目　　录

摘　　要

在当今时代，互联网已成为大众首选的信息源。然而，传统检索方法往往只能返回一连串无序的网页链接，用户不得不亲自进行过滤。由于专业知识的大量涌现，有效信息的甄别对用户来说越来越具挑战性。与此同时，交互式问答系统通过解析查询问题，向用户提供精确且直接的回答，展现出更高的效率与智能度，可以满足当代社会对信息获取的迅速和精准要求。针对医疗领域的具体需求，本文构建了较大规模的通用医疗知识图谱，并基于此图谱搭建了一个问答系统。系统首次提出了 RIGP 模型用于命名实体识别，以及 RISA 模型用于意图识别。本文的主要工作包括：

（1）在复杂多样的实体识别方面，本研究提出了一个新型的命名实体识别模型，名为 RIGP，以解决多种类的实体识别挑战。该模型融合了高效的 RoBERTa-wwm 预训练语言模型与 IDCNN 深度学习网络，以强化文本特征抽取能力。此外，利用 Global Pointer 这一全局指针机制，不仅实现了对实体的高准确率标注，还显著加快了识别速度。

（2）为解决传统问答系统在问句意图识别精度不足的问题，本研究提出了一种新型模型 RISA，它结合了 RoBERTa-wwm 以进行深层词嵌入，并利用 IDCNN 网络捕捉问句的深层语义联系，同时采用 Self-Attention 机制对句间关系进行加权融合，从而显著提升了模型对问句分类的准确性。

（3）实施了一个以医疗知识图谱为核心的问答系统。本研究首先利用知识融合技术，将多源数据高效集成至 Neo4j 图形数据库中，成功构建了一个结构完整的医疗领域知识图谱，并在此基础上构建了智能问答系统。

最后通过实验对该问答系统进行了性能评估，其展现了出色的效果及实用性，可满足用户绝大多数医疗场景下的咨询需求。

关键词　知识图谱；命名实体识别；意图识别；预训练语言模型；问答系统

ABSTRACT

In today's era, the internet has become the preferred source of information for the general public. However, traditional retrieval methods often only return a series of unordered web links, requiring users to filter through them manually. With the emergence of vast amounts of specialized knowledge, the discernment of effective information has become increasingly challenging for users. Meanwhile, interactive question-answering systems, by parsing query questions and providing precise and direct answers to users, demonstrate higher efficiency and intelligence, meeting the contemporary society's demands for rapid and accurate information retrieval. Addressing specific needs in the medical field, this paper constructs a large-scale general medical knowledge graph and builds a question-answering system based on this graph. The system introduces the RIGP model for named entity recognition and the RISA model for intent recognition. The main contributions of this paper include:

(1) In the complex and diverse entity recognition domain, a novel named entity recognition model named RIGP is proposed to address various entity recognition challenges. The model integrates the efficient RoBERTa-wwm pre-trained language model with the IDCNN deep learning network to enhance text feature extraction capability. Additionally, utilizing the Global Pointer mechanism achieves not only high-precision entity annotation but also significantly accelerates the recognition speed.

(2) To address the issue of insufficient accuracy in intent recognition of traditional question-answering systems, a novel model named RISA is proposed. It combines RoBERTa-wwm for deep word embedding and utilizes the IDCNN network to capture deep semantic connections in queries. Meanwhile, the Self-Attention mechanism is employed to weight and fuse inter-sentence relationships, significantly improving the model's accuracy in question classification.

(3) Implemented a question-answering system with a medical knowledge graph as its core. This study first utilized knowledge fusion techniques to efficiently integrate data from multiple sources into the Neo4j graph database, successfully constructing a structurally

complete knowledge graph in the medical domain. Based on this, an intelligent question-answering system was developed.

Finally, the performance of the question-answering system was evaluated through experiments, which demonstrated excellent results and practicality, and can meet the consultation needs of users in most medical scenarios.

Key words: Knowledge graph; Named entity recognition; Intent recognition; Pre-trained language model; Question-answering system

第一章　绪论

1.1 研究背景及意义

在数字时代迅猛发展的今天，我们的生活品质不断提升，信息量爆炸性地增长为我们带来了极大的方便。但这也导致了大家对于高质量信息的追求变得更加迫切。根据中国互联网络信息中心在第 52 届统计报告发布会上透露的数据，2023 年 6 月份中国网民总数已经突破了 10.79 亿，相较于 2022 年 12 月增加了 1109 万，互联网渗透率上升至 76.4%。互联网医疗服务用户规模也扩大至 3.64 亿，占到网民总数的 33.8%，自 2022 年 12 月以来增长了 162 万。随着在线诊治普及，无论是网络上的常见病诊断服务，还是医疗保险报销和网购药品等，地方政府都在出台促进互联网医保发展的政策。这一系列举措为互联网医疗服务的用户增长提供了坚实的支撑。显而易见，互联网已成为大众获取医疗健康资讯和咨询医疗服务的重要渠道[1]。

当下，互联网作为人们探寻医疗信息和健康知识的重要工具，但过多的信息却使得用户选择更加困难，获取到所需信息也成了一个挑战。尽管网络上有大量医疗健康的相关知识，用户却难以高效利用。现行的搜索技术，例如传统搜索引擎，它们的功能局限于对信息的简单索引，也就是用户输入查询词，引擎回馈含关键字的网页，用户需逐一访问以寻找答案。但这些引擎缺乏对用户查询意图深层次的洞察，不能精确把握查询句子的深层含义；同时，它们存储的信息并非结构化知识。尽管现有众多医疗网站，如人民健康网和好大夫在线，它们提供的内容往往过于复杂，用户难以迅速找到所需信息。这些网站大多也不支持自动化问答服务，即便提供了医生咨询服务，通常也涉及额外的费用和医生不能实时在线等问题，效率并不高。

为了解决上述提出的难题，知识图谱技术可提供针对性的解决方案。它不只可以解析用户的搜索意图，还拥有结构化的语义知识库，并且具备逻辑推理的能力。在信息技术快速发展的当下，知识图谱作为一个热门的研究议题，其独到的优点正在满足时代的需求。目前，如 Freebase[2][3]、DBpedia[4]、KnowItAll[5]、WiKiTaxonomy[6]、YAGO[7]等，已经有了涵盖现实世界复杂结构信息的大型知识图谱。这些图谱通常采

用三元组形式来储存信息，如“肺炎”、“肺病”、“内科”等。知识图谱的应用领域极为广泛，覆盖了问答系统和网络搜索等多个方面。目前的知识库已经积累了成千上万的关系类型、数以百万计的实体以及数十亿的三元组。但是，这与现实世界中海量的信息相比，仍显得不足。因此，目前有许多研究工作正在推动知识图谱的智能化拓展。

当下，知识图谱在各领域中被广泛应用，尤以医疗领域为甚。由于医疗数据信息量的日益膨胀和医疗大数据库的构建，医疗信息资源日趋庞杂。这就要求本文必须挖掘和利用这些数据，实现智能化医疗服务，解决核心问题。这正是建立医疗知识图谱[8]的一个关键驱动因素。医疗知识图谱[9]不仅是实现当今医疗行业智能化转型的关键技术手段，而且在智能化处理医疗记录、电子化管理病历、提高医疗服务质量以及医疗知识查询等方面起到了基础性的作用。

在此项研究中，本文注意到医疗数据的高度专业性、语种多样性以及结构上的复杂性，是造成现有医疗知识图谱面临的低效率、众多约束、以及可扩展性不足等难题的根源。为此，探索一种既具有高效性、低限制性，又拥有强拓展能力的医疗知识图谱创建方法成为了本研究的重点。打造此类医疗知识图谱，将为医疗领域的智能化提供关键支持，并且促进人工智能与医疗技术的深度融合，这一点正是本文开展相关研究工作的核心意义所在。

深度学习技术在构建知识图谱方面的应用，相较于传统方式，带来了显著提升，这种进步主要表现在几个关键领域：（1）数据规模与性能关系：与传统方法相比，深度学习方法的一大特点是，数据量越庞大，其性能越优越。在处理庞杂的医疗信息时，这一点尤其凸显。（2）在特征提取方面，将领域专有知识整合到内部提取器中，可以简化数据复杂性并优化机器学习算法的效率。传统方法常常需要将大量时间耗费在特征的提取上，而且这一过程高度专业化，经常需要专家对特征进行定义和编码。相对而言，深度学习旨在直接从数据中抽取更高级别的特征，它避免了针对不同问题重复设计特征提取器的必要性，从而减轻了知识图谱大规模构建的负担，并提高了后续知识库更新时的效率和准确性。（3）深层模型处理：在输入和输出数据之间的联系呈现出高度非线性和复杂性时，传统方法往往束手无策。而深度学习的模型以其强大的非线性建模能力，在处理这些复杂情况时表现卓越。尤其是在面临结构复杂的医疗知识问题和构建医疗知识图谱时，深度学习更体现出了它的优势[10]。

在当今时代，医疗领域的知识图谱建立及其在智能问答系统中的应用已成为自

然语言处理领域内的关键研究课题之一。此领域的研究融合了人工智能、文本处理以及语义理解等多种技术手段。尽管计算机视觉领域已经取得显著进步，相比之下，自然语言处理还需致力于人们生活中大量非结构化数据的分析与处理，并在此方面显得不那么成熟。特别是在 2018 年，深度学习在自然语言处理领域取得重要进展，不仅智能问答[11]技术得以飞速发展，而且涌现出许多创新方法。尽管这些方法在通用应用中展现出卓越成效，但在医疗行业中的应用进展却相对缓慢。针对这一现状，本文旨在基于有效且精确的医疗知识图谱，运用提出的 RIGP 和 RISA 模型为核心，构建先进的智能问答系统。

1.2 国内外研究现状

1.2.1 知识图谱研究现状

自从 1972 年首次提及以来，知识图谱的应用已经有了长足的发展，尤其是在 2012 年，随着谷歌公司推出其标志性的知识图谱产品，这一概念在当代获得了全新的内涵。该知识图谱主要服务于谷歌的搜索引擎，它的性能之所以卓越，部分归功于集成的多种自然语言处理技术。这种图谱构建方式利用节点和边来连接各种信息点，它不仅直观地展现了数据之间的联系，还构筑出一张完备的图谱[12]。当前，作为存储大量人类知识的数据结构，知识图谱在多个场景下展现出其价值，特别是在自然语言处理、搜索技术和智能对话系统等方面的广泛应用。

在国外，医疗、金融及政府部门等各个行业都已认识到知识图谱的价值并予以重视，多个学术研究团体均对知识图谱的发展有显著贡献。比如，斯坦福大学的 Hazy Research 团队所推出的深度学习工具 DeepDive 系统[13]，它能够将杂乱无章或是部分有结构的数据源转换为有序的结构化数据，辅助用户构建及管理知识图谱。同样，位于伯克利的加州大学 AMPLab 也开发了 Apache Spark 知识图谱处理工具[14]，这大幅提升了大型知识图谱的分析及处理能力。而 Google Knowledge Vault 等平台的问世，加速了知识图谱应用的普及，谷歌还推出了一系列构建和管理知识图谱的工具，包括但不限于基于维基百科技术的 Freebase 和 Google Knowledge Graph。Freebase[15]这个项目，力求将众多不同来源的数据整合成有机的结构。尽管包含了超过 4000 万个实体以及几亿的属性与关系，2015 年终止了服务。与此同时，Google Knowledge Graph 作为谷歌搜索引擎的增强功能，为用户呈现了更加详细的搜索信息，它基于知识图谱

技术，结合了网络爬虫和机器学习算法，自动从网络上不同的数据源中提取信息，并在搜索结果中以图形化方式呈现。知识图谱通过辨识实体的特性、相互间的联系以及重大事件，并将这些信息融合到搜索结果当中。

在国内，知识图谱领域实现了重大突破。互联网巨头如百度、阿里巴巴、华为纷纷投入资源研发相关项目。百度将其知识图谱整合至百度大脑中，从而支撑自然语言处理、机器翻译和智能问答等技术。该图谱集成了百度的搜索、地图服务等，为百度大脑提供了扩展的数据资源[16]。而阿里巴巴利用电子商务知识图谱深化了对旗下平台商品、商家和消费者数据的解析，创建了庞大且多元的知识网络[17]。同时，国内高等学府及科研机构也在此领域做出了贡献。例如，清华大学研制了 OpenKE[18]工具，这一开源项目支持广泛的知识表示和图网络模型，并通过 Web 界面使用户能够轻松训练、测试模型并实现实体与关系的可视化。语言与知识计算专业委员会则推出了 OpenKG[19]，这一开放项目集成多源数据，并提供交互式的可视化工具，协助用户构建和管理知识图谱。复旦大学知识工场实验室的 CN-DBpedia 则是基于 DBpedia 的，不仅持续更新扩展国内相关的知识数据，而且特别为中文环境及中国特色领域量身打造了数据模板和结构。CN-DBpedia 已经汇集了大量的实体、谓词和三元组数据，覆盖了诸如人物、机构、地理、文化等众多领域。

1.2.2 智能问答研究现状

自然语言的理解和生成能力是实现有效人机交互的关键，智能问答系统作为该领域的一项主要技术，通过理解人类语言的提问并提供精准答案，为各种查询任务提供了便捷解决方案。该系统在商业、教育等多个场景中展现出其潜在价值。60 年代初期的问答系统以规则和模板为基础，受限于知识内容较少，难以满足复杂需求[20]。然而，随着深度学习等先进技术的涌现，智能问答技术取得了显著进步。虽然在线搜索引擎依旧普及，但是如京东的 JIMI、苹果的 Siri、小米的小爱同学和微软推出的智能问答服务，均展示了智能问答系统在减少劳动力投入、改善用户体验方面的巨大潜力[21]。这些进步标志着智能问答系统正成为与人类互动的窗口，同时也是推动未来智能信息服务发展的核心技术。

智能问答系统助力于用户迅速地获取所需信息，它们与搜索引擎同出一辙但却更具针对性。其独特性体现在能够理解用户所提问题，并准确无误地提炼出所需答案。这种系统的枢纽在于问句分析、信息检索以及答案抽取这三大核心环节，它们共

同应对自然语言处理所带来的挑战。随着信息技术的不断发展，智能问答系统变得越来越重要，它不仅简化了用户获取信息的流程，还有效减少了企业在人力资源上的投入，同时提升了工作的效率。尽管多种智能问答系统可能在构建架构和技术实施上存在差异，但它们的共同使命是解读用户的查询、从知识库中检索数据，并且准确地抽取出回答。

针对智能问答系统在医疗行业中的应用，全球研究者致力于探索如何应用人工智能技术创建稳定、可信赖的健康咨询系统，并努力优化医生与病人间的交流。在此背景下，IBM 的 Watson 医疗智能问答系统代表了一种运用人工智能的、旨在医疗界提供支持的智慧型问答平台[22]。它通过解析医学资料和患者的病历数据，为医护人员及病人呈现出智能化的诊疗辅助与健康监护服务。此外，谷歌推出的“谷歌医生”服务，则聚焦于为医疗咨询提供专家级建议[23]。这一系统把人工智能与机器学习的力量带入医疗界，实现了医疗数据的智能收集与综合，能够基于患者具体情况，提出一系列可能的医疗诊断和治疗建议，并向患者提供必要的医疗资讯及指南。

国内医疗智能问答技术的探索主要集中于消除患者与医生间交流障碍所可能引起的误解。在此领域内，陈梅梅团队[24]运用了基于规则的字典方法，成功构建了一种医疗知识图谱，并采用实体识别技术与基于 LSTM 的属性链接算法来打造了一套即时问答平台，旨在为公众提供服务；另一方面，杨笑然等人[25]分析并对比了融入注意力机制的双向长短期记忆网络模型和记忆神经网络的效率，最终决定选用记忆神经网络来构建问答系统；黄星宇等研究者[26]则提出利用 ALBERT 进行问句语义分析，这一改进显著提高了命名实体识别的准确性，由此开发出的问答系统能更精准地理解并响应用户的需求。

1.3 研究内容

本研究将重点放在医疗领域的应用上，对当前领域知识图谱建立以及自动问答系统的开发中遇到的挑战进行了探讨。本文对知识提取和问句解析的技术手段进行了深度剖析，针对命名实体识别和意图识别的任务，分别提出了 RIGP（RoBERTa-wwm-IDCNN-Global Pointer）和 RISA（RoBERTa-wwm-IDCNN-Self Attention）模型。依托这些研究成果，本文最终实现了基于知识图谱的医疗问答系统的开发。

（1）研究医疗领域知识图谱的构建

首先，通过网络爬取算法，本研究从“人民健康网”收集了大约两万多个相关的医

疗网页内容。接下来，通过数据清洗和格式化的步骤，优化了所采集的医疗信息，并以 json 文件形式输出。借助这批数据，研究中定义了相关实体、属性与关联关系，并据此建立了医疗实体的节点以及它们之间的联系。最后，整理好的数据被有效整合进 Neo4j 图形数据库，完成了医疗领域知识图谱的构建过程。此知识图谱的构建，不仅完整呈现了医疗信息的框架，而且为之后开发的问答系统提供了坚实的数据基础。

（2）研究问答系统的算法设计

本文将问答系统文本语义的处理划分为两个主要部分：命名实体的识别和意图的询问澄清。针对命名实体识别，本文设计了一种结合了全词遮蔽技术的 RoBERTa-wwm、深层次特征提取的 IDCNN 以及全局最优决策的 Global Pointer 的综合模型：RoBERTa-wwm-IDCNN-Global Pointer。这个模型利用 RoBERTa-wwm 以全词遮蔽形式考虑上下文信息以改善词语表征，通过 IDCNN 来捕捉句子级别的信息，再应用 Global Pointer 算法优化实体的最终识别。本文将此模型应用于 CMeEE 数据集，并将结果与目前主流的模型进行比较。测试结果展示了本文提出的模型在准确率上的明显优势。同时，为了处理意图识别，本文提出了 RoBERTa-wwm-IDCNN-Self Attention 模型，它在特定构建的新 CMID 数据集上进行了测试，以符合本文的研究要求。该模型结合了 RoBERTa-wwm-IDCNN 的预训练语言模型，提高了对句子含义的理解，通过 Self Attention 机制对特征进行加权，并使用 softmax 函数完成意图的分类。本文的模型在 CMID 数据集上进行了评估，并显示出在精确度上超越了其它文本分类的主流模型。

（3）基于知识图谱的医疗问答系统的设计和实现

本文采用 Flask 框架，实现了知识图谱与问答系统各自模块的后端整合。同时，前端界面利用 Vue 框架进行了搭建，为用户提供了一个既直观又方便的交互平台以进行问答。系统构建完毕后，通过对其功能的一系列测试，验证了该系统的可靠性。

1.4 论文结构

本论文共分为六章，内容安排如下:

第一章概述了研究的背景及意义，对当前知识图谱与问答系统的学术探讨情况进行了深入剖析，并明确了研究的主题及各个章节的结构安排。

第二章提供了对后续章节中所用重要技术的概括性说明。本章节将详细阐述知识图谱、深度学习模型以及注意力机制等概念，为理解本文的技术细节奠定基础。

第三章聚焦于构建医疗领域的知识图谱。首先，本文采集所需的数据并对其进行预处理，以确保其准确性和可用性。接着，本文基于所得数据信息定义了相应的数据模式。完成这些步骤之后，本文把处理过的数据导入到 Neo4j 这一图形数据库中，并对构建的知识图谱进行了可视化展示。

第四章着重探讨智能问答系统的模型构建。首先，本研究所采用的命名实体识别技术涉及特定数据集的详细描述以及模型架构的阐述。其次将 RoBERTa-wwm-IDCNN-Global Pointer 模型与主流模型进行比较，实验数据显示了本模型在识别任务上的显著优势。接着，文章展开讨论了在医疗意图识别领域中使用的数据集，并通过实验证实，与其他主流技术相比，采用 RoBERTa-wwm-IDCNN-Self Attention 模型的方法显示出更高的识别性能。

第五章中深入探讨了系统的设计与实现过程。对系统需求的充分分析推动了本文合理的系统框架的构建。为了提升用户体验，使其更加方便地进行信息展示和交互，本文选用了 Flask 这一 Web 应用框架，将问答服务包装起来，并通过 Web 界面与用户互动。本文也对系统的各项功能进行了综合测试，这些测试涉及了对命名实体的识别、用户意图的判定以及专业的医疗问答能力。在对上述功能进行严格测试后，将对本系统的问答界面进行展示。

第六章全面回顾了本项研究的成果，并进行了综合性的归纳。同时，深刻剖析了在系统实施阶段遭遇的各项挑战。此外，本章节回顾了文中提及的相关研究，并突出了如何将这些研究进一步扩展和深化的可能路径。作为结尾，本文还讨论了未来研究可能采取的策略和方向，旨在为后续工作者勾勒一幅未来研究的蓝图。

1.5 本章小结

本章首先探讨了知识图谱的理论起源与其实际重要性，并详细阐述了知识图谱的概念框架、演进脉络、应用场景与国内外的研究进展。同时还阐述了从传统方法到深度学习技术在知识图谱领域的相关研究。在最后部分，对本文的工作进行了总结，并对构建的基于医疗知识图谱的智能问答系统中的创新之处进行了深入分析。

第二章　相关理论与技术概述

2.1 知识图谱概述

在人工智能领域，知识图谱是知识工程的一个核心研究点，通过使用领域的知识数据来构建知识体系，显示出了强大的应用潜力。它在加强网络语义搜索的精确度方面表现出了巨大潜力，并且在智能问答服务上展现出了强大的能力，逐渐成为知识驱动型智能服务的重要基石。并且，与大数据、深度学习并驾齐驱，知识图谱对互联网及人工智能领域的革新与进步起着至关重要的作用。

2012 年 5 月 17 日，谷歌推出了知识图谱（Knowledge Graph）[27]这一概念，此举旨在利用其开发出新一代更智能的搜索引擎。这种知识图谱技术孕育了创新的信息获取模式，它提供了针对搜索难题的新视角。知识图谱实质是一个展现不同实体间联系的语义网络，有助于对世间万物及其关联性进行定式化阐释。当前，知识图谱一词通常用以泛称众多大型知识库。示例如图 2.1 所示。

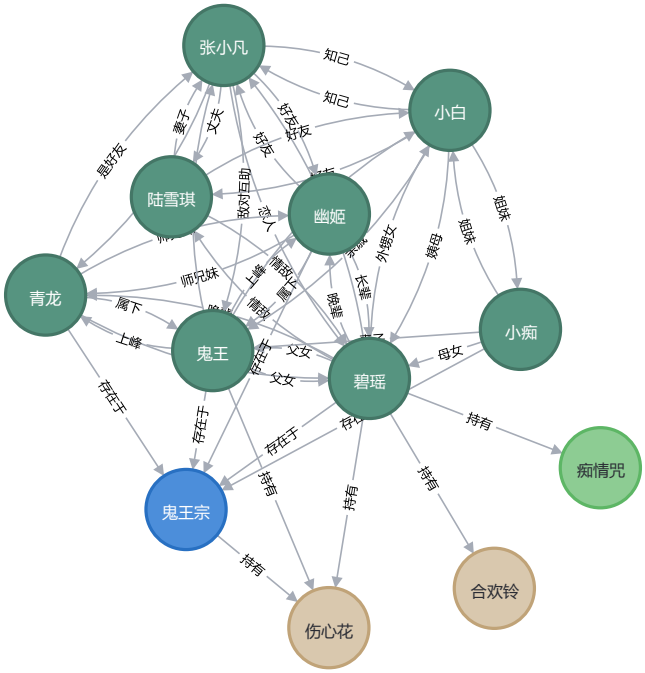

图 2.1 知识图谱示例图

知识图谱的构成要素为三元组，即 G=（E,R,S）。E={e1,e2,e3,…,en}代表其中的实体集合；R={r1,r2,…,rn}代表关系集合，涉及的关系种类总计为|R|个；S⊆E×R×E，代表了由实体和关系构成的三元组集，这个集合蕴含了知识库中的核心信息。在知识图谱里，实体是构成基础的核心要素，而实体之间的互动与联系通过各种不同的关系得以体现。这些联系包括了多样的类型，如实体之间的相互关系、概念及其属性和相应的属性值。在这里，概念可以被视作集群或分类，它们定义了对象的类别或种类；属性描述了对象可能拥有的各项特征或特性；属性值则具体指出了这些特征或特性的具体内容。为了对知识图谱中的每一个概念进行明确界定，本文通常使用一个独一无二的全局 ID 来对其进行标识。实体的内在特质通过其属性与属性值的配对来表征，同时，两个实体间的相互作用和联系通过关系这一纽带来显现，描绘它们相互之间的联结[28]。

2.2 深度学习模型

2.2.1 卷积神经网络

CNN（Convolutional Neural Network），即卷积神经网络，已成为处理图像和自然语言的重要深度学习模型。该网络结构的关键部分为卷积层（Convolutional Layer），其重要性不言而喻。Lecun 等人[29]于 1998 年提出了一种新的梯度反向传播算法，旨在提升文档识别效率。

2012 年，Krizhevsky 等人[30]设计出了著名的 AlexNet 架构，并且在 ImageNet 的图像识别大赛中取得了压倒性的胜利，领先之前的最佳成绩达 11%以上。此后，各位研究者相继提出了多种新的网络结构，不断地打破 ImageNet 比赛的纪录。其中，一些颇具影响力的网络架构包括 VGG（Visual Geometry Group）[31]，GoogLeNet[32]，以及 ResNet[33]。图 2.2 为 Lecun 等人提出的 LeNet-5 的网络架构：

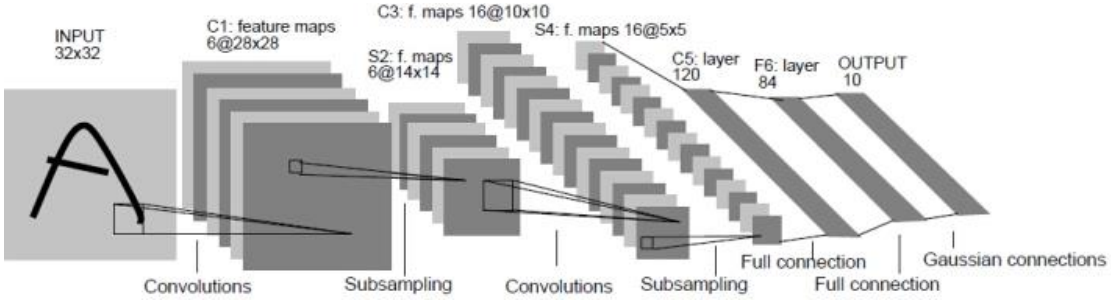

图 2.2 LeNet-5 网络架构图

卷积神经网络（CNN）在处理图像分类的任务中展现了显著的效果。不仅如此，研究人员还努力拓展其在图像处理其它方面的应用范围，包括但不限于物体检测[34][35][36]、语义分割[37]、图像摘要[38]、行为识别[39]等。此外，CNN 在处理非图像数据方面也显示出其潜力[40]。

下面本文针对 CNN 网络中的不同类型的网络层逐一进行介绍。

（1）输入层

LeNet-5 主要针对的是将一幅灰阶图像（Gray Scale）进行数字识别，该图像的分辨率为 32x32 像素。通常，电脑处理的彩色图像使用红、绿、蓝三个色彩通道来表示，每个通道的像素点都有一个特定的数值区间。与此不同的是，灰阶图像只有单一通道，缺乏色彩数据，但其像素值的区间与彩色图像是一致的。

因此，从计算机视角来看，图像仅仅是数值矩阵的集合，正如图 2.3 所展示的，取自于 MNIST 数据集的样本图像。

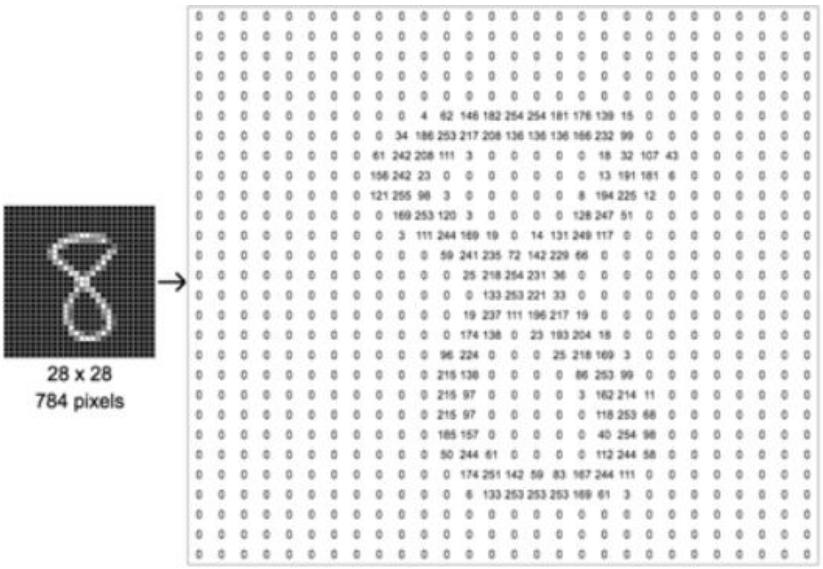

图 2.3 MNIST 样本图像

在对图像送入卷积神经网络（CNN）进行处理之前，通常需要对其像素值进行标准化。鉴于图像中每个像素的最大可能值是 255，通过将这些数值除以 255，本文可以将它们标准化到 0 到 1 之间的区间内。这样的操作有助于网络更高效地处理图像数据。

（2）卷积层

要理解卷积层的工作原理，首先需要掌握卷积这一概念。卷积是一种特殊的数学

运算，涉及两个函数在数学上的结合方式。具体而言，这个操作可以表述为：在给定的函数空间内，卷积运算通过公式(2.1)来实现：

$$(f*g)(x)=\int_{-\infty}^{\infty} f(\tau)g(x-\tau)d\tau \tag{2.1}$$

其中，f,g 两个概率矩阵尺寸匹配。在卷积神经网络(CNN)的场景里，输入数据通常是一个较小规模的矩阵，也被称之为“卷积核”。在这种情况下，卷积的运算方法与之前例子中类似，区别在于需要对矩阵进行旋转，以确保相乘的各元素位置对应一致。同时，矩阵需在对应的输入图像上滑动以计算得到卷积值。如图 2.4 所演示，它详细描述了此类计算的单个步骤。

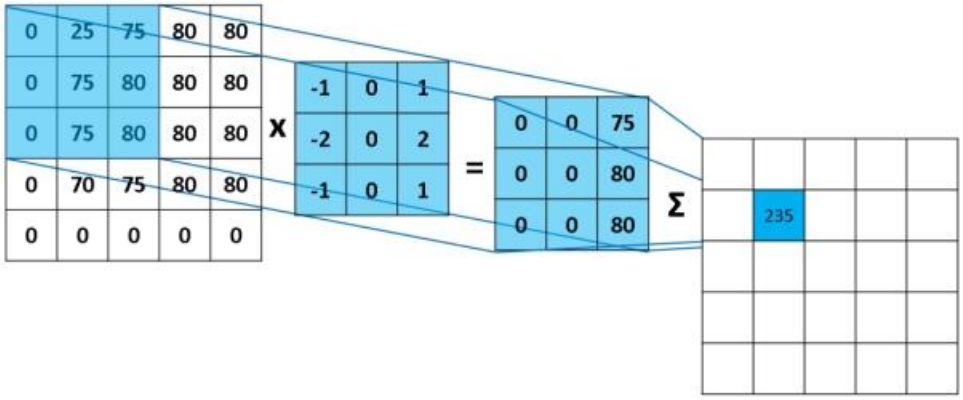

图 2.4 卷积计算步骤

（3）非线性层

通常，在构建 CNN 模型时，多采用 ReLU 函数进行激活，尽管非线性激活层在卷积神经网络以外也有广泛应用，故不再此做过多阐述。

（4）池化层

池化层通过应用池化函数（pooling function）来改善网络的输出结果。这一层的核心作用在于采用一块区域内的输出数据的整体统计信息来替换在同一位置上的原始输出值。主流的池化方法主要有最大池化（max pooling），它提取区域内的最高数值，以及平均池化（average pooling），它计算出区域内的均值。不论采取哪种方法，池化层都能保持对输入数据的某种程度上的稳定性，即使是在其遭受轻微位移的情况下。这种对小范围位移的不敏感性，即局部平移不变性是非常重要的性质，尤其是

当关心某个特征是否出现而不关心它出现的位置时。

（5）全连接层

全连接层（Fully-connected or Dense Layer）的作用在于，它将卷积神经网络中的最终池化输出与输出层所需的预测节点相连。拿手写数字识别任务来说，输出层需区分的是 0 到 9，即 10 个不同数字，若应用 one-hot 编码技术，将会有 10 个输出节点。LeNet 结构中包含两个这样的层，节点数分别是 120 和 84，常规做法是让全连接层中的节点数量按层级减少。重要的是在连接操作之前，最终的池化层输出需要经过一个“flatten”处理，将多维数据转换成一维向量，之后才能与全连接层进行结合。

（6）输出层

针对不同的应用场景，输出层的构成会有所调整。以手写数字辨识为例，若运用 one-hot 编码方式，该层将设立十个神经元，每个神经元对应一个数字。而在处理回归问题或二元分类问题时，输出层通常只设一个神经元。在处理二元分类问题的情形中，本文亦可采纳多分类问题的策略，运用 one-hot 编码来区分两种类别，分别用以表示类别 0 和类别 1。

2.2.2 循环神经网络

循环神经网络（Recurrent Neural Network, RNN）的结构主要包括时间序列模型，而不是基于结构递归神经网络（Recursive Neural Network）。此类网络设计之初，旨在对连续的数据点集合进行有效分析。

（1）网络结构

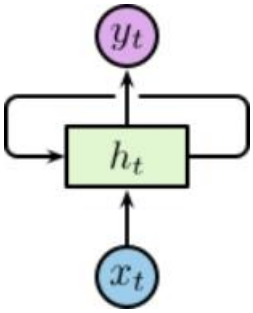

图 2.5 RNN

传统前馈神经网络的工作方式是接收特定输入并产生输出。相对而言，图 2.5 所展示的循环神经网络（RNN）由构成人工神经元网和至少一个反馈环组成。

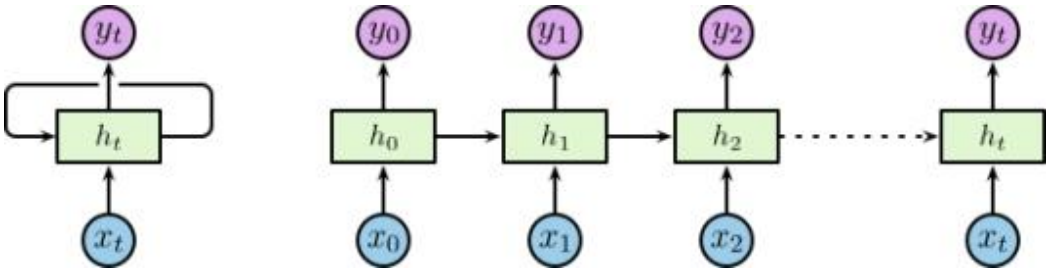

图 2.6 RNN 网络结构

网络的架构涵盖了三个主要部分：输入层x_t负责接收初步数据，输出层y_t则负责产出结果。与此同时，网络中嵌入了包含循环连接的隐含层h_t。为了使网络结构更为清晰，可以将隐含层中的循环链接展现出来，其具体展开结果如图 2.6 所示。

展开的网络框架接收时间序列$\{\ldots,x_{t-1},x_t,x_{t+1},\ldots\}$作为其输入，$x_t \in R^n$，这里的 n 代表输入层的神经元数量。接着，对应的隐含层为$\{\ldots,h_{t-1},h_t,h_{t+1},\ldots\}$，$h_t \in R^m$，m 表示隐藏层神经元的数量。在隐藏层，使用微小的非零值来初始化节点有助于增强网络整体的表现和稳定性[41]。此外，隐藏层构建了系统的状态空间（state space），亦可将其视作 memory [42]。公式如 2.2。

$$h_t=f_H(o_t) \tag{2.2}$$

其中

$$o_t=W_{IH}x_t+W_{HH}h_{t-1}+b_h \tag{2.3}$$

$f_H(\cdot)$为隐含层的激活函数，b_h为隐含层的偏置向量。对应的输出层为$\{\ldots,y_{t-1},y_t,y_{t+1},\ldots\}$，其中$y_t \in R^p$，p 为输出层神经元个数。则：

$$y_t=f_O(W_{HO}h_t+b_o) \tag{2.4}$$

其中$f_O(\cdot)$为隐含层的激活函数，b_o为隐含层的偏置向量。

在 RNN 中常用的激活函数为双曲正切函数：

$$\tanh(x)=\frac{e^{2x}-1}{e^{2x}+1} \tag{2.5}$$

Tanh 函数实际上是 Sigmoid 函数的缩放：

$$\sigma(x)=\frac{1}{1+e^{-x}}=\frac{\tanh(x/2)+1}{2} \tag{2.6}$$

（2）长期依赖问题

通过循环结构链接，循环神经网络（RNN）的隐藏层节点可以构筑记忆能力，其当前隐藏状态的形成受前一时刻状态影响。该结构赋予 RNN 处理和存储跨越较长时间跨度的复杂数据序列的能力。然而，在某些场合，当前任务的处理可能只需倚靠近期数据。例如，在一个语言模型中，其目标是通过分析之前的文本来预言接下来的单词。

如要预测“云朵在天空”一句中的最后一个字，显而易见不需过多考虑遥远的上下文，答案很可能就是“天空”。此时相关信息与预测目标之间的距离较短，RNN 在这种情形下能够高效地利用最新获得的信息。

在某些情况下，预测文本时必须依赖较为广泛的上下文内容。例如，“我在法国成长……我能说一口流畅的法语。”此例中，近文提示本文需要确定的是一种语言。

尽管理论上递归神经网络（RNN）能够处理长距离的依赖问题，但从实际出发，RNN 解决此类问题时表现并不理想。要精确识别出该语言，本文需追溯到“在法国成长”这一更为遥远的信息。

2.2.3 长短期记忆网络

（1）LSTM 网络结构

Hochreiter 和 Schmidhuber[43]提出的长短期记忆网络（Long Short Term Memroy, LSTM）属于 RNN 的一种变体，旨在克服长期依赖难题。其核心优势在于它能够存储并维系长时间跨度的信息流。

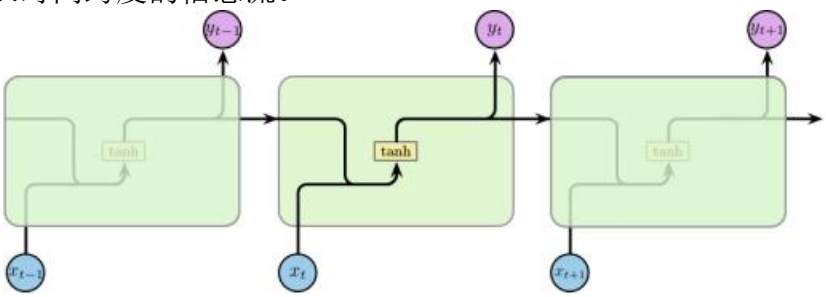

图 2.7 RNN 模块结构图

循环神经网络的结构由一连串相同的单元组成。在典型的 RNN 模型中，这些重复单元结构通常很基础，只由一个具有 tanh 激活函数的隐藏层组成，图 2.7 为对应示意图。

长短期记忆网络（LSTM）呈现出一种链式的形态，其内部重复模块与典型的不相同。在每个模块内，不是仅有一个隐藏层，而是包含四个互有特殊关联的子层，

具体的交互方式可参见图 2.8 示意图：

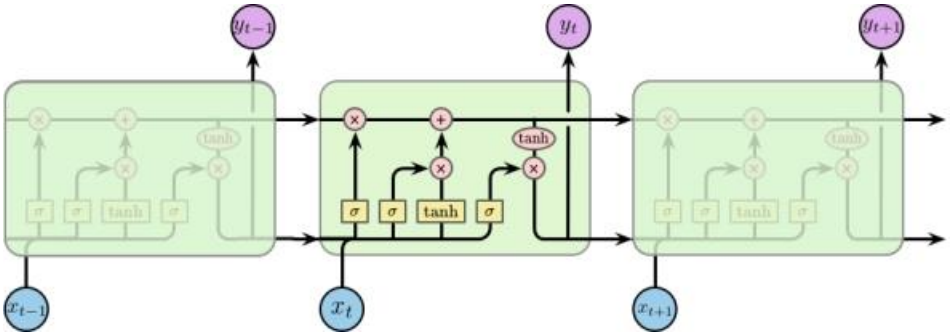

图 2.8 LSTM 模块结构图

以下是一些 LSTM 单元(cell)中将会使用到的符号的简单的说明，如图 2.9 所示：

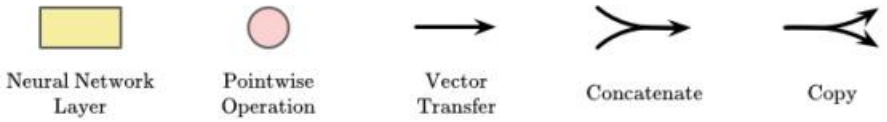

图 2.9 符号说明

本段落描述了一个神经网络中的数据流动和操作情况。为满足要求，进行如下重构：该结构示意图中，线状元素象征了数据从输出单元向其他单元扩散的完整路径。此外，粉色圆形象征对元素进行的单独操作，而黄色长方体代表了通过学习得到的网络层。图中，线条交汇象征节点间的连接，而分岔则表示信息的复制与向不同位置的迁移。

（2）LSTM 单元状态和门控机制

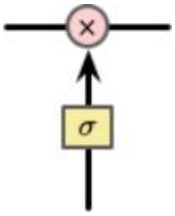

图 2.10 LSTM 门结构示意图

循环神经网络中的长短期记忆模型（LSTM）具备对单元状态进行信息新增或剔

除的功能，这一机制受一种被称作“门（gates）”的构件控制。这些“门”组件作用于挑选性地传递信息，并且它们是由带有 Sigmoid 激活函数的神经网络层以及逐元素的乘法操作共同组成，如图 2.10 所示。

单元状态（cell state），如图 2.11 中的顶部横穿单元的直线所展示，是 LSTM 核心所在。这一状态宛如传输带，在链条上直接、顺畅地移动，且只涉及极少的线性处理。这一机制使得信息得以轻松传递并且保持其原貌。

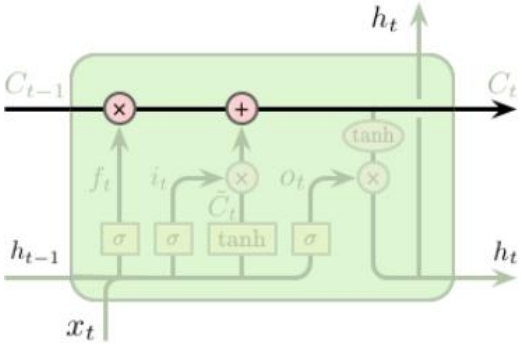

图 2.11 LSTM 单元状态

Sigmoid 函数负责将 LSTM 单元的输出限制在 0 到 1 的范围内，这个值反映了可以传递的信息量。其中，数值 0 暗示着阻断所有信息的流动，而数值 1 则意味着信息能够自由流通。在 LSTM 架构中，三个独特的门结构共同作用于控制这种类型的网络单元的内部状态。

（3）LSTM 工作步骤

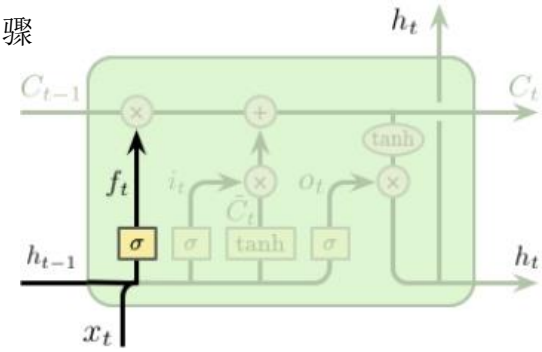

图 2.12 LSTM 遗忘门

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{2.7}$$

长短期记忆网络（LSTM）的首个操作环节是决策单元状态中需舍弃的数据点。此决策过程由一层被称作“遗忘门（forget gate）”的 Sigmoid 神经网络层来执行，如

图 2.12 所示，它接收前一时刻隐藏层的输出h_{t-1}及当前输入x_t作为其输入信号，并产出一个范围在 0 至 1 之间的数值。数值为 1 时表示信息被完全保留，而数值为 0 则意味着信息被完全抛弃，如公式(2.7)所示。如处理语言模型问题时，若出现新的主语，则旧主语的信息应通过此门被抹除，以确保模型后续在代词选择上的准确性。

接下来，需要明确在单元状态上要加入何种新资讯，这个过程分为两个步骤：开始阶段是由一个被命名为“输入门（input gate）”的 Sigmoid 神经层来执行，负责选择性地更新信息。接着，Tanh 神经层将负责制造一个新的候选值$\tilde{C}_t$，该候选值接着会被融合到单元状态。通过整合这两个过程，单元状态便得以刷新。例如，在本文处理的语言模型案例中，新的主语性别信息需被植入单元状态中，以取代被遗忘的旧主语性别信息。具体如图 2.13 和公式(2.8)所示。

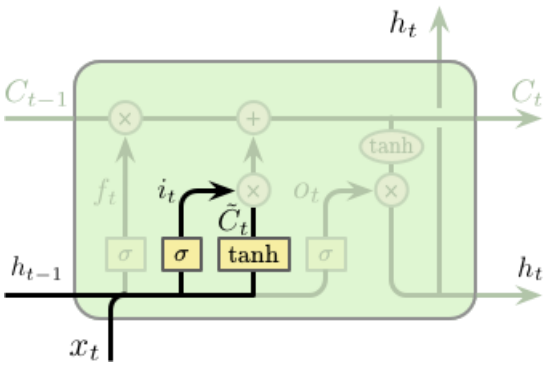

图 2.13 LSTM 输入门

$$\begin{aligned} i_t&=\sigma(W_i\cdot[h_{t-1},x_t]+b_i)\\ \widetilde{C}_t&=\tanh(W_C\cdot[h_{t-1},x_t]+b_C)\end{aligned} \tag{2.8}$$

随后，是对旧单元状态的刷新工作。

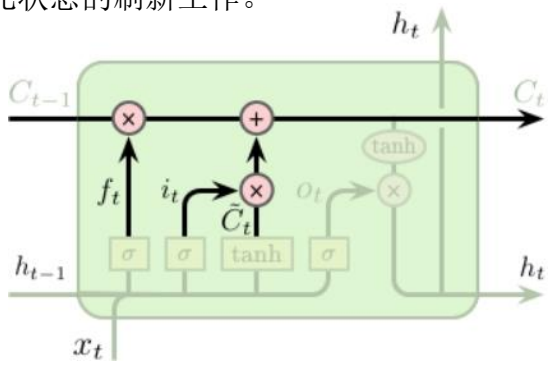

图 2.14 LSTM 更新门

$$C_t=f_t\odot C_{t-1}+i_t\odot\widetilde{C}_t \tag{2.9}$$

通过将旧单元状态与f_t相乘，本文可以精细调节遗忘旧信息的程度，紧接着加上$i_t \odot \tilde{C}_t$的部分负责引入新的单元状态更新。在所开发的语言模型里，这一步骤关键地促成了对旧主语信息的淘汰和新数据的接收，如图 2.14 和公式(2.9)所示。

接下来的步骤是决定单元的最终输出信息，这一信息将是对单元状态的一种筛选形式。初始阶段，通过 Sigmoid 函数的网络层来评估哪些部分的单元状态将被输出。紧接着，本文通过 tanh 函数对单元状态的值进行调整，使其介于-1 到 1 的范围内。调整后的状态值将与 Sigmoid 网络层的结果相结合，以产生本文所需的输出数据。如图 2.15 和公式(2.10)所示

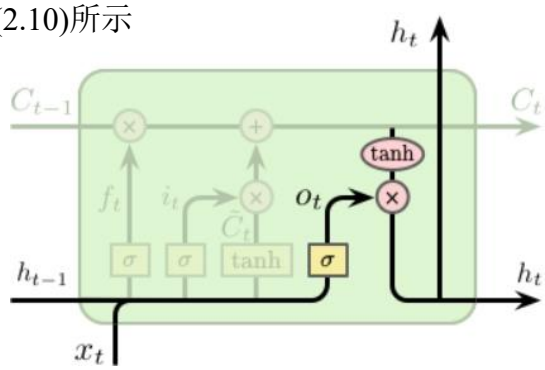

图 2.15 LSTM 输出门

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o)$$
$$h_t=o_t\odot \tanh(C_t) \tag{2.10}$$

2.2.4 序列到序列模型

序列到序列（Sequence to Sequence, Seq2Seq）模型，是处理诸如机器翻译等序列到序列转换问题的技术。

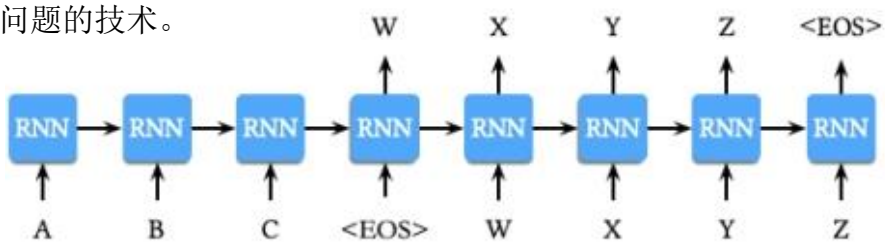

图 2.16 Seq2Seq 网络结构图

Sutskever 等人[44]提出了一种针对机器翻译的基于 Encoder-Decoder 架构的 Seq2Seq 模型，该模型与 RNN 在网络构造上有所不同，具体细节如图 2.16 所示。

模型的相关细节如下：

（1）在数据预处理过程中，本文首先对每个句子进行了特殊处理：在其结尾处插入了一个标记<EOS>，正如所展示的图中所示。随后，本文根据 A、B、C 与<EOS>的特定表示，进一步推算出 W、X、Y、Z 以及<EOS>的相应条件概率。

（2）采用了双 LSTM 模型架构，一个负责处理输入数据序列，而另一个则专注于生成输出数据序列。

（3）采用一个四层架构的 LSTM 模型可以有效增强模型的性能。

（4）将给定的输入序列进行逆序变换，举例来说，对于一个初始的输入序列 a,b,c 及其相应的输出序列α,β,γ，本文期望 LSTM 掌握从逆序后的序列映射至原始输出序列的转换关系 c,b,s->α,β,γ。

在解码模型时，采用了一种自左向右的简易 Beam Search 策略。此策略通过保持一个固定大小为 B 的候选集合来记录当前最优结果。图 2.17 清晰展现了在集合大小为 B=2 的情境中 Beam Search 的实际运作模式：

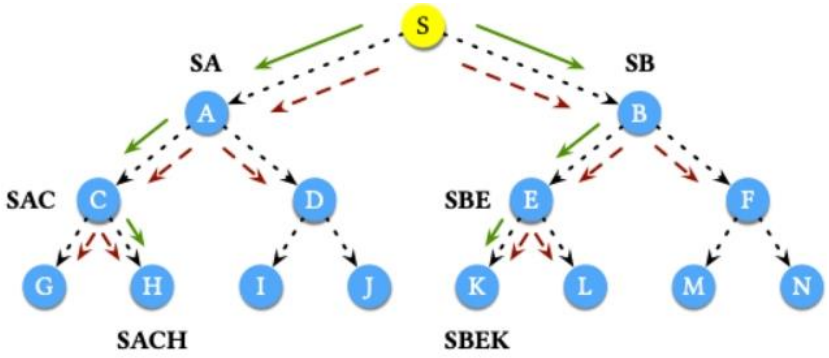

图 2.17 Seq2Seq 解码示意图

在这部分，本文可以看到，红色的虚线箭头标示出了探索过程中可能采取的每一条路径，而绿色的实线箭头则指向了在每一步中概率最高的 Top B 选择。以 S 点作为起始，搜索过程展开如下：

（1）在初步搜索阶段，最优的两个结果被确定为 SA 和 SB，这两个结果被选为前两位。

（2）在第二阶段的探索过程中，本文最终筛选出了 SAC 和 SBE 两个潜在的结果，它们在众多候选项中表现突出，因此被选为前两名。这一轮搜索中本文还考虑了 SAD 和 SBF，但它们并未进入最终的优选名单。

（3）经过第三轮搜索，得到的潜在结果包括 SACG、SACH、SBEK 与 SBEL。

在这些候选项中，根据优先级，本文选择了排名前两位的 SACH 和 SBEK 作为最终结果。随着这一步骤的完成，搜索过程也随之画上句号。

2.2.5 条件随机场

在给定某随机变量 X 的情境下，马尔可夫随机场中的变量 Y 的概率分布可以通过条件随机场（Conditional Random Field, CRF）来刻画。设 X 与 Y 是随机变量，P(Y|X)是给定 X 的条件下 Y 的条件概率分布。若随机变量 Y 构成一个有无向图 G=(V,E)表示的马尔可夫随机场，即：

$$P(Y_v|X,Y_w,w\neq v)=P(Y_v|X,Y_w,w\sim v) \tag{2.11}$$

对任意结点 v 成立，则称条件概率分布 P(Y|X)为条件随机场。其中，w~v 表示在图 G=(V,E)中与结点 v 有边连接的所有结点 w,w≠v 表示结点 v 以外的所有结点，Y_v,Y_u与Y_w为结点 v,u 和 w 对应的随机变量。相关公式见(2.12)。

$$G=(V=\{1,2,\dots,n\},E=\{(i,i+1)\}),i=1,2,\dots,n-1 \tag{2.12}$$

定义不强制 X 和 Y 共享一致的构造模式。普遍看来，X 和 Y 被认为拥有一样的图架构。以下所示为无向图线性链的例证，具体如下图 2.18 所示：

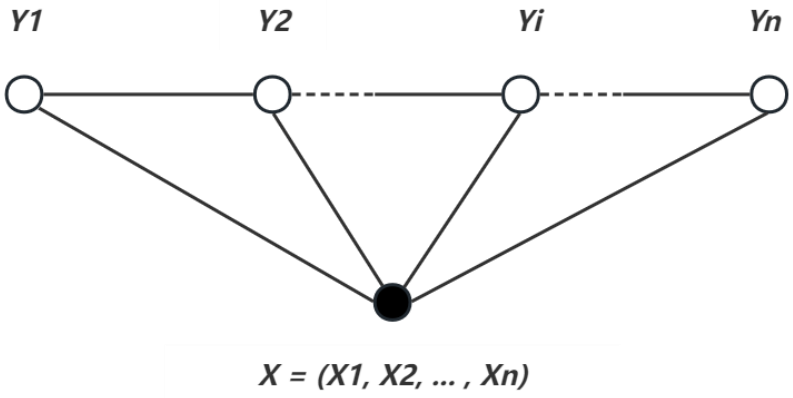

图 2.18 线性链条件随机场

此情况下，$X=(X_1,X_2,\dots,X_n)$，$Y=(Y_1,Y_2,\dots,Y_n)$，最大团是相邻两个结点的集合。

2.3 注意力机制

2.3.1 硬性和软性注意力

Xu 团队[45]在图像标题生成领域应用了注意力机制。研究者们在文章中提出了两种不同的注意力机制：硬性注意力（Hard Attention）和软性注意力（Soft Attention）。

对于硬性注意力而言，令s_t表示在生成第 t 个词时所关注的位置变量，$s_{t,i}=1$表示当第 i 个位置用于提取视觉特征。将注意力位置视为一个中间潜变量，可以以一个参数为$\{\alpha_i\}$的多项式分布表示，并将上下文向量$\hat{z}_t$视作随之变化的随机变量，如公式(2.13)所示。

$$\begin{aligned} p(s_{t,i}=1|s_{j<t},a)&=\alpha_{t,i} \\ \hat{z}_t&=\sum_i s_{t,i}a_i \end{aligned} \tag{2.13}$$

因此，硬性注意力机制能够根据隐状态的概率分布来抽样，以此计算出上下文向量。此外，为了使梯度能够逆向传播，必须运用到蒙特卡罗方法以估算梯度值。

软性注意力机制采用的是直接对上下文向量$\hat{z}_t$的期待值进行计算的方法，具体的计算过程见公式(2.14)：

$$E_{p(s_t|a)}[\hat{z}_t]=\sum_{i=1}^{L}\alpha_{t,i}a_i \tag{2.14}$$

计算剩余分量的方法与 Bahdanau 等学者[46]的研究相似。软性注意力模型的求解过程可以通过传统的反向传播算法实现，并且可以与模型的其他部分同时进行训练，这使得整个流程更为直接和便捷。

图 2.19 所示包含数个图像标题合成的视觉展示案例。图中，被白色标记的部分代表着模型关注的焦点，而与之相连的文本则反映出了相应生成标题里的关键字词。

图 2.19 图片标题生成可视化示例

2.3.2 自注意力

Vaswani 等人在文献[47]中首次引入了一种被命名为 Transformer 的革新性网络架构，它主要以自注意力机制（SelfAttention）为核心。该机制通过对序列内部各元素间的相互关联作用，实现了对句子整体意义的捕捉。编码器和解码器分别由 Self Attention 层和全连接层相互堆叠而成，构成 Transformer 的基本框架，结构如图 2.20。

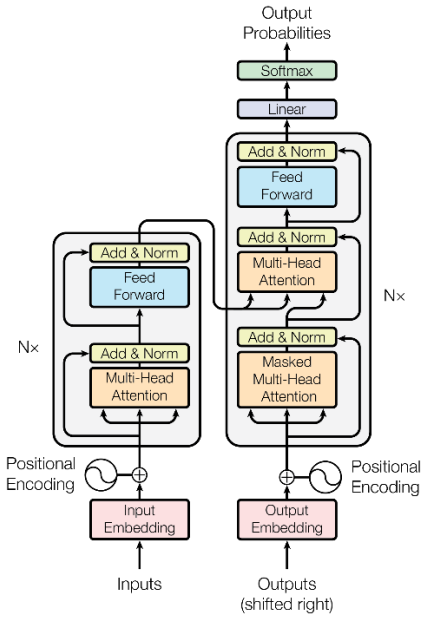

图 2.20 Transformer 模型架构

编码器由 N=6 个相同的网络构成，且每一层均由两个特定的子层组合而成。首先是一个 Multi-Head Self-Attention 层，其次是一个 Position-Wise 的全连接前馈网络。在每个子层后，模型采用了残差连接（Residual Connection）[48]以及层标准化（Layer Normalization）[49]策略。因此，可以将每层的输出表示为 LayerNorm(x + Sublayer(x))，这里 Sublayer(x)代表子层操作。为了允许残差连接的顺畅操作，确保了从嵌入层（Embedding Layer）开始，整个模型中的每个子层包括 Embedding 层的输出都维持着相同的维度。

解码部分同样由 N=6 层级组成，每个层级由三个子模块构成，其新增加的一个子模块专门负责处理来自编码器端的信息流。此外，解码层在自我关注机制上进行了改良，只允许位置 i 的预测依赖于 i 位置之前的输出，以此来确保预测的有效性。与编码器相似，解码器的每个层级都采用了残差链接和层级归一化技术。

2.4 本章小结

本章深度剖析了知识图谱的历史演变及其应用范畴的广度，对其从初步概念到构成要素进行了详细说明，并探讨了深度学习与神经网络在构建自然语言处理的知识图谱和问答系统方面的关键作用。该章进一步详述了在智能问答和知识图谱领域关键且不可替代的算法技术，包括卷积神经网络、长短期记忆网络、序列到序列模型以及注意力机制，并分析了这些技术的核心思想、理论结构以及详尽的数学推导。以上内容为本研究的理论探讨提供了坚实的基础。

第三章　医疗领域知识图谱构建

3.1 知识图谱构建流程

本章主要探讨医疗行业知识图谱的构建。主要集中于三大核心领域：知识抽取、知识融合与知识存储。该知识图谱赖以构建的数据源头涉及医学网络提供的各类数据，包括结构化、半结构化及非结构化数据，尤以前两者为主导。这一过程首先需要对采集的数据进行初步加工，识别出医疗相关的实体、它们之间的关联以及各自特性，进而形成结构化的三元组。完成这些步骤后，相关数据将被有效地存入 Neo4j 图形数据库中。

医疗知识图谱问答系统的设计方案如图 3.1 所示：

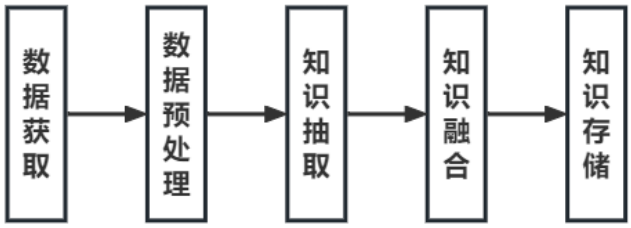

图 3.1 医疗知识图谱构建流程

3.2 知识获取

3.2.1 数据集

本智能问答系统专注于医学领域，并从专门的医药网站、阿里天池的医疗数据集以及官方收集的药品说明文档中提取信息。围绕疾病这一核心，本文成功构建了一个包含大约 58,000 个知识实体，分为 13 个类别，以及大约 270,000 个属于 37 个类别的实体关系的庞大知识图谱。

为构建医疗领域的知识图谱，本研究侧重于从多个数据源中搜集原始资料。数据的形式多样，包括结构化数据、json 文本格式数据、以及自然语言文本等。本文精心挑选了数据来源，例如开放获取的互联网医疗网页信息、药物说明文档以及已经组织

好的知识图谱竞赛数据集。本文的目标旨在增强医疗知识图谱的准确性和丰富程度，并提升终端用户的查询体验。下文将详细阐述本文用以建立知识图谱的各类数据来源。

（1）在“人民健康网”与“百度百科”的疾病专题页面，本文可以获取丰富的半结构化数据，涉及广泛的疾病信息和相应的并发症。这些页面详细记载了关于疾病的治疗相关知识，如何施治、治愈概率、常用治疗药品、责任医疗科室、饮食忌讳以及手术需求等。它们还包括了基础的常识内容，例如标准诊检程序、典型病状、高风险人群、可能伴随的其他疾病、受影响的身体部位和传播方式等。通过应用网络爬虫技术，有效搜集到了这两个平台上有关疾病及其并发症的知识数据。

（2）CMeKG2.0 数据集，CMeKG（Chinese Medical Knowledge Graph）是利用自然语言处理与文本挖掘技术，基于大规模医学文本数据，以人机结合的方式研发的中文医学知识图谱。CMeKG 的构建参考了 ICD、ATC、SNOMED、MeSH 等权威的国际医学标准以及规模庞大、多源异构的临床指南、行业标准、诊疗规范、医学教材与医学百科等医学文本信息。CMeKG2.0 目前包含 1 万余种疾病、近 2 万种药物、1 万余个症状、3 千种诊疗技术的结构化知识描述，描述医学知识的概念关系及属性三元组达 156 万。

本文所依赖的数据来源于两个主要部分：一是通过爬虫技术从医疗相关网站上获取的半结构化信息；二是来自公开可用的知识图谱数据集中的结构化数据。这两者共同构成了构建医疗领域知识图谱所需的数据基础。

3.2.2 数据预处理

数据预处理的主要任务是优化存储于数据库内的数据结构，包括对数据的清理和格式化工作。具体而言，数据清理指对数据格式执行整理，目的是删减无关的信息，如利用正则表达式技术去掉冗余的换行、空白或其他无关符号。此外，数据格式化的步骤旨在促进数据实体的识别，这通常通过构建清晰的键值对来体现，展现为实体名称和对应的实体值。执行这些步骤不仅提升了数据品质和可靠性，而且为后续的研究与数据解析工作提供了坚实的基础。

3.3 知识抽取

经过数据预处理，最终获得关于各类疾病、饮食及病理症状的数据集。在此基础

上，知识抽取的工作就涉及到从这些清晰且基于事实的数据资料中，提炼出有关的实体、实体与关系、以及属性值的信息。因此，下一步本文将把搜集来的数据与CmeKG2.0 数据集整合，在此基础上进行深入分析，进而设计并构建知识图谱中的三元组结构。鉴于不同的数据源可能采用不同的分类体系，本文必须对各个实体和属性信息的不同展现形式进行标准化，同时对数据表内部的字段命名也做出调整，以制定出新的分类标准。考虑到医疗诊断的具体特点，本研究采用了两种方式来构建医疗领域的知识三元组，即“实体-关系-实体”形式和“实体-属性-属性值”形式。

本系统共定义了 58 种标识符，覆盖了 13 种不同的实体类别标签，涉及疾病、科室、症状、制药公司、手术程序、诊断测试、测量结果、药物、食品等领域；另外设有 37 种标识符，用以标记实体与其关联的关系，包括但不限于疾病的分类、相关症状、就诊科室、易感群体、推荐食物、忌吃食物、用于治疗的药物、并发症等；以及 8 种实体属性类别的标签，它们分别指向疾病的名称、简介、传染性、原因、预防措施、治疗方式、疗程以及别称。各类别示例见表 3.1 至 3.3。

表 3.1 部分实体类型示例

实体类型	定义	示例
Department	疾病	中枢神经系统感染、静脉曲张出血、化脓性葡萄膜炎
Disease	症状	入睡困难、风寒咳嗽、肌肉无力、头昏
Food	科室	康复科、耳鼻喉科、中医科、小儿内科
Medicine	药品	多酶片、沃拉帕沙、伊沙匹隆、西帕依固龈液
Symptom	食物	荸荠粉、猪肝、甲鱼、芝麻

表 3.2 部分关系类型示例

关系类型	定义	示例
cure_department	就诊科室	<涎腺结核，就诊科室，传染科>
do_eat	宜吃食物	<白线疝，宜吃，白果（干）>
has_cause	病因	<新生儿肝炎，病因，病毒感染>
has_common_drug	推荐药物	<肥厚型心肌病，推荐药物，华法林>
not_eat	忌吃食物	<鼻前庭炎，忌吃，鸭血（白鸭）>

表 3.3 部分属性类型示例

属性类型	定义	示例
name	名称	水痘
abstract	简介	水痘是由水痘-带状疱疹病毒引起的传染病，主要特征是皮肤上出现水泡、发热等症状。
alias	别名	水泡病、水痘疹
diseased_bodypart	发病部位	全身皮肤
contagion	传染性	高，通过飞沫传播和接触传播

3.4 知识融合

通过实施实体对齐，能够在数据融合时显著降低知识的重复率，从而促进知识图谱的质量提升。具体地，将已经定义好的实体信息应用于知识库的知识结构分析。在识别到多个数据源中引用的是同一实体时，应将这些实体汇集为一候选组。例如，当“胃病”与“胃部病变”被认定为同一实体，相应的操作是在其它数据源中通过模糊查询技术来定位此实体的候选集合。具体操作是对所有实体进行初步的映射，以形成这样的候选组。

该部分采纳了一种结合规则和统计方法的实体对齐策略。该策略涉及对两个实体间字符和语义的近似程度进行量化，目的是将不同实体的别名与标准化的别名数据库相匹配[50]。这一过程将产出两个实体间最终的相似度评分。计算实体间相似度的公式为$\mathrm{sim}(x_1,y_1)$，具体公式见(3.1)。

$$\mathrm{sim}(x_1,y_1)=\omega_1\cdot\mathrm{sim}_j(x_1,y_1)+(1-\omega_1)\mathrm{sim}_s(x_1,y_1) \tag{3.1}$$

其中，字符相似度$\mathrm{sim}_j(x_1,y_1)$是通过计算两个实体的标签来确定的。如果某个实体具有多个替代名称，本文就从中挑选出相似度最高的那个作为代表。与此同时，实体之间的语义相似度$\mathrm{sim}_s(x_1,y_1)$也得到了计量，例如通过创建一系列特定名称集，本文可以得出一个确定的语义相似度值。

之后运用了余弦相似度这一技术来量化实体间的语义相关性。实体首先被转换为向量表征，然后通过测量两个实体所代表的向量夹角的余弦值来评估它们之间的

语义距离。这种方法的具体应用可以参见公式 3.2，该公式说明了如何精确计算两个实体之间的语义相似度。

$$sim_s = \cos(\alpha,\beta) = \frac{\sum_{i=1}^{n}(m_i \times n_i)}{\sqrt{\sum_{i=1}^{n} \cdot m_i^2} \times \sqrt{\sum_{i=1}^{n} \cdot n_i^2}} \tag{3.2}$$

为了衡量两个实体间字符级的相近度，采用了 Jaccard index[51]。该指标主要评估集合间的相似性：一开始通过确定实体 A 和 B 的公共字符数和总字符数来获取其交集并集的比值。计算公式取交集字符数|A∩B|与字符总数|A|+|B|-|A∩B|之比，即得到两实体间的字符相似度。当 Jaccard 系数增加时，可解释为实体对在字符层面上的相似性提高。相关系数的计算方式可以参考公式 3.3。

$$sim_j = \frac{|A \cap B|}{|A \cap B|} = \frac{|A \cap B|}{|A|+|B|+|A \cap B|} \tag{3.3}$$

在对实体进行匹配的过程中，一开始会应用一套既简便又相对准确的规则来执行匹配操作。首先审查实体名称的集合以确定是否可以精准地找到对应的实体。若无法找到，则引入一个预设的阈值来评估候选实体是否有资格被对齐至同一目标实体。此阈值被定为 0.8，仅当实体间的相似度高于此数值时，才认为它们可对齐至相同实体。在众多超过此阈值的候选中，选出相似度最高者作为最终对齐对象。

3.5 知识存储

本部分需要仔细分析医疗行业数据的独有特性，同时顾及后续实际应用场景的需求。知识图谱的存储通常分为两大类：DRF 格式与图形数据库。考虑到实际操作的便捷性及数据检索的高效率，本文决定采用图形数据库作为知识图谱的存储方案。图数据库在映射复杂和相互关联的医疗信息方面，显示出明显的优势。与 Titan、OrientDB 等图数据库相比，Neo4j 以其友好的设计和操作界面脱颖而出，其不仅能够实现快速检索，并且提供了高度可用的分布式架构，同时兼容 Java、Python 等多种流行的编程语言。鉴于以上优势以及在 DB-Engines 平台上的排名领先，本研究选用 Neo4j 图数据库进行知识图谱的存储和可视化展示。

所得的医疗数据被储存在 Neo4j 图形数据库里。数据交互则是通过 py2neo 插件与 Neo4j 建立联系进行的。鉴于庞大的数据体量，本文采用 Neo4j 所提供的 neo4j-import 工具来完成大批量实体与联系的高效输入。图 3.2 展现了部分医疗领域知识图谱的可视化成果。

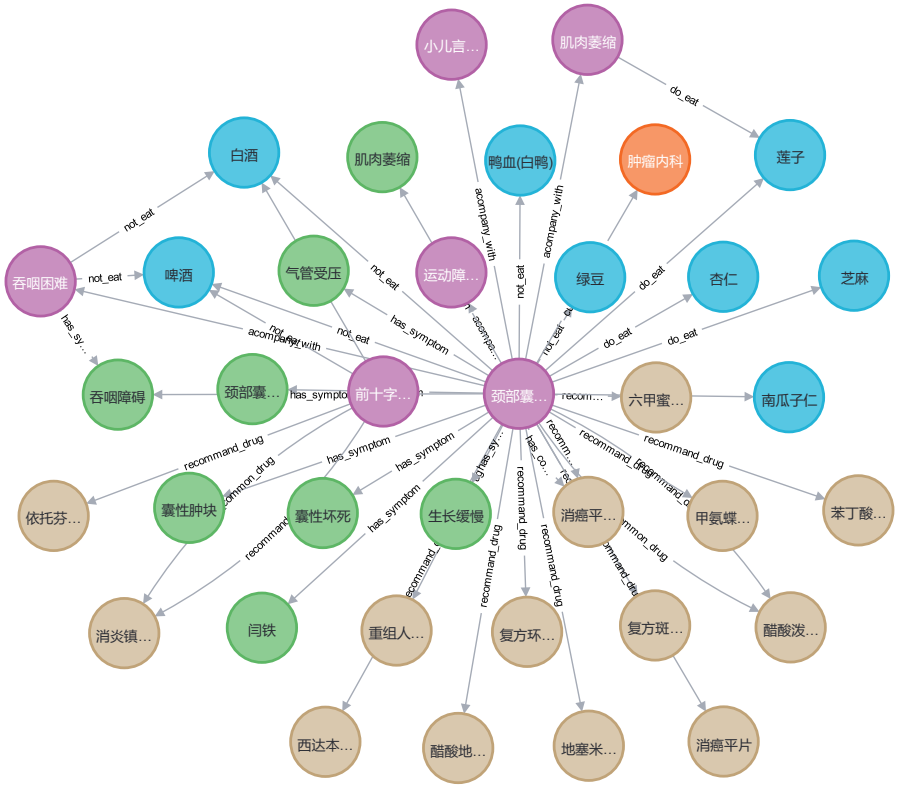

图 3.2 知识图谱部分可视化

3.6 本章小结

本章主要阐述了建立医疗知识图谱的详细步骤。初始阶段，通过网络爬虫从医疗专业网站获取数据，实现初步的数据收集。紧接着进行必要的数据清洗和整理，确保数据质量。其次，细述了医疗知识图谱中实体的识别、属性的归纳及实体之间关系的建立。此外，讨论了如何从原始数据中提取知识以及将来自不同数据源的知识整合到一起的方法。项目的最后步骤包括使用 Neo4j 这一图形数据库技术，对知识图谱进行持久化存储。构建完成的知识图谱囊括了大量医疗防治方面的结构化知识，这为后续开发医疗问答系统提供了坚实基础。

第四章　智能问答模型

知识图谱是问答系统的一个核心应用，本章主要关注知识图谱问答的构建流程。鉴于问答体系中，首要步骤是命名实体识别，它的泛化性及准确度会直接作用于特定领域知识图谱问答性能的整体优劣。因此，针对这一环节，本文提出了一项创新方法：RoBERTa-wwm-IDCNN-Global Pointer（RIGP），此方法基于先进的预训练语言模型开发而成。考虑到该技术在评估中展现出的卓越性能，对于意图识别这一后续任务，本文同样应用了这一基本框架，并结合了自注意力机制，创新性地提出了 RoBERTa-wwm-IDCNN-Self Attention（RISA）模型，实现了预期目标。

4.1 智能问答流程

医疗知识图谱问答系统涉及的关键工作主要集中在三个领域，即信息抽取、自然语言理解和对话管理。在信息抽取领域，本文关注于实现三个细分任务：命名实体识别、实体链接和关系抽取。对于自然语言理解，本文专注于识别用户的意图以及填补对应的槽值。例如，当询问“感冒可以吃哪些药？”时，涉及到的部分步骤如下：

（1）命名实体识别：在自然语言处理过程中，关键是辨认出问句中所含的具体实体，例如在图 4.1 展示的例子里，“感冒”代表的是一个疾病名称，它是需被辨识的实体对象。

（2）实体链接：对于自然语言处理中的实体链接问题，核心任务在于将问句中辨识出的特定实体—例如“感冒”，与已经建立的知识图谱中相对应的实体节点进行匹配，并建立连接。

（3）关系抽取：在实现问答系统时，必须对特定的自然语言提问进行解析，从而识别出与其内容最为接近的关系或属性。比如，对于“感冒可以吃哪些药？”这一问题具有与“疾病”一词具有最高的相似性。

（4）得到结果：系统将筛选出与之最为吻合的关系或特性以及相应的实体或属性值，并将其作为查询问题的答案反馈给用户。

在医疗知识图谱问答系统里，完整的处理流程将依照图 4.1 展开。