基于图模型的存储方式更为适合。故本文选择 Neo4j[48]图数据库作为知识图谱的存储工具。

Neo4j 是一个用 Java 语言开发的图数据库管理系统，也是目前使用的最为广泛的图数据库管理系统。在 Neo4j 中，它会将属性图中的节点、属性和边都以固定长度记录的形式分别存储在不同的文件中。节点记录维护着指向其相邻的边和属性的指针，属性记录维护着指向其所对应的具体属性值，边记录维护着指向其相邻的节点和属性的指针。由于节点记录、属性记录和边记录都是固定长度的，故 Neo4j 在磁盘中读取数据时都能很快计算出偏移量，从而快速得到相应的记录。此外，Neo4j 还提供了 REST API 接口，可以轻松地集成到基于 PHP、NET、Python 和 JavaScript 的环境中。Neo4j 使用 Cypher 语言查询图数据库，Cypher 是描述性的图形查询语言，语法简单，但功能十分强大。

本文使用 Python 提供的第三方库 Py2neo 来操作 Neo4j。将所有的知识图谱数据存入到 Neo4j 中，然后通过浏览器来访问知识图谱。如图 3-6 所示，浏览器的左侧显示了知识图谱中节点和关系的一些信息，右侧的最上方有一个编辑框，在这里可以编写 Cypher 语句对数据库进行查询。例如输入语句为“MATCH (n:Disease{name:'Depressive disorder'})-[]-() RETURN n”，则可以查询出知识图谱中名为“Depressive disorder”的节点，然后展开这个节点就可以查看与该节点相邻的节点以及它们之间的关系。假如我们想知道抑郁症与职场压力之间的一些关系，就可以找到“Depressive disorder”与“Occupational Stress”这两个节点，然后查看两者之间的关系，图 3-7 中展示了这一过程。用鼠标点击连接这两个节点的边就可以查看到这个三元组的所有信息（浏览器的下方）。

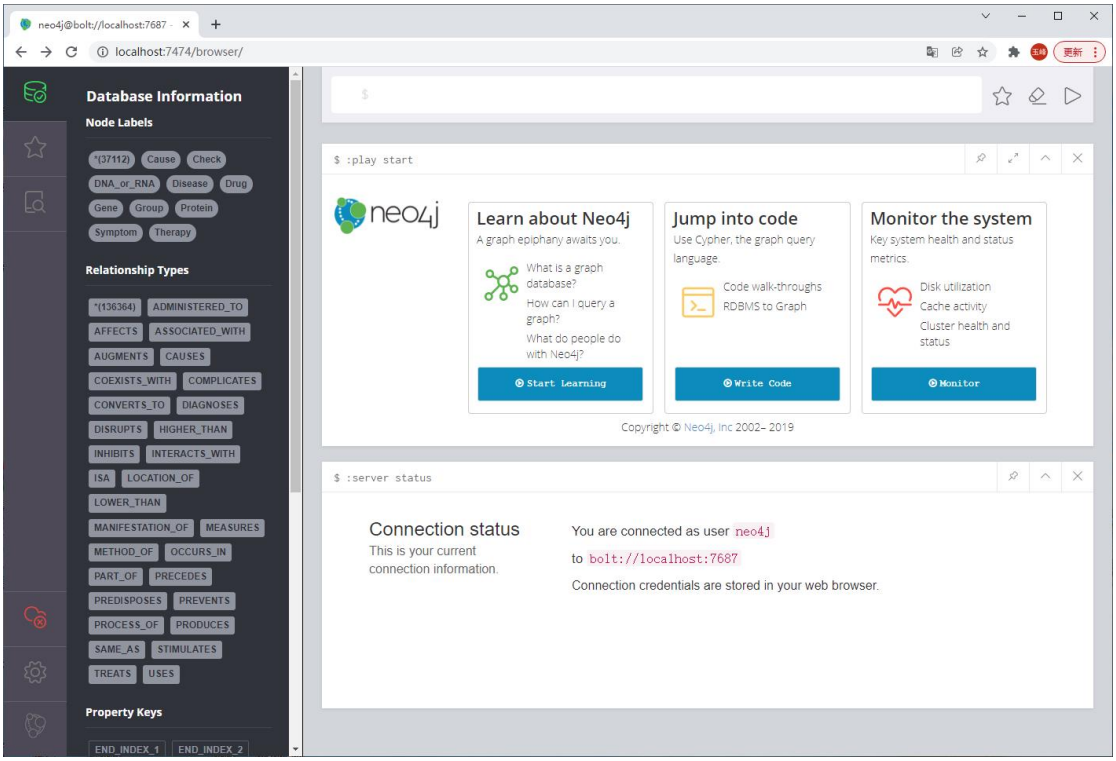

图 3-6 用浏览器访问 Neo4j

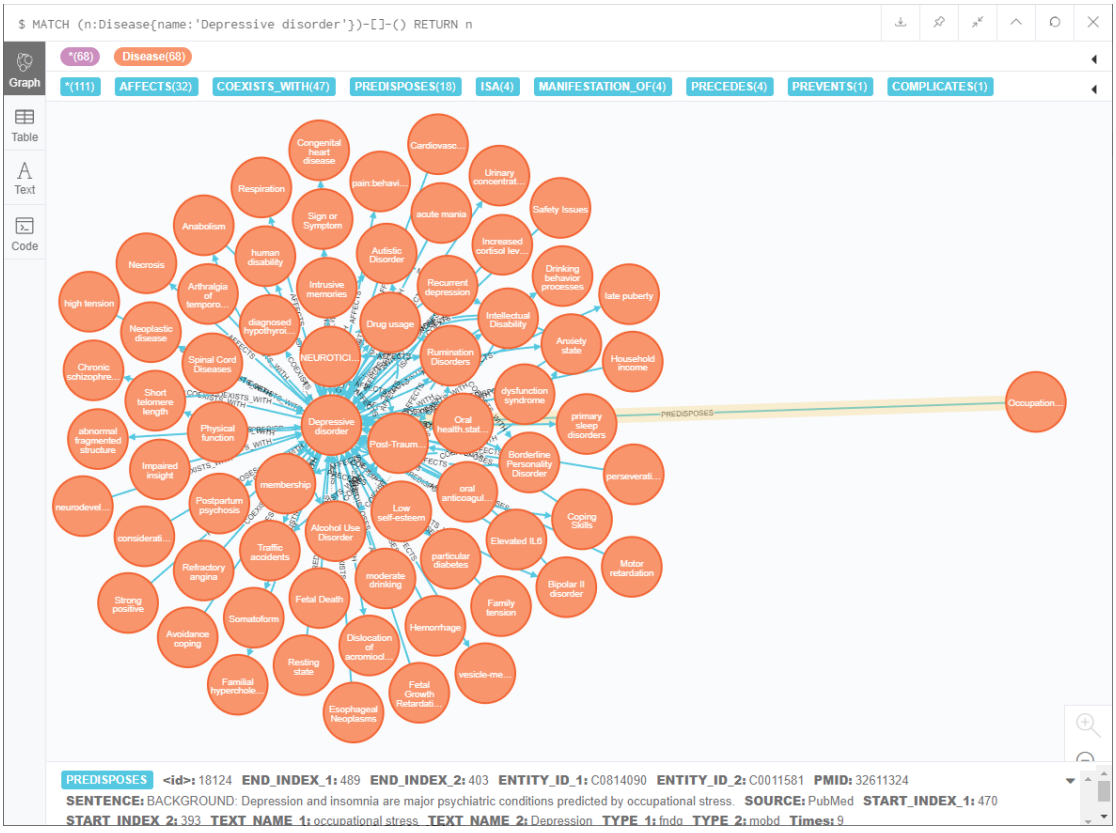

图 3-7 Cypher 语句查询实例

3.7 本章小结

本章详细介绍了基于医学文献的抑郁症知识图谱的构建过程。首先确定使用自底向上的方法来进行构建，然后从本体和关系构建开始，经历了数据提取、数据精炼、数据融合、质量评估和数据存储等步骤之后，最终构建了一个带权重的抑郁症领域知识图谱 DKG。DKG 还存在着两个不足之处，一是 DKG 并不能包含所有的抑郁症领域知识，因为抑郁症领域的医学文献数量众多，很难获取到全部的数据，二是 DKG 的数据质量还有待进一步提高。

第4章 基于知识图谱嵌入的药物发现研究

本章的主要内容为基于知识图谱嵌入技术的药物发现研究。首先介绍了基于人工智能的药物发现方法的研究现状，然后介绍了经典的知识图谱嵌入方法，最后基于本文构建的知识图谱进行了药物发现实验。由于知识图谱不仅能够保留数据的异构信息，还可以获取实体之间的语义关系，故本文的药物发现方法是以知识图谱的结构信息和语义信息为出发点的。本文选取的知识图谱嵌入模型为TransE，RotatE，ComplEx，DistMult，分别用这四种经典的模型对本文构建的知识图谱进行嵌入，然后使用训练好的模型进行药物重定向实验并证明了本文药物发现方法的有效性。

4.1 基于 AI 的药物发现

4.1.1 基于 AI 的药物发现概述

药物发现的主要任务是发现或设计新药，是医药行业中十分重要的研究领域。然而药物发现常常以“高成本，高风险，周期长，低回报”著称，根据《Nature》报道，每一款新药的平均研发成本约为 26 亿美元，同时需要耗费大约 10 年的时间，但是最终只有不到 1/10 的药物可以成功通过并上市[49]。低效率和高成本给药物发现带来了巨大的障碍，如何缩短研发周期、提高效率一直都是药物发现领域致力于解决的问题。此外，之前的一些药物发现方法主要是基于药物靶标进行的，如今药企正在快速消耗着既往发现的新药靶点，这也是阻碍药物发现的一个重要原因。

随着人工智能技术的发展，越来越多的研究人员开始将人工智能技术应用到了医药领域以加速药物发现。在 2019 年 7 月，澳大利亚弗林德斯大学的研究团队宣布利用 AI 技术研发出了一款季节性流感疫苗并进入了临床试验阶段，这是全球首个进入人体试验阶段的使用 AI 技术研制的流感疫苗，而该团队仅用了约两年的时间就研发出来了。又过了仅两个月的时间，加拿大的 Deep Genomics 公司也对外宣布了第一款由 AI 发现的治疗候选药物——名为 DG12P1 的化合物，而这一过程只花费了 18 个月的时间。由此可见 AI 在助力药物发现领域具有极大的潜能。

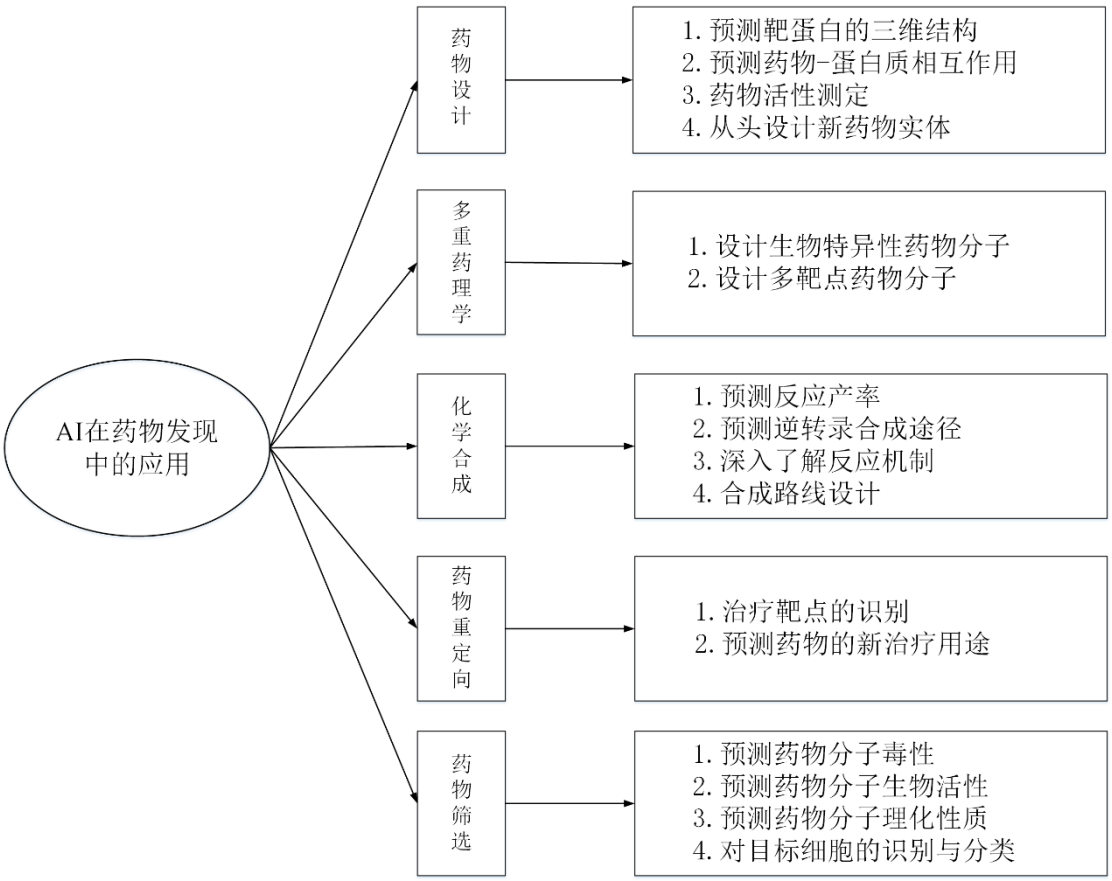

图 4-1 AI 在药物发现领域的应用

目前，已有部分机器学习和深度学习算法被用于多肽合成、虚拟筛选、毒性预测、药物监测和释放、药效团建模、定量构效关系、药物重定位、多重药理和生理活性等药物发现过程[50]。图 4-1 展示了人工智能在药物发现的各个子任务中的应用前景。在药物设计任务中，已经有研究人员引入了 AlphaFold[51]进行蛋白质的三维结构预测并取得了不错的效果，因为人类的多数疾病都是与蛋白质相关的，故在研发、设计药物分子时，需要精确地确定药物作用的靶点（蛋白质）。在药物筛选任务中，也已经有了一些人工智能工具用以预测药物分子的毒性，例如 LimTox[52]、Toxtree[53]和 DeepTox[54]，通过这些工具可以预测药物分子的毒性，有利于提前发现和预防分子的高毒性，从而提高药物的批准上市率。在化学合成任务中，Atomwise 公司研发了一个基于卷积神经网络的系统 AtomNet，该系统目前已经学习了大量的化学知识和研究数据用于化合物的筛选与合成。2015 年 AtomNet 仅用了 1 周的时间就模拟出了 2 种有希望用于埃博拉病毒治疗的化合物[55]。多重药理的任务是设计出能够与疾病相关的多个药物靶标相互作用的单一药物分子，主要针对的是一些复杂疾病，例如癌症、糖尿病、艾滋病等，目前也已经有一些机器学习算法被用于多靶配体的发现。药物重定位（也叫做药物再利

用），该任务的主要内容是基于现有的药物来治疗某些疾病，即对已有的药物进行调查研究，发现新的适应症并将其应用于治疗另一种疾病，目前已经有诸如DrugNet[56]，DRIMC[57]，DPDR-CPI[58]等用于药物重定位的工具。特别是新冠肺炎的全球大爆发以来，世界各国的研究人员都在着力于研发治疗新冠肺炎的药物，而药物重定位是效率最高的一种方式，Hooshmand 等人使用了基于神经网络的方法进行药物重定位，最终为新冠肺炎确定了 12 个最有前景的药物靶点[59]。

4.1.2 知识图谱在药物发现中的应用

上述的药物发现方法都需要依据已有的生物医学知识和研究数据来进行，而现有的生物医学论文、研究报告和实验数据的数量十分庞大且还在快速增长，此外，许多医学知识之间是孤立的，不利于研究人员进行分析。为了更好地将这些异构的生物医学知识和数据进行集成并建立关系，部分研究人员将知识图谱引入到了药物发现领域以辅助和加速药物发现。

早在 2019 年，英国 Bristol 大学的研究人员就提出过一种基于知识图谱嵌入的药物-靶标预测方法，该方法利用生物医学知识库生成药物-靶标知识图谱，再对知识图谱中的实体和关系进行表示学习得到药物和靶标的向量表示，最后预测药物与靶标的关系[60]。2020 年 4 月，英国一家名为 BenevolentAI 的公司借助生物医学知识图谱确定了一种可能用于抑制新冠肺炎感染并减少炎症损伤的药物“Baricitinib”（Baricitinib 是一种用于治疗类风湿关节炎的已获批准的药物），目前该药物已经进入了临床试验阶段。同样在 2020 年，国家超级计算广州中心联合德国阿拉丁公司及挪威奥斯陆大学团队共同开发了一个由基因、药物和疾病构成的高精度多组学生物医药知识图谱系统——PharmKG[61]，该知识图谱整合了OMIM[62]、PharmGKB[63]、DrugBank[64]等多个公共知识库，并进行了精细的数据清洗和实体属性补全，并基于该知识图谱在老药新用和靶标预测两个药物重定位的下游任务中进行了实验，均取得了不错的效果。

4.2 知识图谱嵌入表示模型

本章的工作重点是基于第三章构建的抑郁症知识图谱 DKG，结合知识图谱嵌入（Knowledge Graph Embedding）技术和链接预测（Link Prediction）方法进行药物发现研究。知识图谱嵌入可以将知识图谱中的实体和关系映射到低维向量空间中进行表示，并且能够保留知识图谱中的语义信息和图的结构信息。本章选取了四个经典的知识图谱嵌入模型进行了药物发现实验，包括 TransE[65]，

RotatE[66]，DistMult[67]和 ComplEx[68]，其中 TransE 和 RotatE 是基于翻译的嵌入模型，DistMult 和 ComplEx 是基于张量分解的嵌入模型。下面对这几个经典的嵌入模型一一进行介绍。

4.2.1 基于翻译的嵌入模型

基于翻译的嵌入模型的基本思想是将知识图谱三元组中头、尾实体之间的关系当作头、尾实体的嵌入表示之间的一种翻译。不同翻译模型之间的区别在于如何根据头、尾实体向量所在的空间定义关系向量所在的空间。目前，基于翻译的嵌入模型以模型简单、参数量少、适用于大规模数据集著称，在知识图谱补全任务上有着不错的效果。

1） TransE

TransE 提出于 2013 年，是翻译模型的开山之作。在 TransE 模型中，若三元组(h,r,t)成立，则尾实体t的嵌入向量应该近似等于头实体h的嵌入向量与关系嵌入的向量的和，如图 4-2 所示。具体来说，给定一个由三元组组成的训练集S，三元组表示为(h,r,t)，$h,t \in E$，$r \in R$，其中 E 为实体集合，R 为关系集合。TransE 的训练策略就是使$h+r$与t尽可能地相似，由此可以定义出一个距离函数：

$$\begin{aligned} d(h+r,t) &= [(h+r)-t]^2 \\ &= \|h\|_2^2 + \|r\|_2^2 + \|t\|_2^2 - 2\left(h^T t + r^T(t-h)\right) \end{aligned} \tag{4–1}$$

这里的距离函数使用的是L_2范数，也可以使用L_1范数。

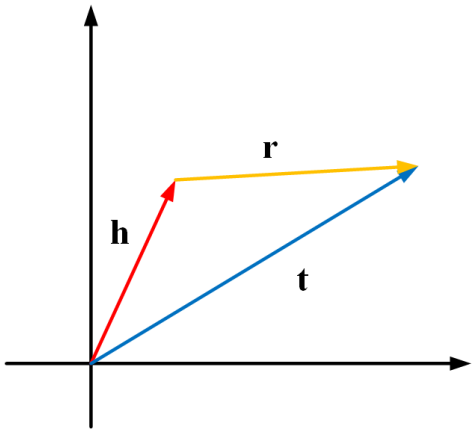

图 4-2 TransE

在训练 TransE 时，除了让$h+r$与t尽可能地相似外，对于不成立的三元组则需要让$h+r$与t的距离尽可能地远，因此需要引入负样本(h',r,t')，其中h',t'表示不属于某个三元组的实体，由此就可以构造出损失函数：

$$L = \sum_{(h,r,t)\in S} \sum_{(h',r,t')\in S'} [\gamma + d(h+r,t) - d(h'+r,t')]_+ \tag{4–2}$$

其中S是正样本集合，S'是负样本的集合，$\gamma(>0)$表示距离因子，$[x]_+$是正值函数，即$x>0$时，$[x]_+ = x$；$x \le 0$时，$[x]_+ = 0$。该损失函数的目标就是尽可能最小化$d(h+r,t)$，同时尽可能最大化$d(h'+r,t')$。

2） RotatE

RotatE 也是一种基于翻译的嵌入模型，提出于 2019 年，该嵌入模型来源于欧拉分解：$e^{i\theta} = \cos\theta + i\sin\theta$。具体来说，RotatE 将实体和关系映射到二维复空间，并将关系定义为头实体到尾实体之间的旋转。对于三元组(h,r,t)，满足$t = h \circ r$，其中$\circ$为 Hadamard 积，$|r_i| = 1$，如图 4-3 所示。则距离函数就可以定义为：

$$d_r(h,t) = \|h \circ r - t\| \tag{4–3}$$

由此距离函数就可以构造出损失函数：

$$L = -\log\sigma\left(\gamma - d_r(h,t)\right) - \sum_{i=1}^{n} \frac{1}{k} \log\sigma\left(d_r(h_i',t_i') - \gamma\right) \tag{4–4}$$

其中，γ是一个固定值，σ是 sigmoid 函数，(h_i',r,t_i')是第 i 个负采样的三元组。

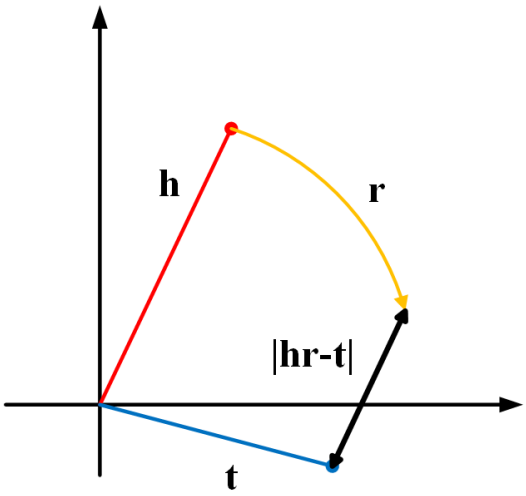

图 4-3 RotatE

4.2.2 基于张量分解的嵌入模型

基于张量分解的嵌入模型可以将关系建模为三维张量 X，而无需依赖任何关于知识图谱的信息。这类模型最早源于双线性模型 RESCAL[69]，RESCAL 将整

个知识图谱看做是一个大的三维张量，在这个三维张量中每个位置表示第 i 个实体与第 j 个实体之间是否满足第 k 个关系，如果满足则将值置为 1，否则置为 0。然后采用张量分解方法将这个三维张量分解为低维实体矩阵和低维的关系张量乘积的形式，分解公式为：

$$X_k = AR_kA^T \tag{4–5}$$

通过分解X_k来学习实体和关系的嵌入表示，整个过程如图 4-4 所示。但是，RESCAL 容易过拟合，并且复杂度会随着关系矩阵维度的增加而增加，很难应用到大规模的知识图谱上，因此研究人员又提出了 DistMult、ComplEx 等模型来解决这些问题。

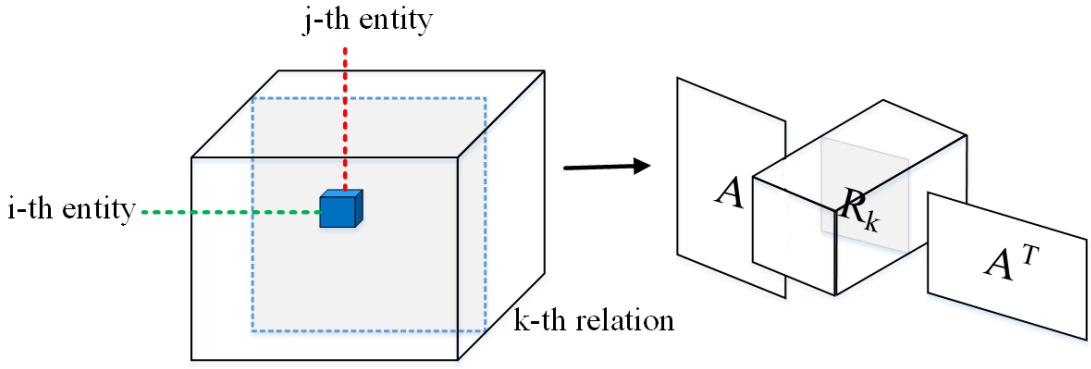

图 4-4 将知识图谱表示为三维张量并进行张量分解的过程图示

1） DistMult

DistMult 提出于 2015 年，是对 RESCAL 的改进模型。RESCAL 中实体和关系之间进行的是矩阵运算，可以捕获更深层次的语义信息，为了衡量实体与关系之间的语义相关性，定义了一个双线性函数作为打分函数：

$$f_r(h,t) = h^T M_r t = \sum_{i=0}^{d-1}\sum_{j=0}^{d-1}[M_r]_{ij}\cdot[h]_i\cdot[t]_j \tag{4–6}$$

其中，h,t为头、尾实体，M_r为关系矩阵。

针对 RESCAL 存在的问题，DistMult 进行了改进，将关系矩阵设计为对角矩阵，从而降低计算量并将参数量降低到与 TransE 相同。由此定义了新的损失函数：

$$f_r(h,t) = h^T diag(M_r)t \tag{4–7}$$

图 4-5 展示了 RESCAL 与 DistMult 的计算方式示意图。

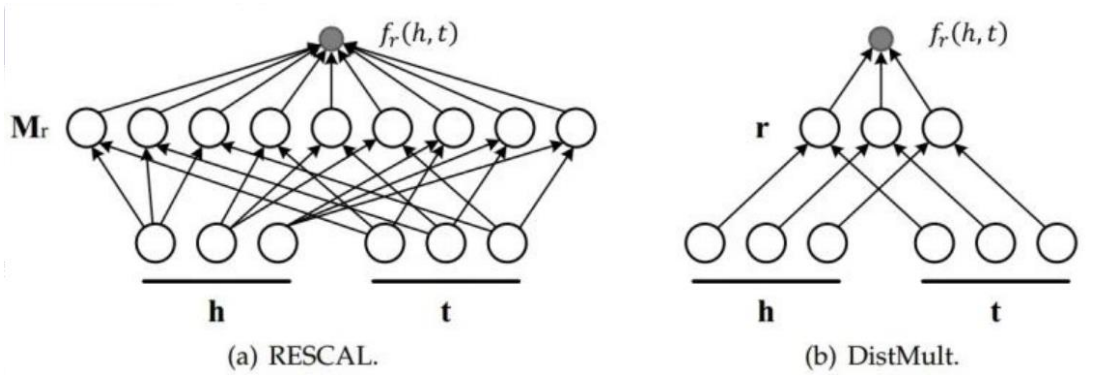

图 4-5 RESCAL 与 DistMult 的计算方式示意图

2） ComplEx

ComplEx 是对 DistMult 的扩展，因为实数向量的点积具有交换性，而知识图谱中大多数关系都是非对称的，故 DistMult 无法很好地建模非对称关系。ComplEx 将 DistMult 扩展到了复数空间，在复数空间中复值向量的实部是对称的，虚部是非对称的，而复值点积（Hermitian dot product）不具备交换性，可以更好地建模非对称关系和反关系。ComplEx 在复数空间中进行矩阵分解以得到实体和关系的嵌入表示，将复值向量点积的实部作为三元组的得分，就可以得到打分函数：

$$f_r(h,t) = Re(h^T diag(M_r)\bar{t}) \tag{4–8}$$

其中，h,t都是复数，$\bar{t}$为t的共轭复数，$Re(\cdot)$表示取复数的实部。

4.3 基于知识图谱嵌入的药物发现方法

本章的药物发现研究仅针对药物重定向中的“老药新用”任务，该任务属于知识发现的范畴，根据知识图谱中蕴含的知识和目前已有的药物去探索这些药物是否存在新的用途。而知识图谱嵌入技术可以将知识图谱中的实体和关系映射到低维向量空间，并且能够保留知识图谱中的语义信息和图的结构信息，以便于通过向量运算进行知识推理和知识发现。因此，本章尝试使用知识图谱嵌入方法进行药物重定向研究。主要思想是先用知识图谱嵌入技术将知识图谱中的实体和关系进行嵌入，给定头实体（某种疾病）和关系（治疗），借助训练好的嵌入模型进行尾实体预测，将药物重定向任务转化为了链接预测任务。

传统的药物发现方法是从药物和靶标的角度去考虑的，通过已有的药物——靶标库中药物与靶标的相互作用关系以及对应靶标与疾病之间的关系去进行药物重定向。例如，已知药物 A 可治疗疾病 B，通过查询药物——靶标库发现药物

A 可作用于靶标 C，而靶标 C 与疾病 D 有着直接的关系，因此可以推断出药物 A 也许能够治疗疾病 D，这就是传统的基于药物——靶标库的药物重定向方法。然而，这种方法存在着两大缺陷，一是药物——靶标库中药物与靶标的相互作用关系十分庞杂，难以进行一一筛选，此外一种药物与多个靶标相互作用和一个靶标与多种疾病相关的情况大量存在，进行药物筛选时会出现大量符合要求的药物，如果每一种药物都去进行医学验证的话是不现实的。二是近年来的药物发现消耗了大量的靶标，传统的方法几乎完全依赖于靶标，因此靶标的减少也会直接影响到药物重定向的效果。特别是，抑郁症目前的发病原因依然不明确，已知的与抑郁症相关的靶标数量有限，传统方法无法发挥作用。

本文的药物重定向方法不同于传统的药物重定向方法，作为非医学专业的研究人员，无法从生物医学的角度去进行研究，因此本文从自然语言的语义角度和知识图谱本身的结构信息出发进行药物重定向研究。在自然语言的语义空间中，具有相同特征的实体之间存在语义上的相似性，就以图 4-6 中 Word2vec 的嵌入结果在语义空间中的表示为例，“Man”与“Woman”之间的距离最近，因为他们之间的相似特征最多，故相似性最高，而“Dog”和“Cat”与“Man”和“Woman”同属于动物，“Grass”属于植物，故“Dog”和“Cat”距离“Man”和“Woman”比“Grass”近。而使用知识图谱嵌入技术对知识图谱中的实体和关系进行嵌入，嵌入的结果除了会包含语义信息，还会包含知识图谱本身的一些结构信息，可以借助这些丰富的信息进行药物重定向。

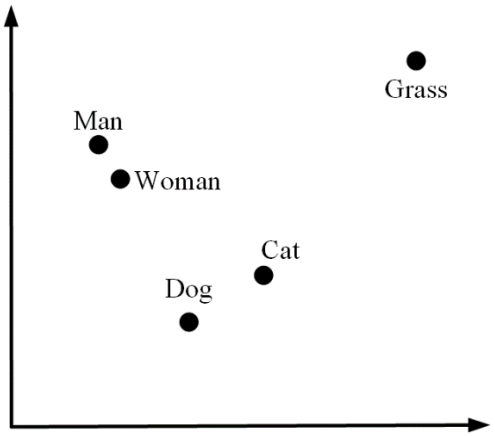

图 4-6 语义空间示例

使用知识图谱嵌入模型对知识图谱中的三元组数据进行嵌入，可以得到实体和关系的嵌入表示，以及训练好的嵌入模型和对应的模型参数。可以利用训练好的模型进行药物重定向，给定一种疾病实体（即实体类型为 Disease），将关系设定为治疗（TREATS），通过模型去预测尾实体，根据尾实体得分的高低来确定可

能的药物。以 TransE 为例，给定一种抑郁症疾病（xxx Depression）作为尾实体t，将关系r固定为治疗（TREAT），则可以根据$t+(-r)\approx h$来估算头实体h，即以尾实体t和关系r的反关系$-r$作为输入，模型将知识图谱中所有的实体依次作为头实体，通过得分函数计算每个实体的得分，并根据得分的高低预测出可能的头实体h（药物）。本文以第三章构建的 DKG 中的全部三元组为数据，分别在 TransE，RotatE，DistMult，ComplEx 四个模型上进行药物重定向实验。

4.4 实验设置及结果分析

4.4.1 实验数据

本章实验的数据为第三章构建的抑郁症知识图谱 DKG 中的全部三元组数据，并将这些三元组平均划分为十份用以进行十折交叉验证实验。需要注意的是，在进行数据划分时需要将形如（药物，治疗，疾病）的三元组筛选出来并通过随机的方式将这些三元组均分为十份，将剩余的三元组也同样通过随机的方式均分为十份，然后将两部分划分结果通过一一对应的方式进行合并，就可以得到划分好的数据。具体来说，共筛选出了 18830 个形如（药物，治疗，疾病）的三元组，通过随机的方式将这些三元组平均划分为 10 份，然后将剩余的 117534 个三元组也通过随机的方式平均分为 10 份，最后将这两部分划分的数据通过一一对应的方式进行合并。最后将这十份数据按照 8:1:1 的比例划分为训练集、验证集和测试集，这样就得到了最终的实验数据。此外，由于本章的实验是针对药物重定向的实验，故每次进行十折交叉验证的时候仅保留测试集中描述药物与疾病的三元组进行测试。

4.4.2 评价指标

本实验采用的评价指标为平均倒数排名 MRR（Mean Reciprocal Rank），Hits@1，Hits@3，Hits@10。其中 MRR 的计算方式为，给定头实体和关系，将头实体和关系的嵌入向量输入到模型中，模型会将知识图谱中的所有实体都当作尾实体来计算得分，并根据得分的高低进行排名，如果排名第一的实体与输入的头实体和关系能够组成知识图谱中真实存在的三元组，则将其值记为$1/1$，以此类推排名为 N 的实体如果预测准确的话则将值记为$1/N$，如果预测结果与头实体和关系不能组成事实三元组的话则将值记为 0，最后将所有的数值相加得到的结果就是 MRR。如公式(4–9)所示：

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i} \tag{4–9}$$

其中，Q为所有实体的数量，$rank_i$为第i个实体命中情况下的排名。MRR 的值越大，表明模型的预测效果越好。Hits@n 的计算方法为，在模型的预测结果中，如果正确结果出现在前 n 个预测结果中则计数加一，最终的计数结果与待预测三元组数目的比值即为 Hits@n，具体计算方式如公式(4–10)所示。

$$Hits@n = \frac{\sum_{i=1}^{N} x_i}{N},\quad x_i = \begin{cases} 0 & rank_i > n \\ 1 & otherwise \end{cases} \tag{4–10}$$

4.4.3 参数设置

在实验过程中，各个模型均使用各自的最优参数。对于 TransE，将 learning_rate 设置为 0.00001，batch_size 设置为 512，共设置了 80000 个 epoch，嵌入维度为 500，超参数$\gamma = 9, \alpha = 1$；对于 RotatE，将 learning_rate 设置为 0.0001，batch_size 设置为 256，共设置了 80000 个 epoch，嵌入维度为 500，超参数$\gamma = 12, \alpha = 1.5$；对于 ComplEx，将 learning_rate 设置为 0.002，batch_size 设置为 256，共设置了 80000 个 epoch，嵌入维度为 500，超参数$\gamma = 200, \alpha = 0.5$，正则项$r = 0.00005$；对于 DistMult，将 learning_rate 设置为 0.001，batch_size 设置为 512，共设置了 80000 个 epoch，嵌入维度为 1000，超参数$\gamma = 200, \alpha = 0.5$；正则项$r = 0.00005$。

4.4.4 实验环境

本章的实验环境如表 4-1 所示。

表 4-1 实验环境

名称	参数
操作系统	Microsoft Windows 10 x64
处理器	Intel(R) Core(TM) i5-9600KF CPU @3.70GHz
内存	32GB
硬盘	500GB
显卡	NVIDA GeForce GTX 1660
CUDA 版本	10.1.120
深度学习框架 PyTorch 版本	1.2.0

4.4.5 结果分析

表 4-2 十折交叉验证实验中每一折的 MRR

模型	MRR									
TransE	0.226	0.232	0.218	0.226	0.233	0.230	0.219	0.209	0.213	0.223
RotatE	0.239	0.244	0.248	0.232	0.244	0.251	0.242	0.235	0.249	0.240
DistMult	0.232	0.219	0.225	0.230	0.240	0.226	0.217	0.219	0.235	0.232
ComplEx	0.222	0.238	0.234	0.233	0.214	0.224	0.215	0.219	0.208	0.217

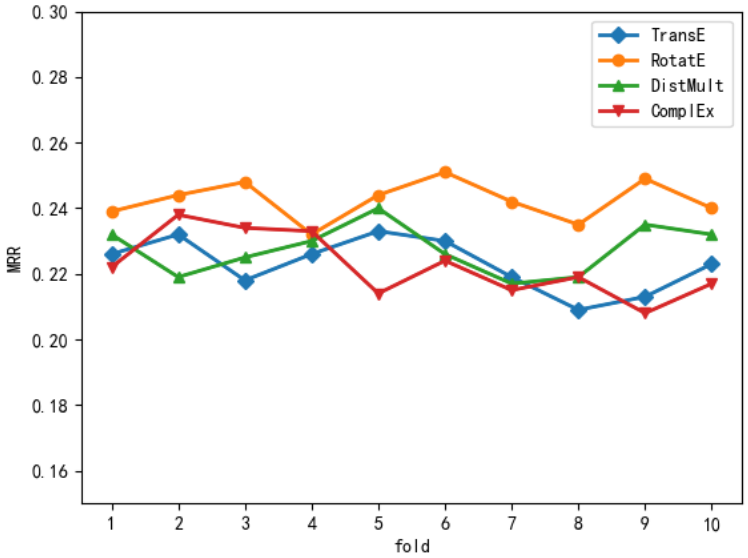

图 4-7 十折交叉验证实验中 MRR 的变化情况

表 4-2 记录了 TransE，RotatE，DistMult 和 ComplEx 各自在十折交叉验证实验中每一折的 MRR 的值，图 4-7 为根据该表中的数据绘制的折线图。从图中我们可以看出这四个模型的预测效果都比较稳定，RotatE 的 MRR 值比其它四个模型都要高，是效果最好的模型。

表 4-3 十折交叉验证实验中每一折的 Hits@10

模型	Hits@10									
TransE	0.418	0.393	0.437	0.424	0.389	0.396	0.435	0.386	0.392	0.396
RotatE	0.431	0.437	0.409	0.437	0.432	0.440	0.443	0.425	0.437	0.432
DistMult	0.377	0.368	0.378	0.364	0.381	0.386	0.369	0.379	0.380	0.375
ComplEx	0.349	0.354	0.363	0.350	0.366	0.344	0.357	0.348	0.360	0.372

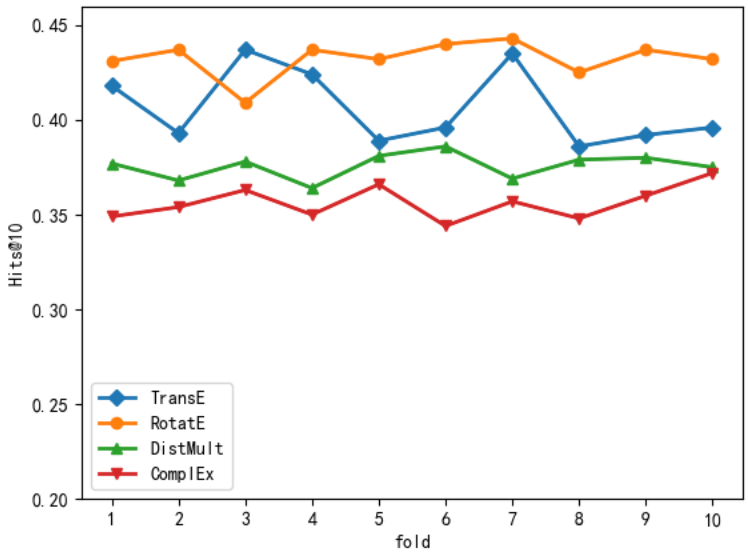

图 4-8 十折交叉验证实验中 Hits@10 的变化情况

表 4-3 记录了 TransE，RotatE，DistMult 和 ComplEx 各自在十折交叉验证实验中每一折的 Hits@10 的值，图 4-8 为根据该表中的数据绘制的折线图。从图中我们可以看出 RotatE 的 Hits@10 处在 0.41 到 0.45 之间，这意味着 RotatE 每次预测结果的前 10 个中都有 4 个以上的实体是命中的。TransE 是效果第二好的模型，但是 Hits@10 上的波动较大，它的 Hits@10 处在 0.38 到 0.44 之间，这意味着 TransE 每次预测结果的前 10 个中大约有 3~4 个实体是命中的。DistMult 的 Hits@10 处在 0.36 到 0.39 之间，ComplEx 的 Hits@10 处在 0.34 到 0.38 之间，这意味着 DistMult 和 ComplEx 每次预测结果的前 10 个中大约有 3 个实体能够命中。

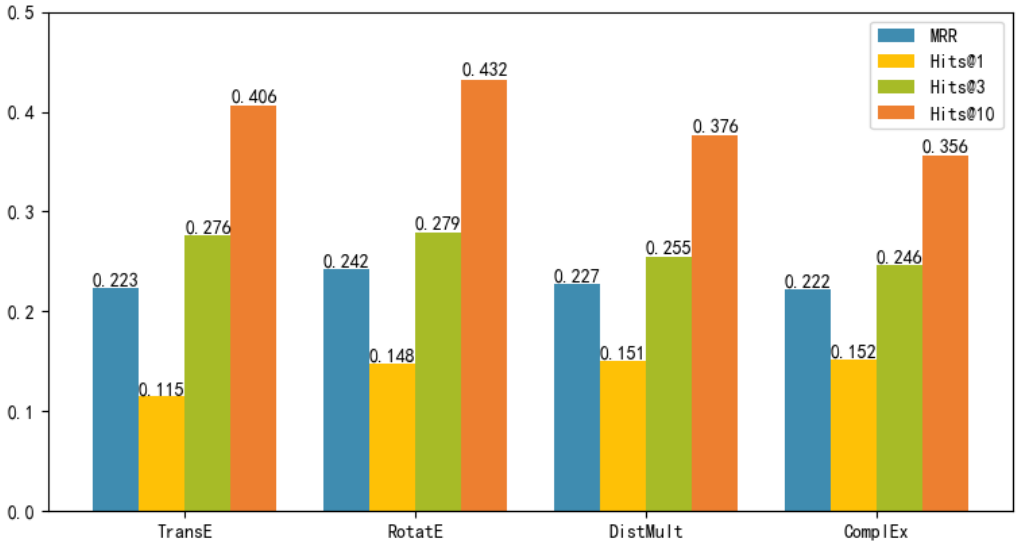

图 4-9 模型在 MRR、Hits@1、Hits@3、Hits@10 上的最终表现

图 4-9 为根据 TransE，RotatE，DistMult 和 ComplEx 在十折交叉验证实验中的最终结果绘制的柱状图，从图中我们可以看出 RotatE 在 MRR、Hits@3 和 Hits@10 三项指标上的结果均优于其他模型，在 Hits@1 上的表现也与 DistMult 和 ComplEx 不相上下。在进行药物预测时，我们需要的往往是预测结果的前 10 个，而不是排在第一的和排名前三的，因为有时候排名第一的结果并不是正确结果，甚至排名前三的结果中也没有正确的答案，故在选择药物预测模型时主要以 Hits@10 的结果为参考，因此 RotatE 比其它模型更适合用于进行药物重定向。

4.4.6 案例分析

本节选取了最后一次十折交叉验证时训练的 RotatE 模型进行案例分析，从测试集中选择一种抑郁症类的疾病进行药物预测，并对结果进行分析。此次实验以“Major Depression”为头实体，以“TREATS”的反关系作为关系（即取 TREATS 嵌入向量的反向量作为实际的关系），去预测尾实体，表 4-4 中展示了使用 RotatE 进行预测的结果。

表 4-4 预测结果

头实体	关系	尾实体
Major Depression	TREATS 的反关系	**Vilazodone** ADAM23 gene IL2RG protein Sequential Treatment **Sertraline** **Psychotherapy procedures** *Quetiapine* CCL7 protein *Oxiracetam* **Perphenazine**

这里列举了预测结果中排名前十的实体，可以看到预测正确的结果有四个，它们分别是第一个、第五个、第六个和第十个。其中第一个、第五个和第十个是治疗抑郁症的药物，而第六个是心理治疗并不是药物，之所以将它当作正确的结果是因为知识图谱中确实存在三元组（Psychotherapy procedures, TREATS, Major Depression），并且“Psychotherapy procedures”的实体类型为 Drug（本文构建的 DKG 的数据中确实存在一些错误，这里就是将心理治疗划分为了药物导致的），

故将其作为正确结果。此外，预测结果中排名第七和第九的分别是 Quetiapine 和 Oxiracetam，其中 Quetiapine 是一种治疗精神分裂症的药物，Oxiracetam 是一种治疗神经功能缺失、记忆与智能障碍的药物。由于测试集中并不存在三元组（Quetiapine, TREATS, Major Depression）和三元组（Oxiracetam, TREATS, Major Depression），故认为这两个预测是错误的，当然这两种药物也有可能能够用于治疗抑郁症，但需要从医学层面进行验证。

4.5 本章小结

本章首先介绍了人工智能在药物发现领域的应用，其次介绍了知识图谱在药物发现领域的相关研究，然后又介绍了具有代表性的嵌入模型，并根据这些嵌入模型进行了药物发现实验。本章的药物发现方法结合了语义信息和知识图谱中的知识，将药物重定向任务转换为了链接预测任务，并设计实验证明了方法的有效性。由于知识图谱中的数据本身存在一些缺陷，导致实验结果并不是很好，但还是能够证明这种药物发现方法的有效性。如果知识图谱的数据质量足够高，模型嵌入的足够好的话，是可以用于药物发现的。

第5章 基于抑郁症知识图谱的智能问答系统设计与实现

构建知识图谱的主要目的在于应用，由于知识图谱中包含着大量的知识，通常被用作知识库，为其它应用提供数据支撑，目前知识图谱最典型的下游应用有智能问答系统、推荐系统等。本章的主要内容是基于抑郁症知识图谱 DKG 和第四章的药物发现方法设计并构建了一个面向抑郁症的智能问答系统，以期为抑郁症的研究提供辅助作用。

5.1 智能问答系统概述

维基百科将问答系统（Question Answering system，QA system）定义为回答人提出的自然语言问题的系统。作为人工智能领域的重要应用之一，问答系统最早可以追溯到著名的人工智能测试——图灵测试，通过人与机器对话的方式来检测机器是否具有智能。20 世纪 60 年代，出现了一个名为 Baseball[70]的问答系统，该问答系统可以回答关于美国棒球大联盟的相关问题，并且可以正确地回答该领域中 90%以上的问题。到了 70 年代，出现了一个可以回答阿波罗探月工程中岩石的地质问题的问答系统——Lunar[71]。自此之后，各类问答系统开始不断涌现。

近年来，随着深度学习的不断发展及其在自然语言处理领域的应用，问答系统也受到了工业界的高度关注。最具代表性的问答系统是 2011 年由 IBM 研发的 Watson 机器人，该机器人在一档智力问答节目中战胜了人类冠军，并且赢得了百万美元大奖。这一事件引起了工业界对问答技术的高度关注，随后各大科技公司开始推出自己的问答系统产品。例如，苹果于 2011 年推出了个人智能助手 Siri，微软于 2014 年推出了虚拟助理 Cortana，谷歌也于 2014 年推出了个人智能助手 Google Now 等。随着智能设备的普及，问答系统已经与每个人的生活息息相关，并且涉及到了各行各业。

问答系统有多种分类，如果根据回答的问题类型可以分为事实型问答系统，是非型问答系统，对比型问答系统，观点型问答系统，对话型问答系统等，而事实型问答系统是问答系统研究和关注最多的问题类型。如果根据知识源的不同进行分类可以分为基于非结构化知识源的问答系统和基于结构化知识源的问答系统，基于非结构化知识源的问答系统主要包括单文档阅读理解（Single-document Reading Comprehension）和多文档阅读理解（Multi-document Reading Comprehension）两种方式，基于结构化知识源的问答系统包括基于关系型数据

库的问答（Relational DB oriented QA，RQA）和基于知识库的问答（Knowledge Base Question Answering，KBQA）。

然而目前的问答系统通常只能回答常识性问题，对于非常识性问题表现不佳，此外，大多数问答系统是基于信息检索的系统，无法进行知识推理，因此常常会显得不够智能。知识图谱的出现很好的赋能了问答系统，使其变得更加智能，因为知识图谱的本质是知识库，并且知识图谱天生具备语义信息，可以很好地进行知识推理，所以基于知识图谱的问答系统不仅可以回答一般性问题，还可以回答具有推理性的问题。

5.2 智能问答系统构建技术

由于本文构建的智能问答系统属于基于知识库的问答（下文中统称为 KBQA）系统，故本节介绍的构建技术主要聚焦于 KBQA 系统构建过程中用到的技术。

KBQA 系统的核心是将用户的自然语言问句转换为相关的系统查询语句，本文的 KBQA 系统使用的数据源为知识图谱，故需要将问句转换为知识图谱上的查询语句，然后根据查询结果组织答案返回给用户。这一过程包括实体链接和属性理解两个部分，实体链接主要是将从问句中识别出的实体链接到知识图谱中的相应实体上，属性理解是识别问题所对应的知识图谱子结构的过程（将问题映射到知识图谱的具体属性上）。例如，对于问句“How to treat mild depression？”，就可以通过命名实体识别技术识别出问句中的实体“mild depression”，并将其映射到知识图谱中对应的实体上。然后进行属性理解，首先识别用户的意图，由问句中的“How to treat”可知用户询问的是如何治疗，再将这一意图与知识图谱中的“Therapy”属性进行对应，就得到了可以回答该问题的知识图谱子结构，最后组织答案返回给用户。这就是 KBQA 系统的典型工作方式，图 5-1 展示了这一过程。

目前，KBQA 的研究方法可以分为三类，分别是基于模板的方法，基于图模型的方法和基于深度学习的方法。基于图模型的问答方法和基于深度学习的问答方法都需要大量带标签的语料数据进行驱动，由于该问答系统针对的是抑郁症方面的知识，目前还没有可直接用于训练的语料数据，因此本章选择使用基于模板的方法进行自动问答，这种方法也是目前使用的最广泛的方法。

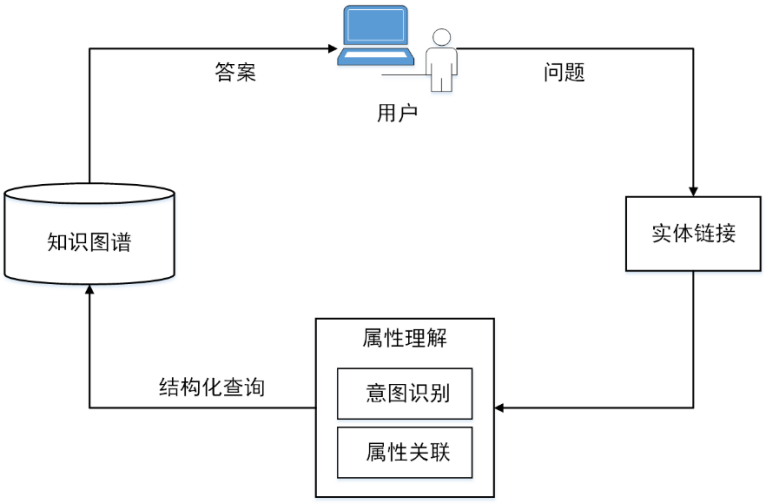

图 5-1 KBQA 系统的典型工作方式

基于模板的 KBQA 方法本质上是预先通过人工的方式定义一些问题模板，然后将用户提出的自然语言问题与模板进行匹配，从而映射为知识图谱上的系统查询语句。这种方式的优点在于，可解释性强、容易理解、具有较好的可控性，不足之处在于人工定义模板的数量有限，难以覆盖到每一种问句情形。从实际情况来看，可控性是应用落地的强需求之一，工业界往往会花费大量的人力来构造大量的模板以表达各种各样的自然语言语义。虽然基于模板的 KBQA 存在诸如表达能力有限、模板数量有限等缺陷，但是可控性更强，更适用于领域问答或者高频问答，因为领域问答和高频问答不同于开放域问答，用户的意图较为固定，通过制定模板可以尽可能多地覆盖到每种问句情形。下面对该方法中的核心步骤进行介绍。

1） 实体链接

对于用户提出的自然语言问句，首先进行命名实体识别，识别问句中的关键词（实体），然后将识别出的实体与知识图谱中的实体进行对应，找到与问句答案相关的知识图谱子结构。由于自然语言表达的多样性，用户输入的问句中的实体可能与知识图谱中的实体名称不一致，故在调用 SemRep API 进行命名实体识别之后，如果不能直接在知识图谱中查找到与之对应的实体，则需要计算识别出的实体与知识图谱中实体的相似性从而进行对应。具体来说，可以预先使用 Word2vec 对知识图谱中的实体进行嵌入表示，然后将识别出的实体也进行嵌入，最后通过余弦相似度计算相似性以确定知识图谱中对应的实体（也可以使用其它的相似度评估方法）。对于嵌入后的实体 $A(x_1, x_2, \dots, x_i, \dots, x_n)$ 与实体 $B(y_1, y_2, \dots, y_i, \dots, y_n)$，可以通过如下的公式进行计算：

$$\cos(A,B)=\frac{\sum_{i=1}^{n}x_i y_i}{\sqrt{\sum_{i=1}^{n}{x_i}^2}\sqrt{\sum_{i=1}^{n}{y_i}^2}} \tag{5–1}$$

根据计算结果取出相似度最高的实体，至此实体链接步骤就完成了。

2） 属性理解

基于模板的 KBQA 中的模板主要针对的是图 5-1 中的属性理解部分，目前主要模板有两类，一类是概念模板，另一类是句法规则。概念模板就是通过正则表达式将问句与用户意图进行匹配，例如，定义模板为“*treat*$disease*”，这一概念模板对应的用户意图为询问某种疾病的治疗方法。对于任何一个问句，如果先出现单词 symptom，后出现概念为 disease（某种疾病，例如 major depression）的实体，就可以与该模板进行匹配从而得到用户的意图。识别了用户的意图之后，还需要进行属性关联，由于本文知识图谱的属性种类不多，可以直接将定义的概念模板与知识图谱中的属性进行关联，即模板“*symptom*$disease*”对应知识图谱中的属性为 ADMINISTERED_TO 或 TREATS。最后结合上一节中的实体链接方法，就可以组织系统查询语句到知识图谱中进行查询，最后将查询的结果组织为自然语言句子返回给用户。至于句法规则，需要专家根据单词和句法来定义模板，这种方式虽然可以回答较为复杂的问题，但是需要花费更大的代价，整体效果来说与概念模板不相上下，因此本文选择使用概念模板。

通过人为的方式定义模板也存在一些缺陷，最大的问题在于能够定义的模板的数量有限，不可能覆盖所有的问句形式。为了尽可能地提高系统的问答能力，本文使用了类似于实体链接的方法进行属性理解，对于模板无法匹配的问句，通过词性标注技术找出句子中的谓词，然后与知识图谱中的属性进行相似性计算从而确定问句的属性。

3） 问答流程

介绍完上述两个步骤之后，就可以实现问答功能了。基于知识图谱的问答的实质就是根据头实体和关系去知识图谱中查找符合条件的三元组，并将符合条件的三元组组织为答案返回给用户。对于无法回答的问题，需要提供人性化提示以提升用户的体验。图 5-2 展示了整个问答的处理流程。

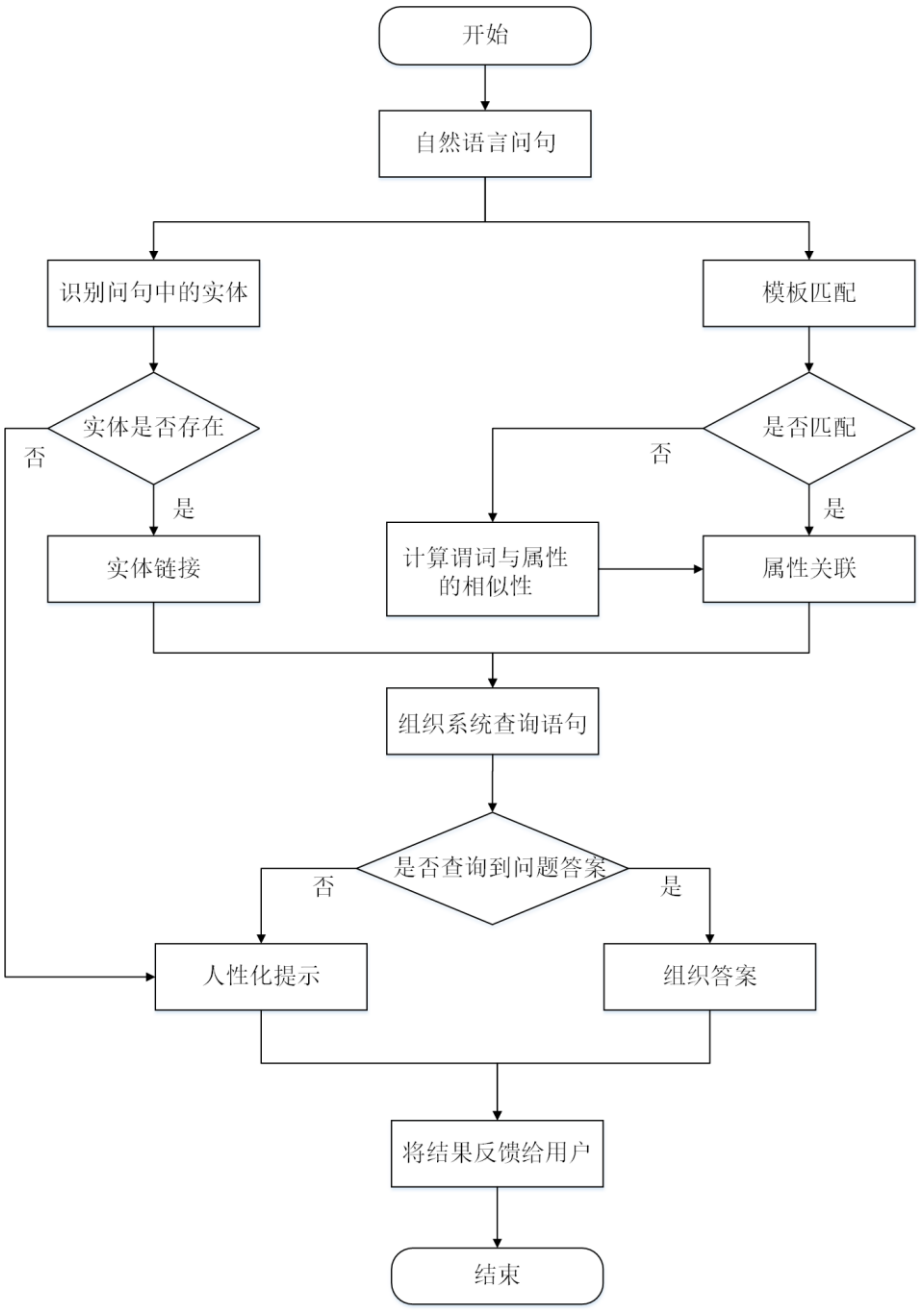

图 5-2 问答流程

5.3 系统功能需求分析

本文的问答系统仅提供针对抑郁症的英文问答服务，根据用户提出的相关问题，搜索答案返回给用户。该系统主要有两个功能模块，第一个功能模块是自动问答，回答用户提出的问题，对于不能回答的问题要做出友好的提示，这一部分在上一小节中有详细介绍。此外，在回答用户问题的同时，还需要给出答案的置

信度，以及答案所依据的文献。第二个功能模块为药物发现，基于第四章的药物发现模型构造该功能模块，构建该模块的主要目的是为了辅助医学研究人员进行药物发现研究。具体来说，用户输入某种抑郁症，系统需要给出可能的治疗药物（给出的药物为其它已知药物，非抗抑郁药），然后由研究人员自行验证。

5.4 系统设计

基于以上两个功能模块，进行了如下的系统架构设计：

图 5-3 系统架构设计

1） 数据层

本文的问答系统使用第三章构建的抑郁症知识图谱 DKG 作为数据支撑，该知识图谱的详细信息在第三章中有介绍，整个知识图谱存储在 Neo4j 中，故数据层使用的数据库为 Neo4j，对应的系统查询语言为 Neo4j 自带的 Cypher 语言。

2） 技术层

技术层中涉及的技术主要有命名实体识别、模板匹配和知识图谱嵌入，其中

命名实体识别使用的是 NLM 提供的 SemRep API，SemRep 在第二章中有介绍；模板匹配使用的是概念模板匹配的方式，需要通过人工的方式制定问题模板，5.2 小节中有详细介绍；至于知识图谱嵌入，本文最终使用的是 RotatE 模型，从第四章的实验结果来看，RotatE 的表现最好。

3） 业务层

业务层包括抑郁症知识自动问答和药物发现两个功能模块。抑郁症知识自动问答主要是根据用户提出的自然语言问句，借助技术层的基于模板的 KBQA 模块将问句转换为 Cypher 语句，然后通过服务器将查询语句发送给后端的图数据库，并根据查询结果组织答案反馈给用户。药物发现借助 RotatE 模型对知识图谱进行嵌入，并根据第四章训练的 RotatE 模型进行药物发现，用户输入某种抑郁症，RotatE 需要根据该抑郁症预测出可能的治疗药物（预测的药物为其它已知药物，非抗抑郁药）并返回给用户。

4） 交互层

交互层是直接与用户对接的可视化界面，也就是系统的前端页面。页面主要包括两个模块，一个模块用于实现问答交互，另一个模块用于进行药物发现交互。前端页面不需要太复杂的功能，只需要提供输入界面和结果展示界面，这一部分主要使 HTML5、Jquery、Ajax 等技术来实现。

5.5 系统实现

本文采用 Java 语言和 SpringBoot 框架进行系统开发，最终实现了一个网页版的抑郁症知识自动问答系统“Depression automatic Q&A System”。

5.5.1 系统运行环境

本文问答系统的运行环境为一般的台式电脑，具体的参数详情如所示。

表 5-1 系统运行环境

名称	参数
操作系统	Microsoft Windows 10 x64
处理器	Intel(R) Core(TM) i5-9600KF CPU @3.70GHz
内存	32GB
硬盘	500GB
集成开发环境	IntelliJ IDEA 2020.2.3 x64
测试浏览器	Chrome、Firefox 等

5.5.2 系统功能实现

基于上述设计构建了一个名为 Depression automatic Q&A System 的问答系统，该问答系统是以网页的形式展现的，在浏览器的地址栏输入 IP 地址和端口号即可访问，系统首页如图 5-4 所示。该系统首页有两个版块，一个版块用于进行抑郁症方面的知识问答，另一个版块用于进行药物发现。此外，该系统还给出了知识图谱的入口，用户也可以直接查看知识图谱。

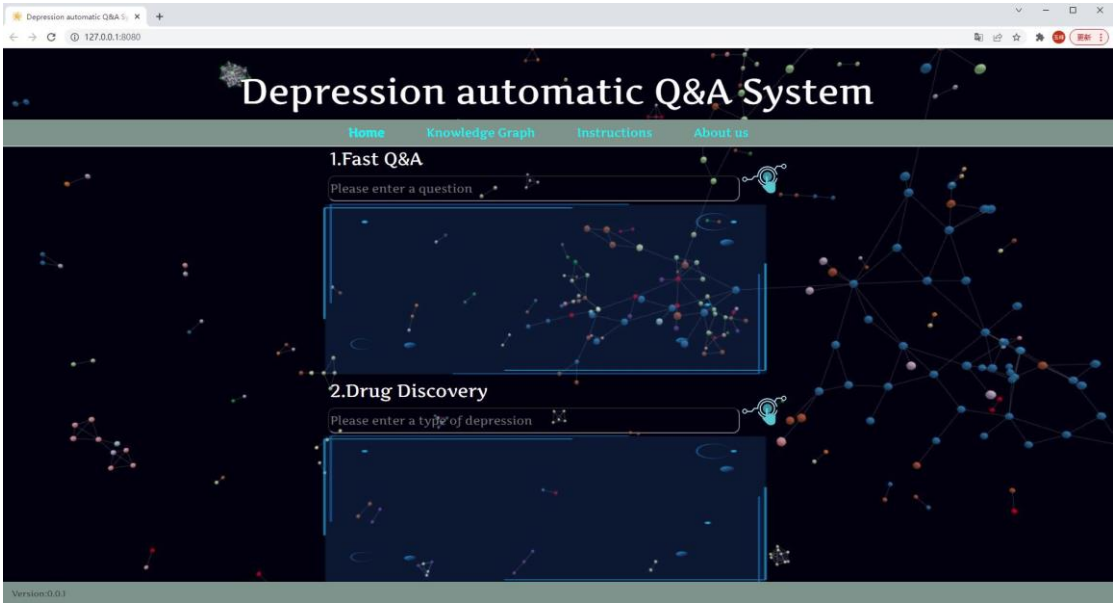

图 5-4 问答系统首页

对于自动问答功能，用户输入英文问句即可得到答案。在图 5-5 中，输入问题“What drugs can treat major depression?”，询问可以治疗重度抑郁症的药物有哪些，系统经过查询列出了 10 个置信度最高的答案。在答案中，不仅回答了能够治疗重度抑郁症的药物的名称，还给出了每种药物与重度抑郁症构成的三元组（药物，治疗，重度抑郁症）出现的频次（本文构建的是带权重的知识图谱，权重就是频次，第三章中有介绍），并且根据频次给出了该答案的置信度，最后还

给出了该三元组的来源，用户可以点击 PMID 去对应的文献中进行查看。图 5-6 是另一个问答示例，询问轻度抑郁症的症状，同样也是给出了置信度最高的十种症状，以及每种症状出现的频次、置信度和文献的 PMID。此外，对于系统不能回答的问题或者无法理解的问题，还需有进行友好的提示，如图 5-7 所示。

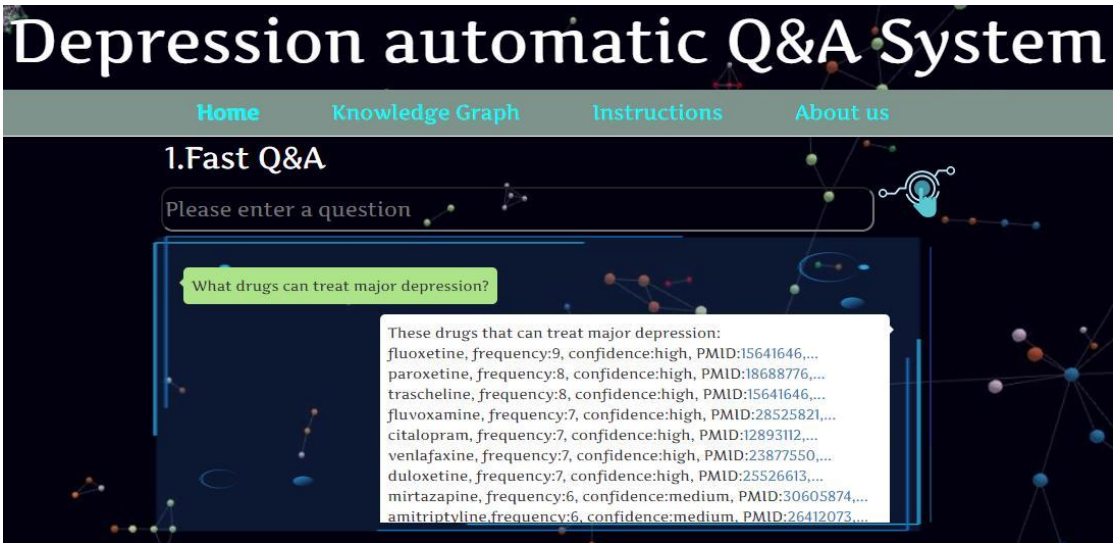

图 5-5 问答示例 1

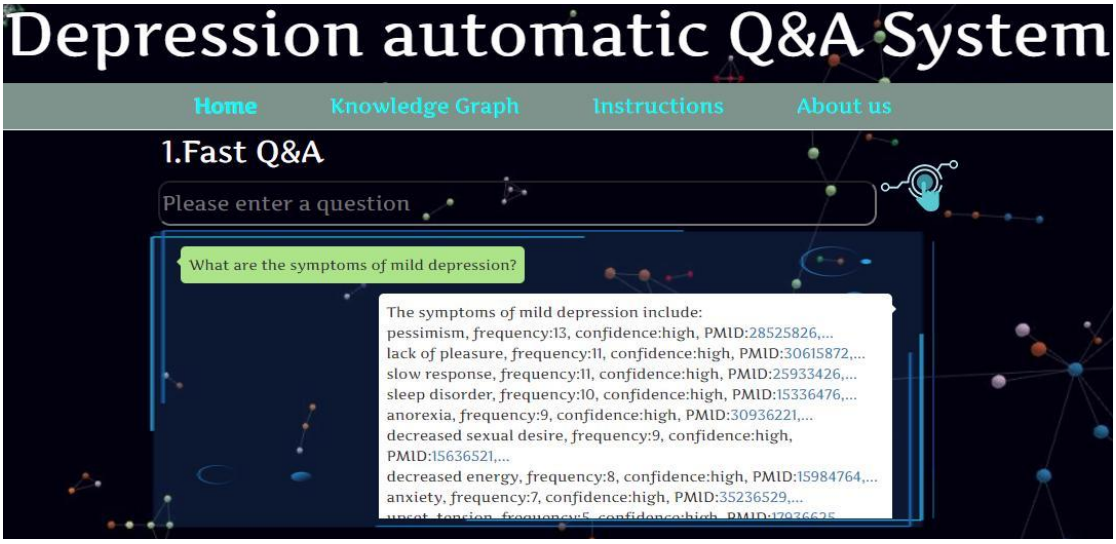

图 5-6 问答示例 2

对于药物发现功能，用户只需要输入某种抑郁症的名称，系统就可以给出可用于治疗该类抑郁症的潜在药物的预测。图 5-8 展示了该功能，用户输入了“Postpartum Depression”（产后抑郁症），系统经过预测和筛选最终给出了四种可能性较大的药物，这四种药物分别是 cinromide（一种抗惊厥药），regorafenib（一种抗癌药物），vilazodone（一种新型抗抑郁药，本文知识图谱中 vilazodone

与 Postpartum Depression 之间不存在关系），phenobarbital（一种镇静、催眠、抗惊厥药）。至于这些药物是否真的能够用于治疗产后抑郁，还需要专业的医学研究人员去进行验证，故该功能目前仅作为测试功能。

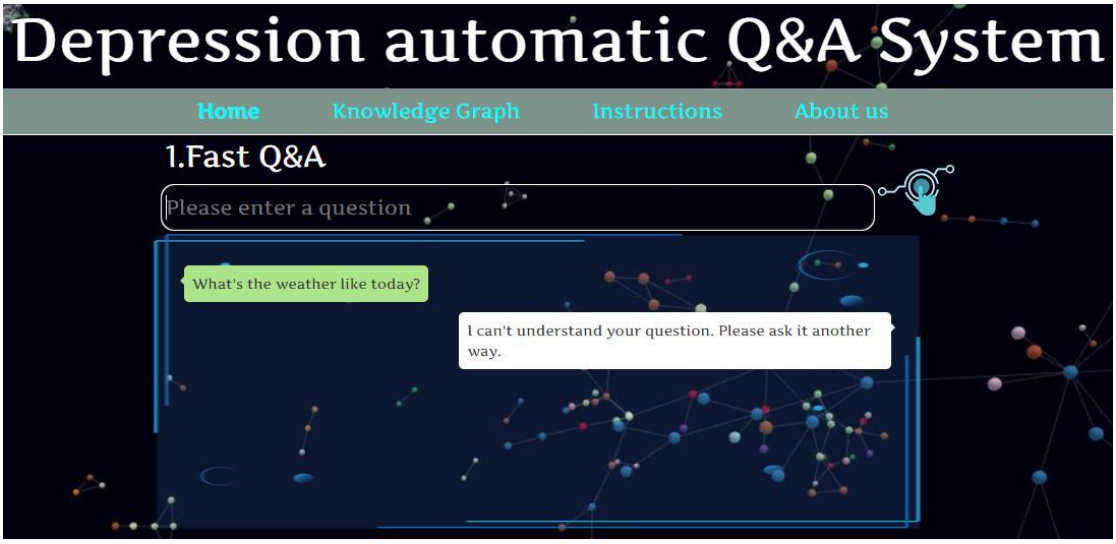

图 5-7 问答示例 3

图 5-8 药物发现示例

5.5.3 系统性能评估

为了更好地评估系统性能，本文征集了五位志愿者，分别从“What”、“How”、“Why”三个角度对系统进行提问，并通过人工测评的方式对系统的回答进行评估。之所以从这三个角度进行提问是因为这三类问题是用户提问最多的，通过这三种问法可以解决用户的大多数问题。本文针对每一类问题各收集了 50 个进行评估，评估结果如表 5-2 所示。从表中可以看出，系统对于“What”类型的问题表现最好，对于“Why”类型的问题表现最差，总体来说系统的平均准确率达到了 70%，说明系统能够回答用户的大部分关于抑郁症的问题。至于有一部分没能

正确回答的问题，可能有以下几个原因。一是命名实体识别不准确，不能很好地与知识图谱中的实体进行链接；二是用户提问过于复杂，无法很好的进行属性理解。三是知识图谱中不包含问题的答案，因为本文构建的知识图谱本身就存在一小部分不准确、甚至是错误的数据。这三方面的原因都有可能导致系统无法正确地回答用户所提的问题。

表 5-2 系统评估结果

问题类型	问题数量	准确率
What	50	76%
How	50	72%
Why	50	64%
平均	—	70.7%

5.6 本章小结

本章首先介绍了智能问答系统及其构建技术，然后进行了系统设计，最终实现了一个抑郁症知识自动问答系统，此外还根据第四章中的药物发现模型加入了药物发现功能。虽然经过测试发现，该问答系统能够有效地回答大部分的问题，但是对于复杂问句表现不好。至于药物发现功能，虽然从实验层面证明了该药物发现方法的有效性，但是实际的预测结果还需要专业的医学研究人员进行验证，故该功能目前还处于测试阶段。

第6章 总结和展望

6.1 总结

面对抑郁症目前的严峻形势，以及生物医学文献数量过于庞大无法有效地进行利用的问题，本文从计算机技术的角度出发，将知识图谱应用到了医学研究领域以辅助抑郁症的研究。总体来说，本文的主要工作可以分为三个部分，第一部分的工作为抑郁症知识图谱的构建，第二部分的工作为使用知识图谱嵌入技术进行药物发现，第三部分的工作结合了前两部分的工作内容设计并实现了一个抑郁症知识的智能问答系统。

在第一部分的工作中，本文提出了一套完整的自底向上的知识图谱构建流程，整个构建流程分为：构建本体和关系、数据提取、数据精炼、数据融合、质量评估、数据存储六个步骤，并基于该流程进行了抑郁症知识图谱的构建。首先从PubMed 网站上获取了 94735 个生物医学文献摘要做为该知识图谱的主要数据来源，然后采用自然语言处理技术对这些非结构化的文本数据进行处理，最后按照构建流程进行了构建，并对构建过程中每一个步骤的任务以及使用的技术进行了详细的介绍。最终构建了一个包含 136364 条三元组，37112 个实体，30 种关系的带权重的抑郁症知识图谱，并且通过质量评估方法计算出该知识图谱中数据的准确率 72.8%。

在第二部分的工作中，本文以知识图谱的结构信息和语义信息为出发点，尝试使用知识图谱嵌入技术进行了药物重定向实验，实验结果证明了该方法的有效性。构建知识图谱的核心目的在于应用，将知识图谱中的实体和关系嵌入到向量空间中然后进行知识推理和知识发现是目前知识图谱领域最热门的应用方式之一。此外，药物发现是目前生物医药领域中的热门研究方向，特别是近几年新冠疫情的肆虐，让人们意识到了缩短药物的研发周期、提高药物研发效率的重要性。故本文选取了 TransE，RotatE，ComplEx 和 DistMult 四种经典的知识图谱嵌入模型，基于构建的抑郁症知识图谱进行了抗抑郁药物发现实验，并通过实验结果证明了该药物发现方法的有效性，同时还发现 RotatE 是表现最好的模型。

在第三部分的工作中，本文结合了前两部分的工作内容，构造了一个抑郁症知识自动问答系统。智能问答系统是知识图谱的一个典型下游应用，而目前基于知识图谱的医学类问答系统还很少，故本文构建了一个抑郁症领域的智能问答系统，一方面希望该问答系统可以向普通用户普及抑郁症方面的知识，另一方面希

望该问答系统可以为医学人员研究抑郁症提供辅助作用。此外，该问答系统还加入了药物发现功能，该功能仅针对于医学研究人员，希望可以借助该功能辅助药物发现研究。当然，该功能目前还处于测试阶段，一方面是知识图谱中的数据量有限，覆盖的知识不够全，会影响模型实际的预测效果，另一方面是因为系统的预测结果无法验证，仅从实验层面证明还不够，还需要医学研究人员从医学层面进行验证，才能真正投入使用。

6.2 展望

本文首先基于生物医学文献构建了抑郁症知识图谱，然后使用知识图谱嵌入技术并基于该知识图谱进行了药物发现研究，最后基于这两部分工作构建了一个抑郁症的智能问答系统。虽然这三部分工作都取得了一定的成效，但也还存在着一些不足。

第一部分的知识图谱构建工作中以生物医学文献摘要为数据来源，虽然获取了接近 10 万个文献摘要，但还是远远不够，不足以覆盖所有的抑郁症知识。此外本文的数据源过于单一，仅考虑了文本数据，书籍、医学网站、研究报告、电子病历等都可以作为数据源，后续也可以考虑加入这些数据源以进一步扩大知识图谱的数据量；本文构建的抑郁症知识图谱的质量还有待进一步提升，除了扩大数据量、丰富数据源之外，后续的研究也可以针对构建过程中用到的命名实体识别方法、关系抽取方法、数据融合方法进行；多模态知识图谱作为时下知识图谱领域最热门的研究方向之一，接下来的研究可以考虑往多模态知识图谱的方向进行，将视频、语音、图像、文字等不同模态的知识融合到一个知识图谱中，借助多模态知识图谱更好地赋能人工智能。

第二部分将知识图谱和知识图谱嵌入技术运用到了药物发现领域，希望能够助力药物发现，虽然通过实验证明了这种药物发现方法的有效性，但是对于药物发现的结果还无法从医学层面进行验证。通过计算机技术辅助药物发现具有十分重大的现实意义，在后续的研究中可以继续朝着该方向前进，将知识图谱与其他深度学习方法相结合进行药物发现会是一个不错的研究方向。

第三部分构建了一个抑郁症智能问答系统，该系统采用了基于模板的 KBQA 方法进行构建，这种方式能够回答大部分的简单问题，但是在复杂问句上的表现不佳。后续的研究可以针对智能问答系统进行，尝试使用其它的 KBQA 方法构建问答系统，例如通过该问答系统收集用户的问题构造语料库，然后使用该语料库作为训练数据，将基于深度学习的 KBQA 方法引入到问答系统中，从而提升

问答系统的效果。此外，该问答系统只能进行单轮对话，后续的研究还可以朝着多轮对话的方向进行。

知识图谱作为下一代人工智能的基石，拥有着巨大的潜力，相信在不久的将来，知识图谱会进入各行各业，为人们的生活带来极大地便利。

参考文献

[1] American Psychiatric Association D S, American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5[M]. Washington, DC: American psychiatric association, 2013. 156-157.

[2] American Psychiatric Association D S, American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5[M]. Washington, DC: American psychiatric association, 2013. 158-159.

[3] World Health Organization, Depression: Let's Talk[EB/OL]. 2018. http://www.who.int/topics/depression/zh/

[4] Huang Y, Wang Y U, Wang H, et al. Prevalence of mental disorders in China: a cross-sectional epidemiological study[J]. The Lancet Psychiatry, 2019, 6(3): 211-224.

[5] 傅小兰，张侃，陈雪峰，等. 中国国民心理健康发展报告(2019-2020)[M]. 北京，社会科学文献出版社，2021, 192-193.

[6] Singhal A. Introducing the knowledge graph: things, not strings[J]. Official Google Blog, 2012, 5: 16.

[7] Schneider E W. Course modularization applied: The interface system and its implications for sequence control and data analysis[J]. Behavioral Objectives, 1973. 21.

[8] Lenat D B, Guha R V. Building large knowledge-based systems: Representation and inference in the CYC project[J]. Artificial Intelligence, 1993, 61(1): 41-52.

[9] Liu H, Singh P. ConceptNet-a practical commonsense reasoning tool-kit[J]. BT Technology Journal, 2004, 22(4): 211-226.

[10] Bollacker K, Evans C, Paritosh P, et al. Freebase: A collaboratively created graph database for structuring human knowledge[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008: 1247-1250.

[11] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A nucleus for a web of open data[C]// International Semantic Web Conference, 2007: 722-735.

[12] Suchanek F M, Kasneci G, Weikum G. Yago: A core of semantic knowledge[C]// Proceedings of the 16th International Conference on World Wide Web, 2007: 697-706.

[13] Vrandečić D, Krötzsch M. Wikidata: A free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.

[14] Xu B, Xu Y, Liang J, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system[C]// International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2017: 428-438.

[15] Niu X, Sun X, Wang H, et al. Zhishi. me-weaving Chinese linking open data[C]// International Semantic Web Conference. Springer, Berlin, Heidelberg, 2011: 205-220.

[16] Wu W, Li H, Wang H, et al. Probase: A probabilistic taxonomy for text understanding[C]// Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012: 481-492.

[17] Chen J, Wang A, et al. CN-Probase: A data-driven approach for large-scale Chinese taxonomy construction[C]// 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE, 2019: 1706-1709.

[18] Ernst P, Siu A, Weikum G. Knowlife: A versatile approach for constructing a large knowledge graph for biomedical sciences[J]. BMC Bioinformatics, 2015, 16(1): 1-13.

[19] Messina A, Pribadi H, Stichbury J, et al. BioGrakn: A knowledge graph-based semantic database for biomedical sciences[C]// Conference on Complex, Intelligent, and Software Intensive Systems. Springer, Cham, 2017: 299-309.

[20] 昝红英, 窦华溢, 贾玉祥, 等. 基于多来源文本的中文医学知识图谱的构建[J]. 郑州大学学报: 理学版, 2020, 52(2): 45-51.

[21] Wang Q, Li M, Wang X, et al. COVID-19 literature knowledge graph construction and drug repurposing report generation[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations. 2021: 66-77.

[22] Ferrada S, Bustos B, Hogan A. IMGpedia: A linked dataset with content-based analysis of Wikimedia images[C]// International Semantic Web Conference. Springer, Cham, 2017: 84-93.

[23] Wang M, Qi G, Wang H F, et al. Richpedia: A comprehensive multi-modal knowledge graph[C]// Joint International Semantic Technology Conference. Springer, Cham, 2019: 130-145.

[24] Speer R, Chin J, Havasi C. ConceptNet 5.5: An open multilingual graph of general knowledge[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. 2017: 4444-4451.

[25] 邓志鸿, 唐世渭, 张铭, 等. Ontology 研究综述[J]. 北京大学学报: 自然科学版, 2002, 38(5): 730-738.

[26] Horridge M, Knublauch H, Rector A, et al. A practical guide to building OWL ontologies using the Protégé-OWL plugin and CO-ODE tools edition 1.0[EB/OL]. University of Manchester, 2004.

[27] Farquhar A, Fikes R, Rice J. The ontolingua server: A tool for collaborative ontology construction[J]. International Journal of Human-Computer Studies, 1997, 46(6): 707-727.

[28] Domingue J. Tadzebao and WebOnto: Discussing, browsing, and editing ontologies on the web[C]// Proceedings of the 11th Knowledge Acquisition for Knowledge-Based Systems Workshop. 1998.

[29] Sure Y, Erdmann M, Angele J, et al. OntoEdit: Collaborative ontology development for the semantic web[C]// International semantic web conference. Springer, Berlin, Heidelberg, 2002: 221-235.

[30] Yadav V, Bethard S. A survey on recent advances in named entity recognition from deep learning models[C]// Proceedings of the 27th International Conference on Computational Linguistics. 2018: 2145-2158.

[31] Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 473-480.

[32] Chen A, Peng F, Shan R, et al. Chinese named entity recognition with conditional probabilistic

models[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 173-176.

[33] Lu P, Yang Y, Gao Y, et al. Hierarchical conditional random fields (HCRF) for Chinese named entity tagging[C]// Third International Conference on Natural Computation (ICNC 2007). IEEE, 2007, 5: 24-28.

[34] Panchendrarajan R, Amaresan A. Bidirectional LSTM-CRF for named entity recognition[C]// Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. 2018.

[35] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 2019: 4171-4186.

[36] 杨笑然. 基于知识图谱的医疗专家系统[D]. 浙江：浙江大学, 2018.

[37] Bach N, Badaskar S. A review of relation extraction[J]. Literature Review for Language and Statistics II, 2007, 2: 1-15.

[38] 庄传志, 靳小龙, 朱伟建,等. 基于深度学习的关系抽取研究综述[J]. 中文信息学报, 2019, 33(12):18.

[39] McBride B. The resource description framework (RDF) and its vocabulary description language RDFS[M]. Handbook on Ontologies. Springer, Berlin, Heidelberg, 2004: 51-65.

[40] Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology[J]. Nucleic Acids Research, 2004, 32(suppl_1): D267-D270.

[41] White J. PubMed 2.0[J]. Medical Reference Services Quarterly, 2020, 39(4): 382-387.

[42] Kilicoglu H, Fiszman M, Rodriguez A, et al. Semantic MEDLINE: A web application for managing the results of PubMed searches[C]// Proceedings of the Third International Symposium for Semantic Mining in Biomedicine. Citeseer, 2008: 69-76.

[43] Rindflesch T C, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text[J]. Journal of Biomedical Informatics, 2003, 36(6): 462-477.

[44] Greenhalgh T. The Medline database[J]. BMJ: British Medical Journal: International Edition, 1997, 315(7101): 180-183.

[45] Rindflesch T C, Kilicoglu H, Fiszman M, et al. Semantic MEDLINE: An advanced information management application for biomedicine[J]. Information Services & Use, 2011, 31(1-2): 15-21.

[46] Angeli G, Premkumar M J J, Manning C D. Leveraging linguistic structure for open domain information extraction[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 344-354.

[47] Levandowsky M, Winter D. Distance between sets[J]. Nature, 1971, 234(5323): 34-35.

[48] Webber J. A programmatic introduction to neo4j[C]// Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity. 2012: 217-218.

[49] Fleming N. How artificial intelligence is changing drug discovery[J]. Nature, 2018, 557(7706):

S55-S55.

[50] Gupta R, Srivastava D, Sahu M, et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery[J]. Molecular Diversity, 2021, 25(3): 1315-1360.

[51] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[52] Cañada A, Capella-Gutierrez S, Rabal O, et al. LimTox: A web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes[J]. Nucleic Acids Research, 2017, 45(W1): W484-W489.

[53] Patlewicz G, Jeliazkova N, Safford R J, et al. An evaluation of the implementation of the Cramer classification scheme in the Toxtree software[J]. SAR and QSAR in Environmental Research, 2008, 19(5-6): 495-524.

[54] Mayr A, Klambauer G, Unterthiner T, et al. DeepTox: Toxicity prediction using deep learning[J]. Frontiers in Environmental Science, 2016, 3: 80.

[55] Wallach I, Dzamba M, Heifets A. AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery[J]. Mathematische Zeitschrift, 2015, 47(1):34-46.

[56] Martinez V, Navarro C, Cano C, et al. DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data[J]. Artificial Intelligence in Medicine, 2015, 63(1): 41-49.

[57] Zhang W, Xu H, Li X, et al. DRIMC: An improved drug repositioning approach using Bayesian inductive matrix completion[J]. Bioinformatics, 2020, 36(9): 2839-2847.

[58] Luo H, Zhang P, Cao X H, et al. DPDR-CPI, a server that predicts drug positioning and drug repositioning via chemical-protein interactome[J]. Scientific Reports, 2016, 6(1): 1-9.

[59] Hooshmand S A, Zarei Ghobadi M, Hooshmand S E, et al. A multimodal deep learning-based drug repurposing approach for treatment of COVID-19[J]. Molecular Diversity, 2021, 25(3): 1717-1730.

[60] Mohamed S K, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings[J]. Bioinformatics, 2020, 36(2): 603-610.

[61] Zheng S, Rao J, Song Y, et al. PharmKG: A dedicated knowledge graph benchmark for bomedical data mining[J]. Briefings in Bioinformatics, 2021, 22(4): bbaa344.

[62] Hamosh A, Scott A F, Amberger J S, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders[J]. Nucleic Acids Research, 2005, 33(suppl_1): D514-D517.

[63] Hewett M, Oliver D E, Rubin D L, et al. PharmGKB: The pharmacogenetics knowledge base[J]. Nucleic Acids Research, 2002, 30(1): 163-165.

[64] Wishart D S, Knox C, Guo A C, et al. DrugBank: A knowledgebase for drugs, drug actions and drug targets[J]. Nucleic Acids Research, 2008, 36(suppl_1): D901-D906.

[65] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]// Advances in Neural Information Processing Systems, 2013, 2787-2795.

[66] Sun Z, Deng Z H, Nie J Y, et al. RotatE: Knowledge graph embedding by relational rotation in complex space[C]// Proceedings of International Conference on Learning Representations, 2019: 1-18.

[67] Yang B, Yih S W, He X, et al. Embedding entities and relations for learning and inference in

knowledge bases[C]// Proceedings of the International Conference on Learning Representations (ICLR) 2015.

[68] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C]// International Conference on Machine Learning. PMLR, 2016: 2071-2080.

[69] Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on multi-relational data[C]// Proceedings of the 28th International Conference on International Conference on Machine Learning. 2011: 809-816.

[70] Green Jr B F, Wolf A K, Chomsky C, et al. Baseball: An automatic question-answerer[C]// Proceedings of the Western Joint IRE-AIEE-ACM Computer Conference. 1961: 219-224.

[71] Woods, W., Kaplan, R. M. and Nash-Webber, B. The lunar sciences natural language information system: Final Report (BBN Technical Report 2378)[R]. Bolt, Beranek and Newman Inc., Cambridge, Massachusetts, 1972.