基于深度学习的中医典籍知识图谱自动化构建研究

金佩

北京科技大学

密　　级：公开

论文题目：**基于深度学习的中医典籍知识图谱自动化构建研究**

学　　号：S20160742

作　　者：金佩

专 业 名 称：软件工程

2018 年 11 月 10 日

基于深度学习的中医典籍知识图谱自动化构建研究

Research on Automatic Construction of Knowledge Graph of TCM classics Based on Deep Learning

研究生姓名：金佩

指导教师姓名：张德政

北京科技大学计算机与通信工程学院

北京 100083，中国

Master Degree Candidate： Pei Jin

Supervisor： Dezheng Zhang

School of Computer and Communication Engineering

University of Science and Technology Beijing

30 Xueyuan Road，Haidian District

Beijing 100083，P.R.CHINA

分类号：TP391　　密　　级：公开

ＵＤＣ：［单击此处键入 UDC 号］　　单位代码：１０００８

北京科技大学硕士学位论文

论文题目：基于深度学习的中医典籍知识图谱自动化构建研究

作者：金佩

指 导 教 师：张德政 教授　**单位：**北京科技大学

指导小组成员：谢永红 副教授　**单位：**北京科技大学

单位：

论文提交日期：2018 年 11 月 10 日

学位授予单位：北 京 科 技 大 学

致　谢

逝者如斯，转眼间研究生的学习生涯已接近尾声，心中有太多的不舍与感触。回首两年半来的点点滴滴，有迷茫，有艰辛，有感悟，有收获，感谢所有指导、支持、关心和帮助过我的老师、同学和亲人。这两年多来，你们的陪伴，给了我良好的学习环境与氛围，让我的研究生生活过得充实而有意义。

首先，向我的导师张德政教授以及谢永红副教授致以最由衷的感谢和最真挚的敬意。感谢两位老师从研究课题的选定、规划、实施到论文的撰写、审阅和定稿，一直以来给予我的悉心指导和无私关怀。倘若没有导师们的指导，就没有我如今理论水平的提高以及论文的顺利完成。在每次的科研或项目讨论会上，我都能感觉到张老师对人生目标的坚定信念，对项目的整体把控能力，对项目进展的高瞻远瞩以及在中医领域的深刻见地，这些深深地影响着我的学术心态。而与谢老师的每次交谈，都能感受到她治学的严谨、为人的真诚以及对生活的热爱，成为我研究生生活的美好回忆。

感谢计算机与通信工程学院的老师们在论文开题、中期报告过程中给予的宝贵建议，特别是刘宏岚老师、支瑞聪老师和李莉老师，你们的每一条谆谆教诲和对我论文的指正，都是我的论文在学术上更加严谨，研究上更为深入。

感谢在实验室一起奋斗的伙伴们，特别是贾麒学长、陈鹏学长以及张铮、丁瑞东和夏超。感谢大家在机器学习、自然语言处理方向上的集思广益和热烈讨论，让我找到了自己的兴趣点。也感谢大家在科研方向上对我的无私帮助和鼓励，你们的广博的思路和见解，增强了我克服困难的信心，给予我不断前进的动力。

同时，也要感谢这两年多来一直陪伴我的同学、室友和家人，你们的陪伴和信任，让我觉得自己不是孤单一人，才使我有充足的时间和精力去奋斗和拼搏。

最后，感谢在百忙之中，抽出时间审阅论文的老师和各位专家，恳请各位老师多多批评指正，并提出宝贵意见。真诚感谢关心和帮助过我的所有人。

摘 要

中医典籍是我国劳动人民在长期与疾病斗争中形成的丰富诊疗经验的总结。在当今智慧医疗的大背景下，从中医典籍中挖掘可理解、可应用的经验知识，构建可描述各种实体之间丰富关系的语义网，形成可实现语义搜索的知识图谱，辅助医生的临床决策，已成为中医数字化进程的重要环节。

为了解决中医典籍语义复杂、理解难的问题，减少人工干预，更加高效准确地获取中医典籍中的知识，实现精准的中医知识检索与推理。本文针对深度学习循环神经网络的结构特点，提出了一种中医典籍知识的快速自动获取方案。该方案中，首先深入总结了中医典籍的语言特点，提出了一种结合无监督学习快速获取的先验知识进行特征构建的方法，其次利用深度学习串行地进行实体抽取与关系抽取。

本文根据中医典籍的语料特征，创新性地将字向量训练与深度学习相结合，提出了中医典籍中中医认识方法、生理、病理、自然、治则治法的多实体识别方案和复合抽取六大类语义和层次关系的关系抽取方案。在实验语料中验证 F1 值分别达到了 85.32%和 90.53%，明显提高了对中医典籍中实体的识别准确率，在一定程度上解决了语料不足的限制。

此外，本文采用不同的深度学习模型处理中医典籍实体识别和关系抽取的任务，并进行了大量的对比分析，为不同任务匹配了恰当的模型和方法。最终，以（实体，关系，实体）的三元组形式进行知识表示，基于 Neo4j 进行可视化展示，完成中医典籍知识图谱自动化构建。

关键词： 知识图谱，深度学习，中医，无监督学习

Research on Automatic Construction of Knowledge Graph of TCM classics Based on Deep Learning

Abstract

Traditional Chinese medicine classics are a summary of the rich experience in the long struggle against diseases. Under the background of intelligent medicine, it has become an important link to excavate useful knowledge from Chinese medical classics, form knowledge graph which can realize semantic search and assist doctors in clinical decision-making.

In order to solve the problem of complex semantics of TCM classics, reduce manual intervention and ultimately achieve the retrieval and reasoning of TCM knowledge, this paper proposes a scheme for fast automatic acquisition of Chinese medical classics knowledge,which aims at the structural characteristics of deep learning cyclic neural network. Firstly, the linguistic features of TCM classics are summarized in depth, and a method of feature construction is proposed, which combines unsupervised learning with fast acquisition of prior knowledge.

According to the corpus characteristics of TCM classics, this paper innovatively combines character vector training with deep learning, and puts forward a multi-entity recognition scheme of cognitive method, physiology, pathology, nature and treatment in TCM classics and a relationship extraction scheme of six categories of semantic and hierarchical relations. In the experimental corpus, the F1 values were verified to be 85.32% and 90.53% respectively. This method obviously improves the accuracy of entity recognition in TCM classics, and solves the limitation of insufficient corpus to a certain extent.

In addition, different deep learning models are used to deal with the tasks of entity recognition and relationship extraction of TCM classics, and a large number of comparative analyses are made to match the appropriate models and methods for different tasks. Finally, the knowledge is represented in the form of triple tuple (entity, relation, entity) and visualized based on Neo4j to complete the automatic construction system of TCM knowledge graph.

Key Words：Knowledge Graph, Deep Learning, The Chinese Medicine, Unsupervised Learning

目　录

插图和附表清单

1 引言

1.1 课题背景与意义

中医药学是中华民族最宝贵的财富，富含朴素的唯物辨证主义哲学思想。目前，促进中医药发展已上升为国家发展战略。2015 年 5 月，国务院办公厅印发《中医药健康服务发展规划（2015—2020 年）》，对当前和今后一个时期的中医药健康服务发展进行全面部署，明确指出 “中医药信息化势在必行[1] ”。2016 年 2 月，国务院印发了《中医药发展战略规划纲要（2016—2030 年）》，明确提出要推动中医药的继承，实施中医药传承工程，将中医典籍文献的整理纳入国家中华典籍整理工程，推动中医古文典籍数字化。

在如今大数据和智慧医疗的大背景下，如何从海量、碎片化、异构的医疗信息中获取所需的知识，更好地为人们服务，成为社会关注的热点。而知识图谱能够更加直观地展示实体之间的关系，为整合和组织医疗知识提供了非常有效的途径，成为了研究热点。从中医古籍中的挖掘可理解的、应用的经验知识，构建可以描述各种中医实体之间丰富关系的语义网，形成能够实现真正语义搜索的知识图谱，使其辅助医生进行临床决策，已成为中医数字化进程的重要环节。

本课题来自于实验室承担的国家重点研发计划重点专项——“大数据驱动的中医智能辅助诊断服务系统”（编号：2017YFB1002300）中的任务 4：“中医临床智能辅助诊断与决策推荐”，其主要工作是构建中医药本体化知识图谱，构建基于中医药大数据的类人认知体系架构和思维机理，研制人机交互的场景化中医临床智能辅助诊断与决策推荐机制。

目前已有的知识图谱大多为通用知识，缺少领域知识，不能满足中医领域的个性化服务需求。此外，中医古文典籍是我国劳动人民在长期与疾病斗争中形成的丰富诊疗经验的总结，深受古代文学及哲学影响，形成了一种以阴阳五行作为理论基础的独具特色的文本表现形式[2] 。语言表述模糊抽象，具有文学色彩，需一定的专业知识进行理解，已有的研究大多还是针对中医医案中的结构化部分实现半自动的文本挖掘和信息抽取，耗费大量人力。该项目前期已收集整理了数十万份中医医案，从中抽取出了症状、方剂、药味药性等知识，可用于辅助诊断。但事实上，如《黄帝内经》之类的中医典籍，才是中医理论的源头，是指导中医诊断的核心依据，也是研究难点所在。

本文在项目已有的基础上，根据中医典籍的语言特征，创新性地将字向

量训练与深度学习相结合，提出了中医典籍知识的快速自动获取方案，采用不同的深度学习模型处理中医典籍实体识别和关系抽取的任务，实现我国现存最早的医学著作——《黄帝内经》粗图谱的自动化构建，从而探究出一个中医典籍知识图谱自动化构建的解决方案。从中医典籍中挖掘出了大量且多样的隐藏知识，对于传承中医的学术思想具有重要的意义，同时补充了中医大数据知识图谱在中医典籍方面的空缺，为辅助诊疗提供知识支撑，处理好术语规范与辨证论治的个性化诊疗模式，解决了中医药信息化的关键问题。

1.2 国内外研究现状

当今社会是个信息爆炸的时代，如何将大量文本数据中抽取有价值的信息，便于查询、检索，已成为研究开发的焦点[3] 。2012 年，谷歌首次提出知识图谱概念，把各种实体和概念整合在一起，以“关系”的视角来分析和研究问题，引起了业界和学术界的广泛关注。知识图谱技术解决了与实体相关的智能问答问题，加速了语义搜索的发展，由此创造出一种全新的信息检索模式，能更好地满足人们的实际信息需求。

1.2.1 知识图谱的研究现状

知识图谱，也称科学知识图谱，是一个结构化的语义网络，利用“实体－关系－实体”这样的三元组，来描述物理世界中的概念及其相互关系。实体间通过关系连接，构成网状的知识结构。通过知识图谱，可以让 Web 实现从仅仅是网页之间的超链接转向包含描述各种实体和实体之间丰富关系的概念链接，支持用户真正实现语义检索，将结构化的知识以图形的方式反馈给用户，用户不必浏览大量网页，就可以准确定位和深度获取知识，从而拓宽了搜索引擎的广度和深度。

2012 年谷歌公司首次推出知识图谱概念[4] ，从维基百科和 CIA 中获取了超过 570 亿个对象和对象之间丰富的关系，为用户提供了更加精确直观的搜索结果。此后，越来越多的知识图谱及相关产品相继出现，目前世界上规模较大的知识库已经多达 50 余种。知识图谱的构建方式有自顶向下和自底向上两种。在发展初期多数企业和机构都是采用和谷歌类似的自顶向下的方式进行构建，即基于维基百科等在线百科知识，从高质量数据中提取本体和模式信息，加入到知识库中，例如 DBpedia[5] 、Freebase[6] 、YAGO[7] 、Omega[8] 和 WikiTaxonomy[9] 等。而伴随着知识抽取与加工技术的不断成熟，现在的知识图谱大多采用自底向上的方式，即从开放网络的数据中提取资源模式，选

择其中置信度较高的新模式，经人工审核之后，加入到知识库中，具有代表性的有 KnowItAll[10] 、TextRunner[11] 、NELL[12] 以及目前拥有概念最多的知识库微软 Probase[13] 和其应用 Satori 知识库。Satori 能够利用已有的上亿级别的实体和关系不断学习，为其搜索引擎必应(Bing)提供更好的搜索支持。

我国对中文知识图谱的研究虽然起步较晚，但也已取得了许多有价值的成果。早期主要采用手工构建的方式，依靠专家知识编写一定的规则，例如中国科学院的知网(HowNet)，其知识的质量较高，但规模较小，有较强的领域限制，且耗费大量人力物力。近年来，如何自动化构建知识图谱成为研究热点和难点。在业界，各家公司纷纷推出了知识图谱的商业应用产品，如百度“知心”和搜狗“知立方”等。“知心”的数据来源于百度百科和用户的搜索日志，从半结构化的数据中抽取相关实体的属性—值。“知立方”从互联网文本中抽取实体间关系，并用语义推理、消歧等技术完善实体信息，从而提高用户搜索意图的识别准确率。在学术界，上海交通大学发布了中文知识图谱研究平台 zhishi.me；复旦大学 GDM 实验室推出的中文知识图谱项目[14] 等。这些项目的知识库规模较大，涵盖的知识领域较广泛。

然而，以上这些知识库大多是通用的常识内容，缺少领域知识。此后虽然相继出现了一些针对特定领域的知识库，如针对社区内容的 FOAF、针对电影内容的 LinkedMDB、专注数学领域且包含实体最多的 WolframAlpha，以及清华大学构建的首个大规模中英双语知识图谱 Xlore、中国科学院基于开放知识网络(OpenKN)构建的“人立方、事立方、知立方”原型系统。但是这些领域知识库的对象多为人名、地名、机构名等公共实体或公众关注的知名实体，并不满足许多特定领域的应用需求，对关系的抽取则局限在上下位等简单的层次关系上。

目前，国内知识图谱的研究领域集中在情报学、教育学、体育学、管理学等为数不多且较狭窄的知识领域之内，研究对象多以结构化较强的现代文献为主，由于中文分词、句法分析等基础自然语言处理技术在专业领域的效果限制，基于有监督和半监督的研究方法占据着国内知识图谱研究领域的半壁江山。如何自动地构建大规模的领域中文知识图谱，实现实体间复杂语义关系的抽取仍是业界研究的热点。

1.2.2 中医知识图谱的研究现状

近年来，医疗数据的飞速增长和知识图谱技术的逐步成熟，如何利用海量医疗数据更好地服务人们得到了广泛的关注，越来越多的企业和机构开始

重视医疗知识图谱的构建。

2015 年 2 月，谷歌公司宣布开始着重医疗资讯在搜索结果页面的地位，把搜索引擎、知识图谱和在线医疗进行深度整合。2016 年，谷歌的医疗知识图谱正式在印度上线，在用户搜索疾病或症状时为他们提供超过 400 种健康状况的数据，这些数据由专业的医疗机构进行审核发布。在线医疗的形式给医疗资源匮乏的国家和地区带去了福音，也极大程度地促进了智慧医疗的发展。由于医学数据碎片化、多样化情况复杂，医学专家们开始致力于构建统一的医学领域标准。目前，国外已经有了一些非常成熟的生物医学领域知识库，比如 Gene Ontology、Disease Ontology、MeSH、OMIM、HPO、CYC 等。随后，美国统一医学语言系统 UMLS 开始整合上述知识库在内的 100 多个知识库，一共收录了 300 多万个生物医学概念和 1200 多万个概念名称，同时提供医疗领域词汇映射，使同一词汇能在不同术语系统中转换，一共整合了 800 多万个 RDF 三元组和 37 万个以上的 RDF 链接。

国内外目前主要的研究方向是针对电子病历和开放网络中医学数据构建医疗知识图谱，虽然国内研究生物医学文本挖掘的起步较晚，但也取得了一些成果，如清华大学的“生物关系信息挖掘评价与融合方法研究与实现”项目与中国科学院上海生命科学院的“生物医药信息数字化决策支持系统”项目等，都极大地为知识图谱在生物医学领域中的应用提供了有力支撑。在工业界，许多公司也推出了基于知识图谱的医学产品，如湖南格尔智慧公司的智慧护理系统、北京康夫子科技有限公司的智能诊断系统等。而在中医领域，早从 2001 年开始，我国便已经开始建立一个统一标准的中医药术语本体系统。中国中医科学院联合全国其他 30 多家中医研究单位，建立了“中医药学一体化语言系统”，共编录了 16 个一级项目，12800 多个类[15]，是最大的传统的中医医药本体，对实现中医药知识的标准化起到了很重要的作用。

随后，学术界也开始对中医垂直领域知识图谱构建与应用进行深入研究。庄力从中医药文献资料中抽取结构化中医临床诊疗信息，实现了中医临床诊疗垂直搜索系统 TCMVES。华东理工大学的王昊奋[16] 利用文本挖掘、关系数据转换以及数据融合等技术，探索中医药知识图谱自动化构建方法与标准化流程，以实现基于模板的中医药知识问答和基于知识图谱推理的辅助开药。中医科学院的贾李蓉[17] 从信息获取、知识抽取、图形化展示几个方面介绍了对中医知识图谱构建的研究工作，并预计开展研究基于中医药知识图谱的检索系统、知识地图和中医药知识图谱维基百科等一系列应用。张梅奎等[18] 构建了针对中医脑疾病的本体知识，张德政等[19] 实现了基于本体的中医知识图谱构建，郝伟学[20] 利用临床病例及百科知识实现了中医健康知识图谱的构建。

由于中医语言较为抽象，带有文学色彩，需要专业的中医知识去理解，目前国内研究大多集中于中医临床文本，如中医医案等，且仅限于中医简单术语的自动化抽取，如症状、病名、脉象、方剂、中药等。在构建中医知识图谱方面，真正落实到位的并不多，大部分研究还停留在理论和实验层次。网上虽然出现了一些中医专家问答系统，但是对于人们提出的各种关于中医的问题，给出的答复并不能使人满意。这说明目前我国还未构建较为专业、完整、可应用于问答系统的中医领域知识图谱，来达到中医知识共享的目的。由于传统中医药学历经了几千年的发展，是在临床诊断、实践中不断扩充的中医知识，使得老中医对中医知识的总结表述上带有主观随意性，对于中医知识没有统一的描述，异构性大，给信息利用以及共享造成了很大的难度。而追根究底，中医起源于《黄帝内经》等中医典籍，归于阴阳五行。《黄帝内经》等中医典籍揭示了中医最根本的理论体系，如果能实现中医典籍知识图谱的自动化构建，将对中医的发展和规范起到很大的促进作用。因此，本文研究的重点和意义就在于实现《黄帝内经》等中医典籍的粗图谱自动化构建，形成中医典籍知识获取的解决方案，从顶层理论知识入手，为中医知识的统一化、结构化提供支撑。

1.3 课题研究方法与内容

本文首先分析研究收集到的 701 本中医典籍的语言特点，以我国现存最早的医学著作《黄帝内经》为核心，结合信息抽取和中医药信息化等领域的研究成果，针对中医典籍命名实体识别和实体关系抽取存在的难点，探索出一种中医典籍知识图谱的自动化构建方案，包括知识获取、知识表示和知识可视化。主要过程为利用自然语言理解技术，如分词、词性标注、句法分析方法等进行中医典籍的领域词表构建和特征提取；结合深度学习循环神经网络的特点，提出自动高效地进行中医典籍的串行命名实体识别和实体关系抽取方法，构建《黄帝内经》粗知识图谱并利用图数据库实现可视化，为中医药大数据知识图谱的补充和扩展打下坚实基础。

(1) 提出了一种中医典籍知识的快速自动获取方案。首先基于无监督学习快速获取先验知识、提取特征，减少人工干预和对专业知识的依赖；其次利用深度学习方式进行中医典籍知识的自动获取。

(2) 分析中医典籍，如《黄帝内经》的语言特色，确定中医基础理论体系，构建中医典籍的领域基础词表。同时，提出了一种结合无监督学习快速获取的先验知识进行特征构建的方法，利用关键词提取、依存句法分析等自然语言处理技术扩充领域词表，规范中医实体的类别和实体之间的关系类别。

(3) 分析了当前主流的深度学习模型的特点，创新性地将中医字向量训练与深度学习相结合，提出基于双向长短时神经网络的实体抽取与关系抽取的串行实现方法，明显提高对中医典籍中复杂实体和关系的识别准确率，在一定程度上解决了语料不足的限制。

(4) 采用不同的深度学习模型处理中医典籍实体识别和关系抽取的任务，并进行了大量的对比分析，为不同任务匹配了恰当的模型和方法。

(5) 实现《黄帝内经》复杂、全面的粗图谱构建，以（实体，关系，实体）的三元组形式进行知识表示，并基于图数据库 Neo4j 实现了领域知识图形化展示。

1.4 论文的组织结构

本文共分为七个章节，各章节主要内容如下：

第 1 章：主要介绍了中医典籍的地位以及目前中医典籍知识的准确率低、缺乏权威性等缺陷，引出大数据及智慧医疗大背景下构建中医典籍知识图谱的优势和重要性，表明了本文研究内容的背景与意义。介绍了知识图谱，特别是中医知识图谱的国内外研究现状，证明本文研究内容的创新性。此外，还介绍了本文的研究方法与论文的组织结构。

第 2 章：主要介绍了知识图谱的关键任务以及现有的实现技术，详细介绍了本文构建中医典籍知识图谱所涉及的相关理论，介绍了 Word2vec、条件随机场、深度学习的神经网络模型，包括 CNN、IDCNN、RNN、BiLSTM 以及注意力机制的原理和适用场景，表明了本文的研究重点及方法的先进性和合理性。此外，还介绍了一种图数据库 Neo4j，用于实现知识图谱可视化。

第 3 章：首先对中医典籍的语言特点进行分析，并根据其特点手工构建了领域基础词表。此外提出了一种结合无监督学习快速获取的先验知识进行特征构建的方法，利用关键词提取、依存句法分析以及迭代思想扩充领域词表，构建了中医典籍以动词为核心的三元组，利用层次聚类提取特征，大大减少了深度学习训练集的人工标注工作量，减少了对领域知识的依赖。

第 4 章：提出了一种基于字向量的 BiLSTM-CRF 的实体抽取模型。模型以中医典籍字向量作为输入，利用 BiLSTM 获取长句中的上下文信息，最终 CRF 根据训练集中的规则完成全局标注。并且对该方法进行算法实现与实验验证，介绍了实验数据集以及评价标准，并且设计了多组对比实验加以分析，验证模型的有效性。

第 5 章：提出了在字级别和句子级别均添加了注意力机制的 BiGRU 关

系抽取模型。模型同样以字向量和位置向量作为实体对的唯一标识。利用注意力机制，为句子中重点的字以及训练集中重点的句子设置权重，充分利用语料，去除噪声影响。最后采用 softmax 对语句向量进行多分类。本章节中也介绍了实验数据集，并且为验证模型的有效性进行了多组对比实验。

第 6 章：主要设计并提出了一个针对中医典籍的知识图谱自动化构建的解决方案，介绍了该方案的整体架构，并详细讲解了知识获取、知识表现和知识可视化三个模块的实现方法，从而实现了《黄帝内经》知识图谱的自动化构建。

第 7 章：总结了全文的研究内容和主要贡献，分析了仍可以改进的地方，并给出了下一步工作计划。

2 相关理论和方法介绍

本章介绍了知识图谱的基本概念与关键任务的技术发展过程，验证了知识图谱技术的优势和技术的合理性；并讲解了构建中医典籍知识图谱所涉及到的相关技术和理论，如词向量、条件随机场、深度学习神经网络模型以及图形数据库的理论与方法等。

2.1 知识图谱技术综述

知识图谱本质上是结构化的语义网络，以图的形式进行存储。存储结构由节点和节点联系组成。在知识图谱中，真实世界中的各种事物被抽象为一个个节点，而各种事物的相互关系则被抽象为节点间的连线。知识图谱提供了一种更为直观的方式观察真实世界中的关系网络。知识图谱起源于万维网之父 Berners Lee 于 1998 年提出的语义网(Semantic Web)和在 2006 年提出的关联数据(Linked Data)，是对现有语义网络技术的一次扬弃和升华。

知识图谱在逻辑结构上分为数据层和模式层。在数据层中，知识以“实体—关系—实体”或“实体—属性—值”三元组的形式存储。所有的三元组相互关联组成了关系网络，构成了知识的图谱。模式层是对知识进行规范整合。因本体库可以通过对规则、约束进行定义从而实现知识的规范化，所以常用本体库对模式层进行管理。知识图谱的构建方式一般分为两种：自顶向下的方式和自底向上的方式。自顶向下的构建方式是基于本体构建的方式，以结构化程度高的百科类等网站为数据源，从中抽取本体和规则约束，填充到知识库中；而自底向上的构建方式是直将收集的数据通过模式识别、制定规则等方式，从中识别实体、属性以及关系，然后加入到知识图谱中。

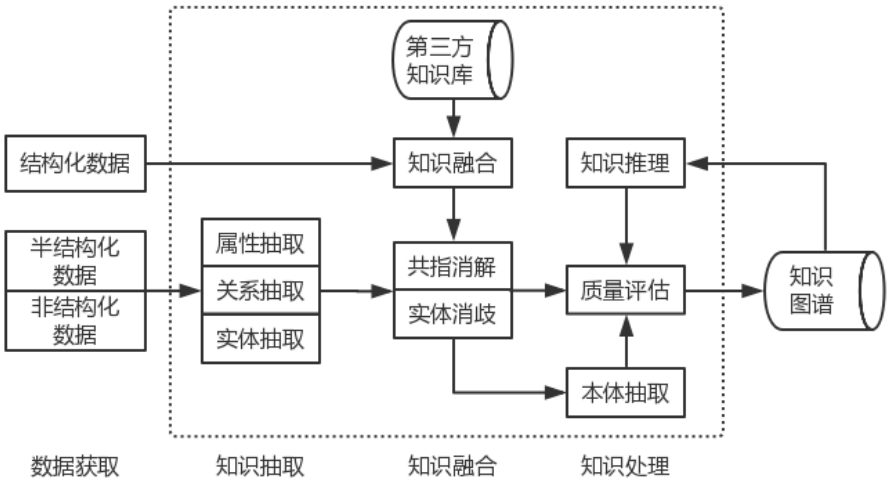

图 2-1　知识图谱技术架构图

由图 2-1 可知，方框内知识的抽取、融合与处理三步骤是知识图谱构建的核心。而结构化的数据因为规范化程度较高，可以较为容易的从中抽取知识；半结构和非结构化数据规范性较差，难以直接获取知识，因此需要借助属性抽取、关系抽取、实体抽取等一系列操作提取出知识的实体和关联，然后存入知识库中。知识图谱的构建过程是一个迭代更新的过程，包括知识抽取、知识融合和知识加工。本文以知识抽取为研究重点，将中医典籍知识图谱构建技术归纳为四部分，即知识获取、知识表示、知识存储和知识可视化。

2.1.1 知识获取

知识获取的关键人物是命名实体识别和实体关系抽取，旨在通过信息抽取技术从百科类数据、文档类数据以及网页数据等抽取结构化信息（实体、实体间的关系）。当前的知识获取方法主要分为两种，一种是先进行实体抽取，再进行关系的抽取的串行知识获取方法；另一种联合抽取方法，对于给定的句子通过联合抽取模型直接得到实体三元组。如何实现知识获取自动化，在减少人工干预的同时，提升信息抽取的准确率，一直以来都是业界和学术界的研究热点。

1) 命名实体识别

命名实体识别(named entity recognition, NER)是指从文本数据集中自动识别出文本中出现的专有名称和有意义的数量短语并加以归类，最初在第 6 届信息技术研讨会(MUC-6)[21] 上作为一个任务被提出。实体抽取的质量（准确率和召回率）对后续的知识获取效率和质量影响极大，因此是信息抽取中最为基础和关键的部分。在生物医学领域比较集中针对医学文献中的基因、蛋白质、药物名、组织名等相关生物命名实体识别[22] 。在中医药领域，主要的研究集中在中医术语的自动识别，即从中医文献中识别出症状、病名、脉象、方剂、中药等术语实体，以便供进一步查询或分析使用。例如，范岩[23] 从中药复方的临床文献进行复方、疾病名称的抽取。但由于中医典籍是中医的起源，表述与现代汉语区别较大，且涵盖的实体范围更广，本文在研究《黄帝内经》的基础上，提出中医典籍实体的类别，将从生理、病理、诊法等方面更全面地对中医典籍进行实体识别。

传统的命名实体识别的方法主要可以分为两大类：基于规则的方法和基于统计的方法。早期的研究主要针对特定领域，采用人工规则的方式。Fukuda[24] 基于术语名称所在的上下文的词法特点，采用一定的类启发式规则，实现蛋白质名称的抽取，在 80 篇 Medline 文摘上试验取得 98.84%召回率和

94.7%准确率。然而基于规则的方法不仅需要耗费大量人力，而且可扩展性较差，难以适应数据的变化，在大数据时代下已退出了历史舞台。

随后，越来越多的人开始尝试基于统计的方法，其无需构建领域专有的模式，而能有较强可扩展性，主要包括最大熵模型(Maximum Entropy, ME)、隐马尔科夫模型(Hidden Markov Model, HMM)[25] 、条件随机场模型(Conditional Random Fields, CRF) [26] 和支持向量机(Support Vector Machine, SVM) [27] 等。在生物领域，Okanohara[28] 使用改进的半监督条件随机场模型进行生物命名体的识别，识别蛋白质、DNA 和 RNA 等命名实体。在中医领域的研究不多，主要有王世昆[29] 等人采用基于条件随机场的方法对明清古医案中症状、病机进行了自动识别标注，验证了该方法效果优于最大熵及支持向量机; 张五辈[30] 等人利用条件随机场对《名医类案》进行了术语抽取试验，其结果准确率 83.11%，召回率 81.04%，F1 值 82.06%；孟洪宇[31] 等人利用条件随机场对中医古籍的《伤寒论》进行中医术语抽取，准确率 85.00%、召回率 68.00%、F1 值 75.56%。相比于 MUC 所报道的关于新闻领域信息抽取 90%以上的 F1 值。可以看出针对中医药领域，特别是中医典籍的实体识别更具有挑战性。此外，统计的方法识别性能很大程度上依赖于特征的准确度，需要大量标注好的语料，成本昂贵；受分词效果的制约，通用性差，并且无法解决单词前后长距离的依赖。

加拿大多伦多大学的 Hinton 教授[32] 提出深度学习的概念，在全球掀起一次热潮。相对于以 CRF 为代表的传统机器学习方法，深度学习模型可以充分逼近任意复杂的非线性关系，而且有很强的鲁棒性、记忆能力、非线性映射能力以及强大的自学习能力，在原始字符集上提取高级特征，大大降低了人工标注的影响，在实体识别任务上收到了广泛的成功，成为了目前实体识别任务的主流方法。国内也开展了大量利用深度学习方法进行中文命名实体识别的研究，识别的主要是人名、地名和机构名等通用实体。Yonghui Wu[33] 等人首先用深度神经网络从大量未标注的语料中训练词向量，然后用另一个深度神经网络进行命名实体识别，在生物学语料上 F1 值达到了 92.8％，超过了最好的 CRF 模型。Zhiheng Huang[34] 等人使用双向循环神经网络和条件随机场模型进行命名实体识别，在 CoNLL2003 数据集上 F1 值达到了 90.10%。

在中医领域主要利用深度学习方法对电子病历、中医医案及中医文献进行实体识别，识别类别多为症候、病名、穴位和药方等。例如，张帆[35] 利用深度神经网络模型从胃癌、糖尿病、哮喘、高血压四类疾病语料中抽取疾病、症状、药品、治疗方法和检查五类实体，准确率 88.03%、召回率 82.34%、F1 值 85.08%；薛天竹[36] 利用双向 LSTM 对电子病历进行术语抽取，F1 值达到

了 88.36%。步君昭[37] 等人利用循环神经网络和条件随机场结合的方法, 抽取生物医学文献中的药物名，F1 值达到 88.76%，效果优于常用的条件随机场算法和标准的循环神经网络方法。从调研结果看来，基于深度学习进行命名实体识别效果要优于传统方法，而进行中医古文的实体识别技术还鲜有研究，本文创新性地探究深度学习技术在中医典籍上的多实体的识别效果。

2) 实体关系抽取

文本经过处理得到一系列离散的命名实体后，还需要从相关语料中提取出实体之间的关联关系，通过关系将实体联系起来，才能够形成网状的知识结构。研究关系抽取技术的目的，就是解决如何从文本语料中抽取实体间的关系这一基本问题。领域实体之间的关系有两种：层级关系（分类关系或上下位关系）和非层级关系（语义关系）。层级关系决定着知识图谱的深度，非层级关系标志着知识图谱的广度。

对于领域实体层级关系的自动获取，国内外主要有两种方法：基于统计和词典的方法、基于模式的方法。基于词典和统计的方法通过对词典中概念定义形式上的规律性进行挖掘，然后寻找给定概念的上位词，精度高是其明显的优点。2008 年，Sumida A[38] 等人借助英文版维基百科进行上下位关系抽取，取得了 75.3%的准确率。然而，中医领域目前并没有专业、全面的领域词表，还不能满足特定领域计算的要求。

基于模式的方法主要分为两类：Hearst pattern 和自扩展 Bootstrapping 模式。Hearst pattern 采用将语言学和自然语言处理相结合的技术，使用词法和语法分析获取上下位关系模式，然后利用模式匹配获取上下位关系。2006 年，刘磊[39] 等人提出一种基于“是一个”模式的方法，取得了一定成功。这种方法虽然精度高，但是召回率很低，移植性很差。自扩展模式是从少数几个种子抽取模式开始，通过迭代发现新的抽取模式，将置信度高的模式合并到当前模式集；2012 年，Tian F[40] 等人提出基于 Bootstrapping 模式的上下位关系抽取，取得了不错的效果。到目前为止，基于自扩展模式及其变形，一直备受业界青睐。雷春雅[41] 等人利用自扩展与最大熵结合对旅游领域的四大类实体关系进行了抽取。

由于语言结构的复杂性，抽取实体之间的语义关系是构建知识图谱的巨大挑战之一。国内外目前对于领域实体语义关系的自动获取主要采取 3 种方法：基于特征的方法、基于核函数的方法和基于深度学习的方法等。基于特征的方法在语义关系抽取领域已取得了较好的效果。2005 年，Guodong Z[42] 等人利用 SVM 作为分类器，分别研究词汇、句法和语义特征对实体语义关系抽取的影响。2007 年，董静[43] 等人对句法特征进行划分得到新的句法特

征，采用 CRF 模型进行训练，提高了实体关系抽取性能。该方法的特点在于构造分类效果好的持征和选取合适的机器学习模型，特征抽取主要依赖于人工选择，特征的好坏会直接影响到关系抽取的性能。

基于核函数的方法是通过构造核函数，隐式地计算特征向量内积，从而得到关系实例之间的相似性。刘克彬[44] 等人借助知网提供的本体知识库构造语义核函数，在开放数据集上对 ACE 定义的 6 类实体关系进行抽取，准确率达到了 88%。但该方法依赖于自然语言处理过程，误差累积极大地影响抽取性能。且在同一个文本中识别两个实体间关系，常由于长距离实体或隐式关系的特点，或不具有足够的领域特征而无法准确识别出领域实体的关系。

基于深度学习的方法很好的解决了长距离的问题，将低层特征进行组合，进而形成更加抽象的高层特征，用来寻找数据的分布式特征表示，避免了利用 NLP 工具对语料进行各种预处理，人工选取特征等步骤，很好的改善特征抽取过捏中的误差累积问题。目前，基于深度学习的实体关系抽取方法在性能上相比传统的方法要好，取得了相当多的成果。Zhang 等人通过使用循环神经网络(RNN)做关系抽取，更好的利用了实体的上下文信息。Zeng[45] 等人用卷积神经网络(CNN)进行关系抽取。2012 年，陈宇[46] 等人基于深度学习的中文名实体关系抽取方法，取得了不错的效果；2013 年，Liu[47] 等人提出的将同义词词典和语义知识融入卷积神经网络，并结合词汇特征，在英文的关系抽取任务中取得了成功。

在医学领域，已经有一部分人开始尝试利用深度学习的方法进行关系抽取，但局限于药名与病名，症候与治法等简单关系的抽取，扩展性较差。例如，曾东火[48] 利用卷积神经网络进行药物之间的关系识别，证明了效果优于传统的 SVM 方法。冯钦林[49] 利用卷积神经网络抽取疾病-症状和病症-治疗物质两类关系，F1 值分别为 83.64%和 86.74%。蒋振超[50] 利用逻辑回归和 LSTM 抽取药物之间的关系。郑洁琼[51] 提出了基于动态拓展树的双向 LSTM 框架，在生物医学文本中进行关系抽取，得到了 58.15%的 F1 值。杨晨浩[52] 根据 I2B2 提出的电子病标注规范，利用 RNTN 对中文电子病历进行了关系抽取。而在中医古典籍上的关系抽取，几乎还是一片空白。

2.1.2 知识表示

知识表示是为描述世界所做的一组约定，是知识符号化、形式化、模式化的过程，主要研究计算机存储知识的方法，其表示方式影响系统的知识获取、存储及运用的效率。然而医学数据种类繁杂、存储方式不一、电子病历

格式和标准不同、经常涉及交叉领域等特点，导致医学领域与其他领域在知识表示 方面有所差异，同时也给医学领域的知识表示带来极大的挑战。早期医疗知识库运用的知识表示方法有谓词逻辑表示法、产生式表示法、框架表示法、语义网表示法等，这些方法由于表示能力有限且缺乏灵活性，不再作为主要的知识表示方法，更多是作为医学知识表示的辅助或补充。

本体表示法以网络的形式表示知识，即以（实体，关系，实体）三元组来表示相关联的两个节点，在知识图谱提出之后逐渐得到认可。它借鉴了语义网表示法，但又有所区别，本体关注的是实体固有特征，比后者更聚焦、更深入，因而也具有更大的发展潜力。而本体的描述语言也多种多样，主要有 RDF 和 RDF-S、DAML、OWL 等。使用本体表示医学术语可以提升数据整合能力，建立强大、可互操作的医疗信息系统；满足重用共享传输医疗数据的需求；提供基于不同语义标准的统计聚合。医学领域本体的构建需要深入分析医学术语的结构和概念，才能将晦涩甚至是跨语言的医学知识有效地表达出来。

2.1.3 知识可视化

知识获取和整合后需要储存于某一介质，用于后续的查询和可视化展示。在传统的数据存储领域，关系型数据库是数据存储的主流，它具有很好的理论基础，而且具备很高的数据一致和安全性，能够利用简单的数据结构表达比较丰富的语义信息。然而，随着互联网技术的快速发展，数据量在不断的增加，关系型数据库在解决复杂的关系查询时出现越来越多的缺点，比如，消耗资源大、速度慢等。针对关系型数据库出现的送些缺陷，非关系数据库(NoSQL)就应运而生。

图形数据库对节点和节点间复杂关系的良好支持，成为了知识图谱存储的首选。图形数据库是非关系数据库中的一种，它可以将基本元素按照一定的结构存储起来。图形数据库主要包括 3 个基本元素：节点(Node)、关系(Relationship)和属性(Property)。图 2-2 是一个简单的图模型：

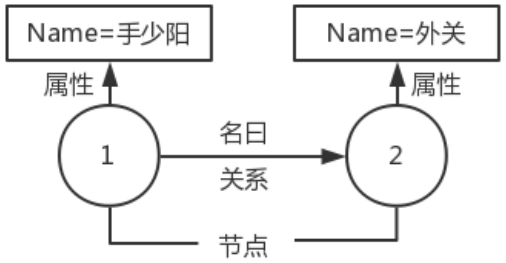

图 2-2 图模型示例

在上图中，两个圆点代表的是结点，中间带箭头的线代表关系，结点和关系旁边的注释代表属性。图形数据库可以将点、线和面存储起来，可以直观的将实体及实体间的联系展示出来，是构造和展示结构化信息的有效方式。

Neo4j 是一个成熟的具有较高性能且很稳定的图形数据库，是一个嵌入式的、基于磁盘的、具备完全的事务特性的 java 持久化引擎，具有以下特点：

(1) 具有完整的 ACID 支持。ACID 是数据库事务能正确执行的基本要素，包括原子性、一致性、隔离性和持久性，是保证数据一致性的基础。Neo4j 完整的支持 ACID，可以保证数据的准确性。

(2) 高可用性。Neo4j 可以轻松的和任何应用进行集成，不受业务约束。

(3) 可扩展性。Neo4j 可以进行分布式的集群部署，可以轻易进行扩展。

(4) 可以高速的检索数据。Neo4j 提供的遍历工具，可以高效的检索数据，提高数据检索效率，理论上可达到每秒上化的检索量。

(5) 简单便捷的安装方式。只要装有 JDK 的机器都可以进行安装。

Cypher 是 Neo4j 的查询语言，类似于关系型数据库的 SQL 语言，可对图形数据库进行增删改查的图形查询语言。而且 Cypher 不仅语法非常的简单，而且功能却十分强大，可以实现 SQL 难以完成的功能。完整的 Cypher 查询语句通常由如下结构构成：

(1) 匹配子句：通过对数据库中的实例进行匹配查询，从而获取满足查询条件的数据。

(2) 条件子句：作用类似 SQL 中的 WHERE 子句，通常作为匹配子句的组成部分，用于条件筛选。

(3) 返回子句：用于返回指定的信息字段。

(4) 创建子句：用于创建节点、关系或属性等。

Neo4j 作为功能强大的图形数据库，不仅可以通过 Cypher 添加数据，还支持从 CSV 或者其他关系型和非关系型数据库批量的导入数据。另外，Neo4j 还提供有面向 JAVA 和 Python 的 API 可以直接调用，用户还可以方便的通过 API 进行数据操作和应用开发。Neo4j 作为一个无框架数据库，在开始添加数据之前，并不需要定义表和关系，一个节点可以具有任何属性，任何节点都可以与其他任何节点建立关系。Neo4j 中的数据模型隐含在它存储的数据中，而不是明确地将数据模型定位为数据库本身的一部分，它是对想要存入数据库的数据的一个描述，而不是数据库的一系列方法来限制将要存储的内容。

2.2 词向量研究综述

词嵌套(Word Embedding)，俗称“词向量”，可将词语表征为高密度的低维实数向量，可以很好的表征词语之间的词法、句法及语义方面的信息，目前使用相当广泛。在传统的自然语言处理任务中，通常采用 One-hot Representation 方法进行词汇的向量化转换。One-hot Representation 首先创建一个词表库，采用稀疏表示法，把文档中所有的词进行顺序编号，把每个词都表示成一个向量，向量的维度和词汇表中词的个数相同，每个词对应的向量只有一个维度 1，其余的全是 0。但该方法存在着“词汇鸿沟”，词语之间相互独立，难以捕捉词与词之间的依赖关系，同时还很可能发生维度灾难。

1986 年 Hinton 提出一种词向量表征方法 Distributed Representation，该方法通过训练将每个词映射成 K 维的实数向量（K 为模型中的超参数），通过计算计算词之间的距离（比如余弦相似度、欧氏距离等）来判断它们之间的语义相似度。2003 年，Bengio 等人提出了基于前馈神经网络的语言模型(NNLM)，该语言利用句子前面出现的词作为上下文信息来预测下一个单词，该模型时间复杂度高，训练速度慢。后期很多学者致力于减少该模型的训练时间。2011 年，Mikolov 等提出了递归神经网络语言模型(RNNLMs)，通过简化神经网络将句子表示成一个向量。2013 年，Google 研究员开源了可将词高效地表征为实数值向量的向量工具 Word2Vec。

在自然语言处理时，将中文的字或词作为特征，利用 Word2Vec 就可以把特征映射到 K 维向量空间，可以为文本数据寻求更加深层次的特征表示。其利用深度学习的思想，通过训练，把对文本内容的处理简化为 K 维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似。

Word2Vec 主要包括 CBOW（Continuous Bag-Of-Words，即连续的词袋模型）和 Continuous Skip-Gram Model 两种不同的方法。这两种方法都利用人工神经网络作为它们的分类算法，针对大规模语料进行向量转化，具有较低的时间复杂度。起初，每个单词都是一个随机 N 维向量。经过训练之后，该算法利用 CBOW 或者 Skip-gram 的方法获得了每个单词的最优向量。前者根据连续的词汇共现来进行建模，后者则采用跳跃式的词共现进行建模，模型结构如图 2-3 所示。

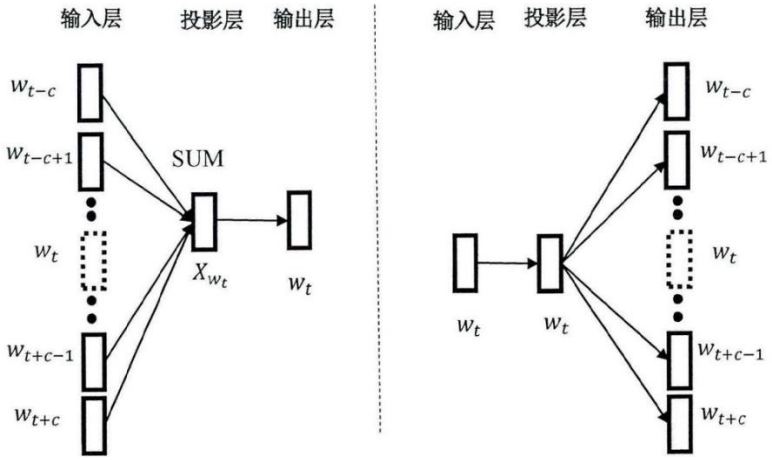

图 2-3 CBOW 模型结构与 Skip-Gram 模型结构

图 2-3 可知，这两个模型都包括输入层、投影层和输出层。CBOW 模型的目标是在已知当前词 w_t 的上下文 Context(w_t)的情况下预测当前词，而 Skip-Gram 模型的目标则是在已知词 w_t 的前提下预测其上下文 Context(w_t)。此处上下文定义如公式(2.1)所示：

$$Context(w_t) = w_{t-c},...,w_{t-1},w_{t+1},...,w_{t+c} \tag{2.1}$$

其中，c 是 w_t 前后考虑的词的个数。许多研究证明，在识别词语之间的语义关系方面，Skip-Gram 模型有着更好的效果，因此，本文使用 Google 的开源工具——Python 版 Word2Vec 并且采用 Skip-Gram 模型进行词向量化训练。Word2Vec 中 Skip-Gram 模型的目标函数为：

$$G = \sum \log p(Context(w_t) \mid w_t) \tag{2.2}$$

$$p\left(Context(w_t)|w_t\right) = \prod_{u \in Context(w_t)} p(u|w_t) \tag{2.3}$$

其中，Context(w_t)表示与词 w_t 的距离小于 R 的上下文，R一般取 5 到 10。Skip-Gram 模型用哈夫曼树(Huffman Tree)表示输出层的结果，一个词对应树上的一个叶子节点，每一个非叶子节点表示选择该节点左右子节点的概率值。从树的根节点出发，存在可访问到任意词 w_t 的路径。此外，Word2Vec 采用了 Hierarchial Softmax 和 Negative Sampling 两种方法对条件概率函数进行了巧妙的构造，在得到目标函数后，采用随机梯度下降求解模型的最优参数。

2.3 条件随机场综述

条件随机场(Conditional Random Fields, CRF)最早由 Lafferty[53] 等人于 2001 年提出，可以看成是一种无向图模型或马尔科夫随机场，可用于标记和切分结构化数据的统计框架模型，包括网格型结构、树型结构和序列型结构。

它既能够进行序列概率的全局归一化，又能够自由设定序列的特征函数标记序列，避免了对输出序列做条件独立性假设，很好地解决了标注偏置问题，拟和现实数据，因此被广泛使用于如分词、词性标记、命名实体识别等自然语言处理的相关应用中。

条件随机场模型是在给定随机变量 X 的条件下，随机输出变量 Y，目标是构建条件概率模型 P(Y|X)，满足马尔科夫性：

$$P(Y_v \mid X, Y_w, w \neq v) = P(Y_v \mid X, Y_w, w \sim v) \tag{2.4}$$

式中 $w \sim v$ 表示无向图 $G = (V, E)$ 中与结点 v 有边连接的所有结点 w，$w \neq v$ 表示结点 v 以外的所有结点，Y_v，Y_u 与 Y_w 为结点 v，u 与 w 对应的随机变量。条件随机场使用势函数和图结构上的团来定义条件概率 P(y|x)。链式条件随机场主要包含两种关于标记变量的团，即单个标记变量化$\{y_i\}$及相邻的标记变量$\{y_{i-1}, y_i\}$，可按照以下的参数化进行计算：

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} w_k f_k(y_{i-1}, y_i, x, i)\right) \tag{2.5}$$

其中，

$$Z(x) = \sum_{y} \exp\left(\sum_{i,k} w_k f_k(y_{i-1}, y_i, x, i)\right) \tag{2.6}$$

在命名实体识别任务中，最常用的是线性链条件随机场(linear chain CRF)。给定观测序列 X=$\{X_1, X_2, X_3 \ldots X_T\}$和与之相对应的标记序列 Y=$\{Y_1, Y_2, Y_3 \ldots Y_T\}$，Y 的条件概率分布 P(Y|X)构成条件随机场，即：

$$P\left(Y_i \mid X, Y_1, Y_2, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_T\right) = P\left(Y_i \mid X, Y_{i-1}, Y_{i+1}\right), i = 1, 2, \ldots, T \tag{2.7}$$

其结构如图 2-4 所示:

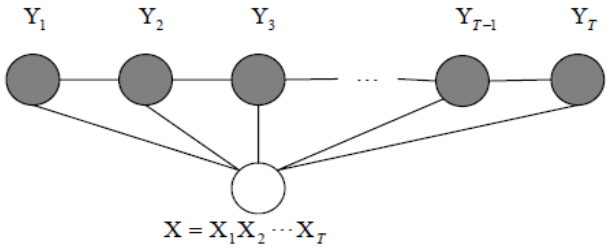

图 2-4　线性链条件随机场

对于线性链条件随机场，随机变量 Y 取值为 y 的条件概率有如下形式：

$$P(Y \mid X) = \frac{1}{Z(X)} \exp\left(\sum_{i=1}^{T} \sum_{k=1}^{K} w_k f_k(t, Y_t, Y_{t-1}, X)\right) \tag{2.8}$$

其中，

$$Z(X)=\sum_{Y}\exp(\sum_{i=1}^{T}\sum_{k=1}^{K}w_k f_k(t,Y_t,Y_{t-1},X)) \tag{2.9}$$

公式(2.8)中，$f_k(t,Y_t,Y_{t-1},X)$ 表示了当给定输入序列中的位置 t 和输入 X，当前位置的标记 Y_t 和前一个位置的标记 Y_{t-1} 时的第 k 个特征值，w_k 为特征权重，Z(X)为归一化因子，求和是在所有可能的输出序列上进行的。条件随机场模型利用前向-后向算法进行不同序列位置的条件概率和特征期望，使用拟牛顿法等极大化似然估计求解模型参数，利用 Viterbi 算法进行动态规划解码测试序列数据。

条件随机场自被提出以来，便被序列标注问题相关研究者进行了广泛的研究和应用，也有诸多相关的开源 CRF 工具包。CRF 一度是最出色的实体识别模型，在中文实体识别的效果上有所提升，但一直未有突破。在原理上，条件随机场仍存在着一些缺陷：它假设了当前的状态只与前面的几个状态或者词有关，与更前面的状态和词相互独立，因此不能抽象出语言中长距离特征。其次，只考虑过去的状态对当前状态的影响，而没有考虑到后面发生的状态也可能对当前状态产生影响。此外，它的实体识别是以词或字单位进行的。分词的准确率会很大地影响命名实体的效果。

由于面向中医典籍的命名实体识别的研究相对匮乏，且没有公开的大规模权威中医语料库，因此本文将利用 CRF 得到的模型和训练结果作为实验的基线。在本文中，X 是输入的已经标注好的中医典籍实体。Y 是与每个字对应的语义信息，语义信息由三种状态，B 为是中医术语实体的开始，I 为是中医术语实体的中间部分，O 为非语义信息。学习时，利用训练数据集通过极大似然估计或正则化的极大似然估计得到条件概率模型 P(Y|X)；预测时，对于给定的输入序列 x，求出条件概率 P(y|x)最大的输出序列。

2.4 深度学习的神经网络模型

深度学习，也叫深度神经网络，其网络结构包含大量相连接的神经元，主要是相对支持向量机、条件随机场、最大熵等浅层方法而言的，浅层的机器学习算法依靠人工经验去抽取样本特征，这些特征往往是单层特征，没有层结构；而深度学习算法通过对原始的特征进行变换，可自动学习得到层次化的组合特征，更有利于进行分类和特征提取。

深度神经网络主要分为三类：反馈神经网络、前馈神经网络以及双向神经网络。前馈神经网络是由多个编码器层叠加而成，常见的有多层感知机和卷积神经网络等。反馈神经网络是由多个解码器层叠加而成，常见的有反卷

积网络和层次稀疏编码网络等。双向神经网络是将多个编码器层和解码器层叠加得到，常见的有深度玻尔兹曼机和深度信念网络等。目前，深度神经网络最主流的两个架构为卷积神经网络(Convolutional Neural Network, CNN)和循环神经网络(Recurrent Neural Network, RNN)。而闸门机制的进步缓解了基础 RNN 的一些限制，最终形成两种主流的 RNN 类型：长短期记忆单元(Long Short-Term Memory, LSTM)和循环门单元(Gated Recurrent Unit, GRU)。

2.4.1 卷积神经网络

卷积神经网络(Convolutional Neural Network, CNN)是目前应用最为广泛的一种深度学习结构，在数字图像处理领域取得了巨大的成功，从而掀起了深度学习在自然语言处理领域的狂潮。

1) 经典的 CNN 模型原理

卷积神经网络是一种深度前馈人工神经网络，在图像分类领域做出了巨大贡献，是当今绝大多数计算机视觉系统的核心技术。Krichevsky 等人于 2012 年将 CNN 用于图像识别并取得惊人的效果；LeCun 等人利用卷积神经网络成功解决了手写体数字识别问题。Collobert[54] 等人于 2011 年将卷积神经网络引入到了自然语言处理中的许多任务中，获得了很好的表现; Shen[55] 等人利用卷积神经网络解决信息检索中的语义分析问题; Kalchbrenner[56] 等人利用卷积神经网络对句子进行建模，并且提出一种新的池化(pooling)方式。2014 年，Kim[57] 利用卷积神经网络进行句子分类，在各个数据集上取得非常好的效果，其 CNN 模型结构如下图 2-5 所示。

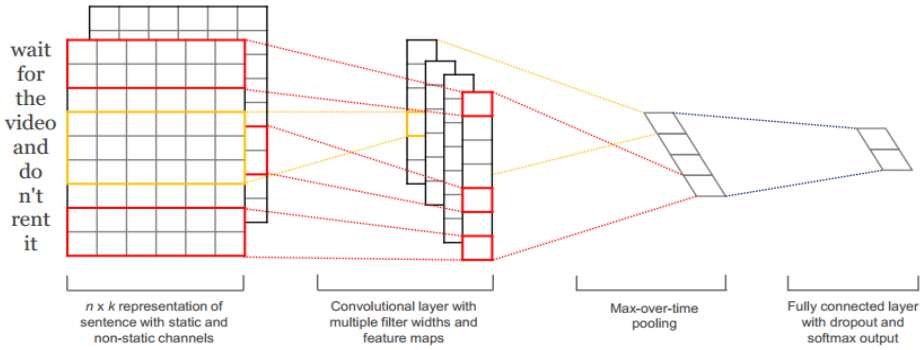

图 2-5 Kim.Y 的 CNN 模型结构图

卷积神经网络通常包括以下 3 个层次：卷积层、池化层和输出层。卷积层(Convolutional layer)主要作用是特征提取，每层卷积层由若干卷积单元组成，每个卷积单元的参数都是通过反向传播算法优化得到的。第一层卷积层可能只能提取一些低级的特征如边缘、线条和角等层级，更多层的网络能从

低级特征中迭代提取更复杂的特征。前一层的局部区域连接到神经元，神经元提取出该区域的特征，最后加上激活函数，常用的激活函数包括 ReLU 和 Tanh 等。ReLU 又称为线性整流，具体函数为：

$$f(x) = \max(0, x) \tag{2.10}$$

池化层(Pooling layer)，又称采样层，主要作用是特征映射，映射函数一般有 max 函数和 average 函数，即常用的方式有最大池化(Max Pooling)和均值池化(Mean Pooling)。通常在卷积层之后会得到维度很大的特征，池化层将特征切成几个区域，取其最大值或平均值，得到新的、维度较小的特征。输出层一般为全连接层，其后按任务不同加上一个激活函数层。图 2-5 的输出层结构为：全连接层(Fully-Connected layer)用全连接的方式把所有局部特征结合变成全局特征，计算最后每一类的得分作为分类器的输入，且使用 dropout 技术防止隐藏层单元自适应，从而减轻过拟合。分类器一般选用激活函数，二分类选用 sigmoid 函数,多分类选用 softmax 函数。

卷积神经网络这种特殊的网络结构使得它具有如下显著的优点:

(1) 卷积层及池化层的交替叠加使得 CNN 对于局部微小特征非常敏感;

(2) 特征提取和模式分类同时进行，并同时在训练中产生; 池化操作在提炼突出特征的同时压缩了数据的维度;

(3) 利用局部感受野和权值共享减少网络的训练参数、降低了模型的复杂度，使得其适应性更强。

一般在自然语言处理任务中，采用一维卷积核，但卷积核大小其实也是二维的，只不过可变动的只有步长这一维，另一维和词向量宽度一样一般设为不可变。池化层会对数据进行降维压缩，信息会丢失，为了不让有效的信息丢失太多，一般的做法是增加卷积核的数量。卷积网络非常利于 GPU 的并行化加速，因为同一特征映射面上，神经元共享权值。卷积网络相对来说速度快，可以处理很长的文本，非常适合提取空间结构特征。

2) 改进的 CNN 模型（IDCNN）

对于序列标注任务来说，CNN 在计算性能上有着绝对的优势，因为它的计算成本不随着输入大小而增加，而是随着层数的增加而增加，直到达到硬件的内存和线程的限制。但是它的劣势在于，卷积之后末层神经元可能只是得到了原始输入数据中一小块的信息。为了覆盖到输入的全部信息就需要加入更多的卷积层以及防止过拟合函数，带来更多的超参数，整个模型变得庞大和难以训练。近几年来，研究人员不断提出对传统 CNN 的改进模型。

2016 年，Yu 和 Koltun[58] 对卷积层进行了改进，提出了“空洞”卷积

(Dilated Convolutions)方法，解决了池化过程导致很多信息损失的问题，在图像分割领域取得了很好的效果。它直接去掉池化下采样操作，而不降低网络的感受野。具体使用时，dilated 宽度会随层数的增加而指数增加，即各层的参数数量相同，随着层数的增加，参数数量线性增加，而感受野(receptive field)指数增加，可以很快覆盖到全部的输入数据。F_{i+1} 中各元素的感受野大小为：

$$F_{i+1} = (2^{i+2} - 1) \times (2^{i+2} - 1) \tag{2.11}$$

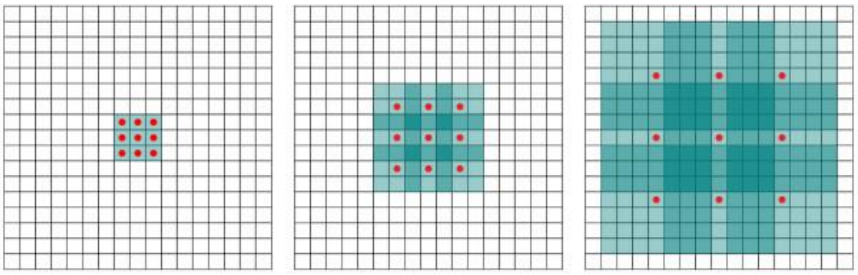

图 2-6 图像领域的 “空洞”卷积原理图

上图 2-6 中，图(a)采用 1-dilated convolution，对 F_0 操作得到的 F_1，F_1 中各元素的感受野为 3×3。图(b)采用 2-dilated convolution，对 F_1 操作得到的 F_2，F_2 中各元素的接受野为 7×7。图(c)采用 4-dilated convolution，对 F_2 操作得到的 F_3，F_3 中各元素的接受野为 15×15。

随后，Kalchbrenner[59] 等人将该方法运用到了机器翻译任务上。该模型中，句子建模时输入是以句子的字符级别开始的，堆叠了 2 层之后随着卷积核所能覆盖的范围扩展，不断地去交互信息，同时还能够保证原始的输入信息不被丢失。堆叠式的“空洞”卷积神经网络可以轻松整合来自整个句子或文档的全局信息。但是，简单地增加堆叠“空洞”卷积的深度会导致相当大的过拟合。因此，Strubell[60] 等人于 2017 年提出了迭代“空洞”卷积神经网络(Iterated Dilated CNNs, IDCNN)。它不再多次都使用相同的小块“空洞”卷积，而是每次迭代将最后一个应用的结果作为输入，以循环的方式重复使用相同的参数，从而提供了广泛的有效输入宽度和理想的泛化能力。

IDCNN 的输入是一个长度为 T 个向量的序列，输出一个每个类的分数序列。将 dilated 宽度为 δ 的第 j 个 dilated 卷积层表示为 $D_{\delta}^{(j)}$，则第一层是一个 dilated-1 卷积为 $D_1^{(0)}$，它将输入转换为一个表示 i_t：

$$i_t = D_1^{(0)} x_t \tag{2.12}$$

接下来，将指数增大的扩张宽度的卷积 L_c 层应用到 i_t 中，将越来越宽的上下文折叠到每层的嵌入表示 X_t 中，令 $r()$表示 ReLU 激活函数：

$$c_t^{(j)} = r(D_{2^{L_c-1}}^{(j-1)})c_t^{(j-1)}) \tag{2.13}$$

并且将一个最后的 dilation-l 层添加到堆叠中：

$$c_t^{(L_c+1)} = r(D_1^{(L_c)}c_t^{(L_c)}) \tag{2.14}$$

将这个“空洞”卷积的堆叠定义为块 B(•)，其输出分辨率等于其输入分辨率。为了在不过度拟合的情况下融入更多的上下文信息，不引入额外的参数，反复使用块 B L_b 次。从 $b_t^{(1)} = B(i_t)$ 开始：

$$b_t^{(k)} = B(b_t^{(k-1)}) \tag{2.15}$$

我们应用一个简单的仿射变换 W_o 到最终的表示来为 x_t 中的每个 token 获得每个类的分数：

$$h_t^{(L_b)} = W_o b_t^{(L_b)} \tag{2.16}$$

2.4.2 循环神经网络

循环神经网络(Recurrent Neural Network, RNN)在前馈神经网络的隐含层加入了一个环路，使得计算当前隐含层能够同时接收到当前的输入和上一个隐含层的输出两个方向的信息，从而具有了记忆功能，解决了传统神经网络无法根据先前事件预测下一事件的难题。因此，RNN 被广泛应用于处理和预测序列数据，在过去几年中，在语音识别，语言建模，翻译，图片描述等问题上取得显著的效果。典型的 RNN 展开结构如图 2-7 所示。

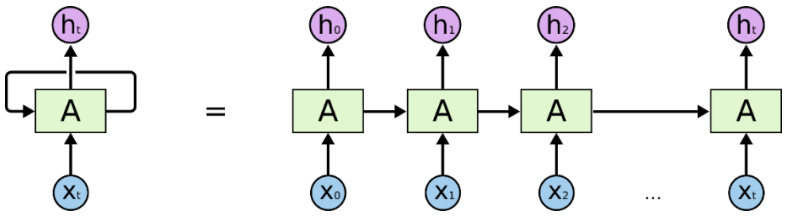

图 2-7 循环神经网络结构

X_t 为当前时刻的输入，A 为模型处理部分，h_t 为当前时刻的输出，其计算公式如下所示，W 和 b 为相应的权重矩阵。

$$h_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2.17}$$

理论上 RNN 能学习无限长的序列，然而 Mikolov[61] 等人经过实践发现是影响 RNN 存在着“梯度消失”的问题，距离当前节点位置较远的节点信息不断被稀疏，使得距离当前位置较远的节点信息在实际应用中并不能被利用到。随着神经网络层数的增加，“梯度消失”情况会越来越严重，反向传播一层，梯度衰减为上一层的 0.25，层数多了后，底层神经元基本接收不到信号。

1) 长短期存储单元（LSTM）

1997 年，Hochreiter[62] 等人提出了控制信息更为精细的长短期存储单元(Long Short-Term Memory, LSTM)。它是 RNN 的特殊类型，其记忆单元(Memory Cell)与输入门(input gate)、输出门(output gate)、遗忘门(forget gate)相连接，进而控制和更新各个门单元的相关参数进行模型的学习和训练，调整信息衰减、更新、去留的程度，使得存储单元能够获得距离较远的历史信息，有效解决了“长期依赖”问题。LSTM 单元的结构如图 2-8 所示：

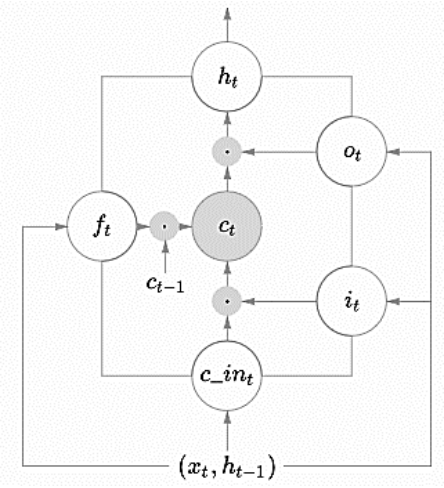

图 2-8 LSTM 单元结构图

其中，中心的 c_t 即为记忆单元 cell，从下方输入(x_t ,h_{t-1})到输出 h_t 的一条线即为细胞状态，f_t，i_t，o_t 分别为遗忘门、输入门、输出门。LSTM 通过门控单元可以去除或添加信息到细胞状态的能力，可以有选择地决定信息是否通过，它由一个 Sigmoid 神经网络层和一个成对乘法操作组成。该层的输出是一个介于 0 到 1 的数，表示允许信息通过的多少，0 表示完全不允许通过，1 表示允许完全通过。LSTM 的结构流程如下：

首先，决定从细胞状态中丢弃什么信息。遗忘门 f_t 判断过去记忆 c_{t-1} 的重要程度，进而判断让过去的记忆内容多大的程度参与新记忆的生成。

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2.18}$$

下一步，确定什么样的新信息被存放进细胞状态中。输入门 i_t 通过 Sigmoid 来判断当前的单词的重要程度，进而判断让它以何种程度参与生成新的记忆。同时，用一个 tanh 层用来生成新的候补记忆单元 c_in_t，把这两部分产生的值结合来进行更新。

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2.19}$$

$$c\_in_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2.20}$$

接下来，进行细胞状态的更新，得到当前时刻记忆单元 c_t。

$$c_t = f_t * c_{t-1} + i_t * c\_in_t \tag{2.21}$$

最后一步，决定模型的输出。首先通过输出门 o_t 来确定细胞状态的哪个部分将输出出去。接着把当前时刻细胞状态通过 tanh 进行处理，并将两者综合考虑，仅仅输出确定的那部分，得到隐藏层最终输出 h_t。

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{2.22}$$

$$h_t = o_t * \tanh(c_t) \tag{2.23}$$

Sigmoid 函数的输出是不考虑先前时刻学到的信息的输出，tanh 函数是对先前学到信息的压缩处理，起到稳定数值的作用，两者的结合学习就是递归神经网络的学习思想。

2) 双向长短期存储单元（BiLSTM）

传统的 LSTM 单元一般是前向的，即只考虑了过去序列对当前的影响，而无法利用后文信息进行知识学习，导致对模型的效果造成负面影响。而在自然语言处理的研究任务上，无论是字、词、短语，它们所处的上下文语境都对序列标注问题有着很大帮助。因此，引入双向长短期存储单元(Bidirectional LSTM, BiLSTM)模型，它能够联结了上文和下文两个方向的 LSTM 单元在同一时刻的输出并给出最终包含上下文信息的隐含层输出，进而提升整体模型的性能。图 2-9 是一个典型的双向 LSTM 的结构示意图。

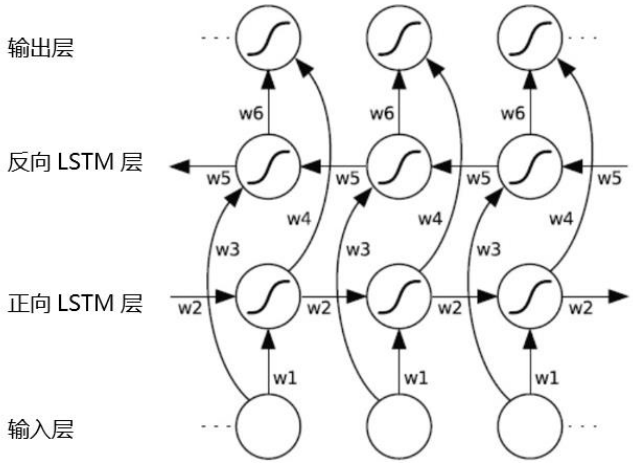

图 2-9 双向 LSTM 单元结构图

3) LSTM 的变体（GRU）

LSTM 有许多变种，改变较大有 2014 年 Cho 等人提出的循环门单元(Gated Recurrent Unit, GRU)。GRU 把 LSTM 中的遗忘门和输入门用更新门来替代，并且把细胞状态和隐藏状态 h_t 进行合并。GRU 模型比标准的 LSTM

模型简单，是非常流行的变体，其单元内部结构具体如下图所示，其中，z_t 是更新门，r_t 是重置门，决定上一时刻隐藏层输出 h_{t-1} 对当前时刻隐藏层输出 h_t 的影响程度。

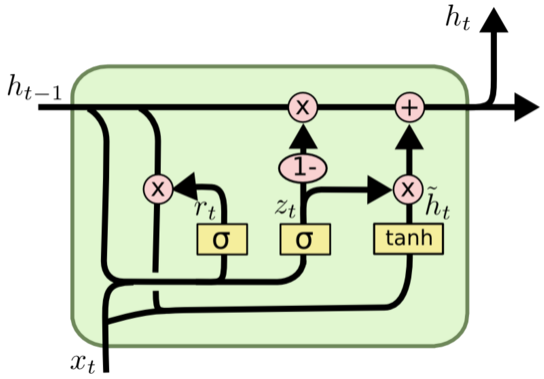

图 2-10　GRU 模型内部结构图

GRU 在计算当前时刻新信息的方法，和 LSTM 也有所区别：

$$z_t = \sigma(W_z[h_{t-1}, x_t]) \tag{2.24}$$

$$r_t = \sigma(W_r[h_{t-1}, x_t]) \tag{2.25}$$

$$\tilde{h}_t = \tanh(W \cdot [r_t \times h_{t-1}, x_t]) \tag{2.26}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \tilde{h}_t \tag{2.27}$$

2.4.3 注意力机制

注意力机制(Attention)[63] 最早是在视觉图像领域提出来的，模仿人看图像时，人的注意力总是集中在画面中的某个焦点部分，而对其它部分模糊处理，然后不断地调整目光的焦点。而人们阅读文本时，其实也是如此。大量实验证明，将 Attention 机制应用在机器翻译，摘要生成，文本理解等自然语言处理的问题上，也取得了显著的成效。Attention 机制引入了权重的概念，其核心就是一个编解码的过程。

1) 编码-解码模型(Encoder-Decoder)

编码，就是将输入序列转化成一个固定长度的向量；解码，就是将之前生成的固定向量再转化成输出序列。Encoder 和 Decoder 均可进行自由组合，常见的有 CNN、RNN、BiRNN、GRU、LSTM 等。传统的 Encoder-Decoder 模型在处理机器翻译任务时，需要根据一个句子 X=$\{x_1, x_2, ..., x_m\}$，得到另一个句子 Y=$\{y_1, y_2, ..., y_n\}$，其经典框架如图 2-11 所示。

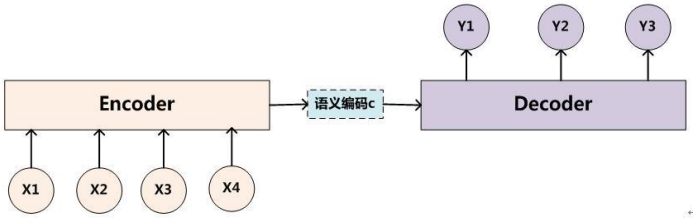

图 2-11 Encoder-Decoder 模型框架

先对输入句子X进行语义编码得到中间语义C，而F为非线性编码函数：

$$C = F\left(x_1, x_2, ..., x_m\right) \tag{2.28}$$

之后，再利用中间语义 C 和之前已输出的序列 $y_1, y_2, ..., y_{n-1}$ 来生成 i 时刻要生成的单词序列 y_n，其中 G 为解码函数：

$$y_i = G\left(C, y_1, y_2, ..., y_{n-1}\right) \tag{2.29}$$

传统编解码模型框架的缺点在于编解码之间的唯一联系是一个固定长度的语义向量 C。那么输入文本的每一个单词 x_m 对输出文本的单词 y_n 的贡献都一样，这显然不合理。无论之前的句子有多长，最终都要被压缩成一个固定长度的向量中去，这使语义向量无法完全表示整个序列的信息，且句子的长度越长，最终生成的向量中损失的信息越多，导致解码准确度下降。

2) 注意力模型(Attention)

注意力模型为解决传统编解码框架的问题，不再要求编码器将所有输入信息都放进一个固定长度的向量之中，而是在每次输出的 y_n 之前，根据之前输出的 $y_1, y_2, ..., y_{n-1}$ 来确定哪个输入的 x 应该获得更多的关注，即给每个输入 x 分配一个注意力权重，再根据这个结果更新中间语义 C_n，并利用这个中心语义和之前的 y 来得到当前应该输出的 y_n。

$$y_1 = f_1\left(C_1\right) \tag{2.30}$$

$$y_2 = f_1\left(C_2, y_1\right) \tag{2.31}$$

$$y_3 = f_1\left(C_3, y_1, y_2\right) \tag{2.32}$$

注意力模型如下图 2-12 所示：

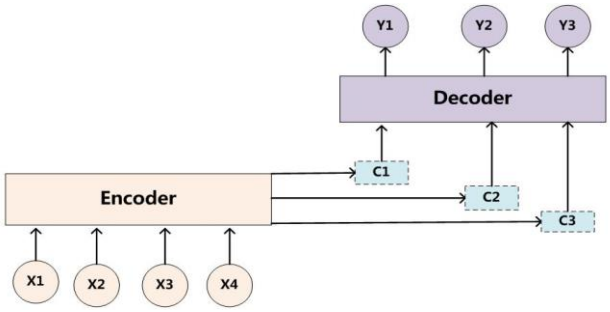

图 2-12 Attention 模型框架

一般编解码器默认选择为 RNN，那么上述公式实际会变化为下式，其中 s 为输出神经元的隐状态：

$$p(y_j \mid y_1, y_2, ..., y_{j-1}, X) = f_{12}(C_j, s_j, y_{j-1}) \tag{2.33}$$

$$s_j = f_{11}(C_j, s_{j-1}, y_{j-1}) \tag{2.34}$$

$$C_j = \sum_{i=1}^{T_x} a_{ij} f_2(x_i) \tag{2.35}$$

其中 T_x 是输入 x 的数量，在 RNN 网络中 $f_2(x_i)$ 往往是输入 x_i 后的隐状态 h_i，所以在 RNN 中有：

$$C_j = \sum_{i=1}^{T_x} a_{ij} h_i \tag{2.36}$$

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{i=1}^{T_x} \exp(e_{kj})} \tag{2.37}$$

那么关键就是 e_{kj} 的计算，方法有很多种，综合来看和上一步输出神经元的隐状态 s_{j-1}，第 i 步输入神经元的隐状态 h_i 两者的相关性有关。那么就利用 score 函数来计算 s_{j-1}，h_i 的关系分数，如果分数大则说明关注度较高，注意力分布就会更加集中在这个输入单词上。南京大学的张冲[64] 给出了一个比较有代表性的公式：

$$e_{kj} = score(s_{j-1}, h_i) = v \tanh(Wh_i + Us_{j-1} + b) \tag{2.38}$$

其中，v、W、U 均是要训练的权重矩阵。

3 中医典籍先验知识快速获取方法

目前，命名实体识别、关系抽取在公开领域取得了很好的效果，但是基本都是对人名、地名、机构名这三类实体的识别。因为这三类实体均已有较为完备的词表，并且不太需要具备领域知识，具有较强的延展性，使得中文分词效果好，效果能接近人工标注的水平。由此可见，获取领域词表，进行特征提取是知识图谱构建过程中的必要部分。中医领域已有的词表局限于方剂名、药材名、病症名和经脉穴位，并不适用于如《黄帝内经》这样的中医理论典籍，且构建中医领域词表需要较强的领域知识，耗费人力，工作量大，效果不佳。本文提出了一种快速获取中医典籍先验知识、提取特征的方法。

3.1 中医典籍的语言特点

中医典籍是中医药学的精髓所在，其体系复杂，种类繁多。根据《全国中医图书联合目录》统计中医典籍多达 12124 种，主要分为医经（如《黄帝内经》、《难经》、《神农本草经》）、方药（如《金匮要略》、《本草纲目》、《备急千金要方》）、各科临床经验（如《伤寒论》、《诸病源候论》、《妇人大全良方》、《外科证治全生集》）、医案（如《名医类案》）等等。经过阅读中医典籍总结发现，其语言表述独具特色，获取知识存在很高的难度。下面以我国最经典而久远的中医典籍——《黄帝内经》为例，进行分析。

词汇方面，中医典籍中的术语多采用嵌套结构，如使用大量双字格、三字格和四字格术语。而四字格术语含义极其丰富，内部成分之间的语法关系复杂，又可分为述宾词组、定中词组、状中词组、主谓词组和联合词组，很难区分确切的中医术语。

表 3-1　中医典籍的词汇特征

词汇特征	具体含义	样本实例
双字格术语	由两个单字组成的双字结构	肝厥、血气、阴虚
三字格术语	由一个单音节字和一个双音节词组成	肝藏魄、髓之府
述宾词组	由述语和宾语组成，成分间是支配与被支配关系	调和气血、清热化痰
定中词组	由名词作定语修饰名词中心语	血分热毒、膀胱湿热
状中词组	由状语和述宾词组组成的结构	芳香化浊、渗淡利湿
主谓词组	由主语和谓语组成的四字词组	心肾相交、痰蒙心包
联合词组	由语法地位平等的多个部分组成，如并列和递进	津枯血燥、散寒祛湿

句式方面，中医典籍受中华文化和哲学思想影响，句式结构复杂，大部分为复合长句，结构松散，有非常多的修饰成分和内容层次。如“*余闻上古之人，春秋皆度百岁，而动作不衰；今时之人，年半百而动作皆衰，时世异耶？人将失之耶？*[①]”句子常由一连串并列谓语构成。因为中医典籍在阐述医理时，注重强调客观事实和事物，且常采用“取象比类”的方式，譬如取五行比五脏、取阴阳比生死等，利用其它关联事物的性质特征对人体，乃至万物中蕴含的中医知识进行解释分析。

此外，同一词汇存在着多种含义，例如“主”字，在“心主脉”中意思为“控制”，在“心之合脉也，其主肾也”中意思为“治疗”。总的来说，中医典籍的语言存在以下特点和难点，导致分词困难，难以快速获取领域词表。

（1）存在着大量的生僻词、通假字，如“五脏”和“五藏”、“六腑”和“六府”、“四肢”和“四支”、“月真”等；

（2）词语歧义、一词多义、多词一义情况严重，如动词“主”共有 5 个不同的含义，动词“出”有 16 个含义，名词“水”也会有多个不同含义；

（3）句法上采用了大量修辞手法，多用类比推理，阅读起来深奥晦涩；

（4）词汇方面存在大量嵌套形式，且内部的语法关系相当复杂；

（5）中文的书写不似英文那般有空格作为词语的分隔符，因此中文的词边界常常模糊不清，很多中医理论典籍甚至连标点符号都没有；

（6）中医知识源于经验积累，每个人对中医内容的总结阐述有自己的习惯，因此没有统一的标准。

3.2 人工构建中医典籍领域词表

下面依旧以《黄帝内经》为例，详细介绍《黄帝内经》的语言特点和可较快速人工构建领域词表的方法。

《黄帝内经》是我国的医药之祖，奠定了我国古代的医学发展基础，有《素问》和《灵枢》两个部分，各 81 篇，共约 23 万字。其中的中医思想发源于远古时期，成熟在春秋时期，数千年来始终是中医的研究重点，地位崇高。《黄帝内经》较全面地阐述了中医学的核心思想和理论体系，涵盖的知识广博，思想深邃，其中还融入许多道家思想，是医与道的结合。在语言表述方面，其语言精美，论述精炼，富有浪漫主义色彩，示例如下：

① 《黄帝内经》上古天真论 https://so.gushiwen.org/guwen/bookv_964.aspx

“黄帝[②]曰：阴阳者，天地之道也，万物之纲纪，变化之父母，生杀之本始，神明之府也。治病必求于本。故积阳为天，积阴为地。阴静阳躁，阳生阴长，阳杀阴藏。阳化气，阴成形。寒极生热，热极生寒；寒气生浊，热气生清；清气在下，则生飧泄，浊气在上，则生(月真)胀。此阴阳反作，病之逆从也。”

——《素问·阴阳应象大论》

《黄帝内经》中采用了多种修辞手法，除了现在也使用较多的比喻、比拟、借代、对偶等修辞方法，还用到了联珠、辟复、互文、讳饰等比较生僻的修辞手法。具体的修辞手法和示例如下表所示：

表 3-2　中医典籍的修辞手法

修辞手法	示例句子	所属章节
明喻	目裹微肿，如卧蚕起之状，曰水。	《素问·平人气象论》
暗喻	太阳为开，阳明为阖，少阳为枢。	《素问·阴阳离合论》
借喻	开鬼门，洁净府，五阳已布，疏涤五脏。	《素问·汤液醪醴论》
拟人	肝恶风，心恶热，肺恶寒，肾恶燥，脾恶湿。	《灵枢·九针论》
对偶	清气在下，则生飧泄；浊气在上，则生嗔胀。	《素问·阴阳应象大论》
联珠	东方生风，风生木，木生酸，酸生肝，肝生筋。	《素问·五运行大论篇》

1) 动词及固定句式

虽然《黄帝内经》语法、句法复杂，但从表 3-2 中可以看出，其章节名包含着中医理念，因此首先将各篇章节名处理归纳为实体，获得 162 个中医术语。此外，丰富的修辞手法和大量的动词、固定句式也为中医术语的获取提供了模板。如比喻中，常以“如”、“若”、“犹”、“譬”、“似”、“象”、“为”作为喻词；拟人中，代表厌恶的“恶”、代表运动趋势的“走”、代表欺侮的“侮”；在联珠中的“生”等动词，出现得十分规律，使得典籍充满韵律感。

例如：*“东方生风，风生木，木生酸，酸生肝，肝主目。在天为风，在地为木，在体为筋，在脏为肝，在色为苍，在音为角，在声为呼，在窍为目，在味为酸，在志为怒。怒伤肝，悲胜怒，风伤筋，燥胜风，酸伤筋，辛胜酸。*

南方生热，热生火，火生苦，苦生心，心主舌。其在天为热，在地为火，在体为脉，在脏为心，在色为赤，在音为徵，在声为笑，在窍为舌，在味为苦，在志为喜。喜伤心，恐胜喜，热伤气，寒胜热，苦伤气，咸胜苦。”

——《素问·阴阳应象大论》

② 《黄帝内经》阴阳应象大论 https://so.gushiwen.org/guwen/bookv_968.aspx

根据文本中存在的相似结构的句式，我们可以在领域知识不足的情况下，根据固定句式定位动词，如“生”、“主”、“伤”、“胜”、“在天为”、“在地为”等，将动词两端的名词 “东方”、“风”、“木”、“南方”、“热”、“火”等作为中医术语，并且根据句子的映照性推测，“东方”与“南方”、“风”与“热”、“木”与“火”为同类实体。其中“在天为”、“在地为”等为嵌套三字格，进一步拆分可知，“天”也为实体，且与“玄”存在语义关系，由此规律整理整理，快速获取了 1005 个名词实体词表和 105 个动词形成的动词词表。词表的部分结果如下表所示：

表 3-3 固定句式部分实体词表

概念	实体 1	实体 2	实体 3	实体 4	实体 5	实体 6
五脏	肝	心	脾	肺	肾	-
五方	东方	南方	中央	西方	北方	-
五色	苍	赤	黄	白	黑	-
五味	酸	苦	甘	辛	咸	-
五行	木	火	土	金	水	-
六腑	胆	小肠	胃	大肠	膀胱	三焦
六气	风	热	湿	燥	寒	火
三阴三阳	太阳	少阳	阳明	厥阴	少阴	太阴
十二皮部	关枢	枢持	害蜚	枢儒	关蛰	害肩
穴位	天突	人迎	扶突	天窗	天容	天牖
病名（痿）	痿厥	痿躄	脉痿	筋痿	肉痿	骨痿
病名（痹）	肺痹	肾痹	肝痹	骨痹	行痹	食痹
病症	皮槁	脉凝泣	爪枯	筋急	头痛巅疾	下虚上实
治法	砭石	按蹻	毒药	灸焫	微针	汤液
部位	胸中	膈中	女子胞	脑	寸口	心中

从结果可以看出，主要可以快速总结出的包括阴阳、五行，包含脏腑、经脉、穴位、部位的人体生理部分以及病名、病症等等。而由《黄帝内经》中固定句式得到的动词词表如下：

一字动词 24 个：主、伤、食、当、刺、荣、走、胜、恶、出、应、治、藏、归、入、宜、候、禁、合、为、则、生、欲、曰；

二字动词 48 个：络于、生于、通于、发于、出于、客于、结于、注于、伤于、在于、属于、并于、入于、藏于、根于、溜于、因于、病在、俞在、过在、厥在、所谓、是谓、此谓、此为、谓之、发为、名曰、为上、成为、

当病、病名、出焉、则梦、欲如、命曰、名为、其色、其音、其虫、其令、其变、其味、其类、其畜、其谷、其臭、其数；

三字动词 32 个：入通于、藏精于、传之于、合入于、内会于、开窍于、受气于、入舍于、禀气于、移热于、移寒于、治之以、病名曰、其华在、在色为、在音为、在志为、在脏为、在声为、在体为、在气为、其性为、其用为、其化为、在天为、在窍为、在地为、其充在、其政为、其志为、其德为、其眚为、其色为；

四字动词 1 个：在变动为。

2) 特殊标点及数字概念

此外，《黄帝内经》中还存在着一部分形式不同的结构相似的句子。例如：

“五味所入③：酸入肝、辛入肺、苦入心、咸入肾、甘入脾，是为五入。

五脏化液：心为汗、肺为涕、肝为泪、脾为涎、肾为唾。是为五液。

五脏所藏：心藏神、肺藏魄、肝藏魂、脾藏意、肾藏志。谓五脏所藏。

五脉应象：肝脉弦、心脉钩、脾脉代、肺脉毛、肾脉石。谓五脏之脉。

五劳所伤：久视伤血、久卧伤气、久坐伤肉、久立伤骨、久行伤筋。是谓五劳所伤。”

——《素问·宣明五气》

利用《内经》中存在的特殊标点符号，譬如“：”“、”“——”等，总结出冒号前后的实体为“是”的解释关系，由此归纳了一部分具有层次概念的实体。并且《内经》中存在着大量的数字，如上述例子中出现的“五味”、“五脏”、“五脉”、“五劳”，还有“三焦”、“六腑”、“七星”、“八风”、“九针”等，这些都属于总结性的高层次概念，其下拥有对应的中医术语。因此我们根据数字，对内经中的实体进行归纳，扩充了 285 个实体词表。

表 3-4 数字规律总结的部分概念词表

类别	内容
量词	一升、二剂、三丈、四寸、七尺、九斗、三舍、三度、五里、六刻、三周
时间	三日、一夜、一备、一纪、一月、十二时、十二辰、一岁、三年、六期
动名词	一盛、一夺、一逆、一合、三变、三常、三实、四过、五禁、五运、五决
数字“一”	一阳、一阴、一脏、一候、一节、一痈、一经
数字“二”	二穴、二火、二输、二十五俞、二十八宿、二十八脉、二十八会、二八星
数字“三”	三阳、三阴、三品、三水、三焦、三椎、三针、三部、三气之纪、三百六十五节气、三百六十五节、三百六十五穴、三百六十五络

③ 《黄帝内经》宣明五气 https://so.gushiwen.org/guwen/bookv_986.aspx

表 3-4　数字规律总结的部分概念词表（续）

数字“四”	四时、四气、四季、四淫、四街、四肢、四难、四关、四海、四厥、四野
数字“五”	五行、五脏、五风、五体、五脉、五痹、五形志、五肤俞、五丸、五趾、五音、五声、五络、五位、五星、五官、五阅、五使、五输、五俞
数字“六”	六腑、六府、六气、六经、六元、六经脉、六律、六分、六痛
数字“七”	七损、七诊、七窍、七节、七焦、七椎、七星、七疝
数字“八”	八远、八风、八益、八节、八纪、八溪、八正、八俞、八痛
数字“九”	九州、九窍、九脏、九候、九气、九节、九野、九焦、九针、九谊、九宫
数字“十”	十二节、十二从、十二经络脉、十二邪、十二经脉、十二经水、十二官、十二分、十二疟、十二部、十二藏、十二原、十五络

针对《黄帝内经》的语言特色，通过章节名、特殊标点、数字以及固定句式，归纳整理快速得到了部分实体词表和动词关系词表，用于指导中医实体的分类和关系的分类。

3.3 基于层次聚类的种子实体获取

上文中按照文本特点手工构建中医领域词表的方式能得到质量较高的基础词表，但仍较为困难费时；此外，对中医术语进行分类，构建层次概念仍需要依靠较深的领域知识，难以保证正确性。因此本文提出一种基于关键词提取与层次聚类，不依赖领域知识的先验知识快速获取方法。

1) 关键词抽取

首先利用结巴分词和 TF-IDF 算法抽取中医典籍中的名词性关键词。TF-IDF 算法是将 TF(Term Frequency，词频)和 IDF(Inverse Document Frequency，反文档频率)相乘，用于反映一个词对于语料中某句话的重要性。若某个词 t 在一句话 d 中出现的频率高，则 TF 高；在其他句子中很少出现，则 IDF 高，那么这个词具有很好的类别区分能力。关键词抽取的流程如下：

(1) 输入预处理后的语料；

(2) 结巴分词和词性标注，保留符合要求、满足指定词性的词作为候选词；

(3) 将候选词添加到词频词典中，出现的次数加 1；

(4) 遍历词频词典，获取每个词的 TF 值，并除以词频词典中的次数总和，分别计算得到每个词的 TF-IDF 值；

(5) 根据每个词的 TF-IDF 值降序排列，将指定个数的关键词存入列表中；

(6) 统计关键词词频，只保留满足阈值的关键词；

(7) 输出语料的关键词和词频，从高到低排序。

实验中，语料为从网络中爬取的 701 本中医典籍，输入为合并后经过数据预处理的文本，如繁简体转换、统一标点符号、特殊字符处理等。根据古文的语言习惯，中医典籍中存在着大量无用的助词、虚词等，如“之”、“乎”、“者”、“则”、“而”、“焉”等，影响分词的准确率，因此我们构建了停用词词表，在结巴分词时将其清除。同时构建了结巴分词自定义词表，用于提高中医术语识别的效果。自定义词表来源于 3.2 章节中人工整理得到的领域基础词表，融合了从网络中爬取相关词表，如搜狗细胞词库中的方剂、穴位等词表，百度百科中的穴位、治法等词表以及中医养生网站中病症类词表。

参数设置方面，由于该方法主要用于扩充以名词为核心得领域基础词表，因此将关键词词性限定为名词。同时将关键词词频的阈值为 5，即只保留词频超过（包括）5 的关键词，共获取了 15053 个关键词，按词频降序排列。其中，频率最高的是“太阳”一词，共 10558 次，随后依次为阳明、阴阳、少阴、少阳、阳气，出现 5000 次以上，随后是五脏、甲乙、桂枝、太阴、岐伯，出现 4000 多次。

2) 层次聚类

在获取名词性关键词后，本文提出以词向量为基础，将关键词表示成 N 维向量，利用层次聚类算法对关键词进行分类，从而得到有用的层次概念，为特征提取提供依据。层次聚类算法是对给定的数据进行一层一层的分解聚合，具体可分为凝结法和分裂法。凝结是一种从底向上的策略，将每个对象都作为一个类别，然后根据对象间的相似度不断合并，直到所有对象都在同一个类或者满足终止条件；而分裂是自顶向下的策略，将所有对象都归为一个类别中，然后逐渐分裂为越来越小的类别，直到每个对象自成一个类别或是满足某个终止条件。实验中，我们采用自底向上的凝结法，具体算法如下所示：

算法 1：凝聚型层次聚类算法

输入：关键词样本点的 N 维向量矩阵，聚类类别距离的阈值 *f*

输出：聚类结果

1: *将样本集中的每一个关键词样本点都当做一个独立的类簇；*

2: ***repeat：***

3: *计算两两类簇之间的距离，找到距离最小的两个类簇 c1 和 c2；*

4: *若类簇 c1 和 c2 之间的距离小于阈值 f，则合并为一个类簇 c；*

5: *重新计算合并后的类簇 c 到其他类簇之间的距离*

6: ***util：*** *任何两个类簇之间的距离均大于阈值 f*

两两类簇之间的距离计算方法有多种，常见的有(1) 最小距离：取两个类中距离最近的两个点的距离；(2) 最大距离：取两个类中距离最远的两个点的距离；(3) 均值距离：将两个类各自的两两样本间距离的平均值；(4) 平均距离：将两个类中所有的样本两两间的距离之和，除以两个类簇中的样本数量的积得到均值。该方法中，两个类簇之间的距离采用了平均距离的计算方式，其中两个样本之间的距离采用的是余弦相似度。

数据准备上，本文仍旧采用了网络爬取的合并后经过数据预处理的 701 本中医典籍，延续使用了关键词抽取过程中使用的停用词词表和自定义词表，将语料进行结巴分词。然后利用开源的词向量训练工具 gensim word2vec 训练得到每个词的词向量，维度设置为 200 维；从中筛选得到关键词的词向量，对其进行层次聚类。在参数设置方面，为得到较少的类别，将聚类类别距离的阈值 f 设置维 0.3，最终得到 13 类，其部分结果如下表所示：

表 3-5 层次聚类的部分实验结果

类别	内容
关键词 A	岐伯、黄帝、雷公、伯高、仲景、张云、闻人、张景岳、常氏
关键词 B	阴阳、阳明、太阳、少阴、太阴、厥阴、少阳
关键词 C	五脏、六腑、经脉、脏腑、孙络、溪谷、十二经脉、人身
关键词 D	气血、荣卫、皮肤、骨髓、筋脉、皮毛、精气、肌肉、血脉、腠理
关键词 E	桂枝、甘草、柴胡、人参、半夏、麻黄、芍药、生姜、茯苓、大枣
关键词 F	缺盆、肩胛、阴股、膂上、小指、内踝、发际、枕骨、环唇、季胁
关键词 G	饮食、用力、饮酒、内伤、不节、贼风
关键词 H	少商、少宫、少羽、太征、太角、戊申、己丑、辛巳、辰星
关键词 I	寒、热、大、小、数、滑、绝、短、微、弱、沉、浮、虚
关键词 J	头痛、风寒、热病、气虚、脓血、咳喘、疮疡、筋挛、巅疾、逆气

由上表可知，层次聚类的区分度较为明显并且准确率较高，即使没有深厚的领域知识，也可以对结果进行大致的判断和分类，例如关键词 A 为人名类、关键词 B 为阴阳类、关键词 C 为脏腑经脉类、关键词 D 为人体组成部分、关键词 E 为草药方剂、关键词 F 为人体部位、关键词 G 为病因、关键词 H 为天干地支等概念、关键词 I 为脉象表征、关键词 J 为病症。这样的聚类结果可以用于快速地获取层次概念，高效地提取特征，为实体分类打标签，弥补了领域知识不足的缺陷。

3.4 基于依存句法分析的关系抽取

经过上文中的方法，我们已经快速构建出了一部分先验知识，但是仍欠缺高质量的语义关系，虽然手工整理了一部分，但是扩展性不好，效率不高。目前对于关系的抽取，大多数人使用卷积神经网络训练模型，但是这需要大量的标注数据，仍旧依托于领域知识。因此这里采用了无监督的关系抽取方法，基于依存句法分析以动词为核心构建三元组。

依存分析指的是利用句子中各成分间的依赖关系，展示出句子的句法结构。依存分析的核心思想是，认为句子中的核心是动词，而其他的成分都被核心动词支配。也就是说，依存分析不依靠词语的位置去识别句子中的语法成分，而是分析各词语间的语义修饰关系，获取了“主谓宾”等信息，解决了句子特征远距离的问题。在中医典籍中，谓语是动词或者动词短语的动词谓语句占了很大的比重，依存分析的关系类型及对应代号如下表所示：

表 3-6　依存分析标注关系

关系类型	依存弧	例子
动宾关系	VOB	我送她一束花(送→花)
间宾关系	IOB	我送她一束花(送→她)
定中关系	ATT	红苹果(红←苹果)
状中结构	ADV	非常美丽(非常←美丽)
动补结构	CMP	做完了作业(做→完)

以《黄帝内经》为例进行预处理，按照标点符号“。”、“？”、“!”分句，利用上文已获得的实体词表进行筛选，保留包含两个及两个以上实体的句子。本实验调用了哈尔滨工业大学开源的自然语言处理包——LTP 工具，其中的分词 WS、词性标注 POS、句法分析 Parser、实体识别 NE 四个模型进行依存句法分析，得到中医典籍中以谓词为中心的三类事实三元组，包括主谓宾、定语后置的动宾关系、含介宾关系的主谓动补关系。将新识别的动词添加进原有的基础动词关系词表，采用 Bootstrapping 思想，进行迭代运算，筛选出语料中其他存在动词关系表中动词的句子，再次利用依存句法分析得到三元组，将新识别的实体存入原有的基础名词词表中，再次迭代，直到不再能够识别出新的实体和动词，从而快速扩充了中医典籍的动词词表和名词词表，可用于层次聚类进行分类，提取特征，从而得到了中医典籍的先验知识。

总的来说，中医典籍种子知识的快速获取方法整体流程如下图所示，主要分基于层次聚类的实体获取和基于依存句法分析的关系抽取两个模块，其中关系抽取模块为一个迭代过程，可扩充实体词表和关系词表。

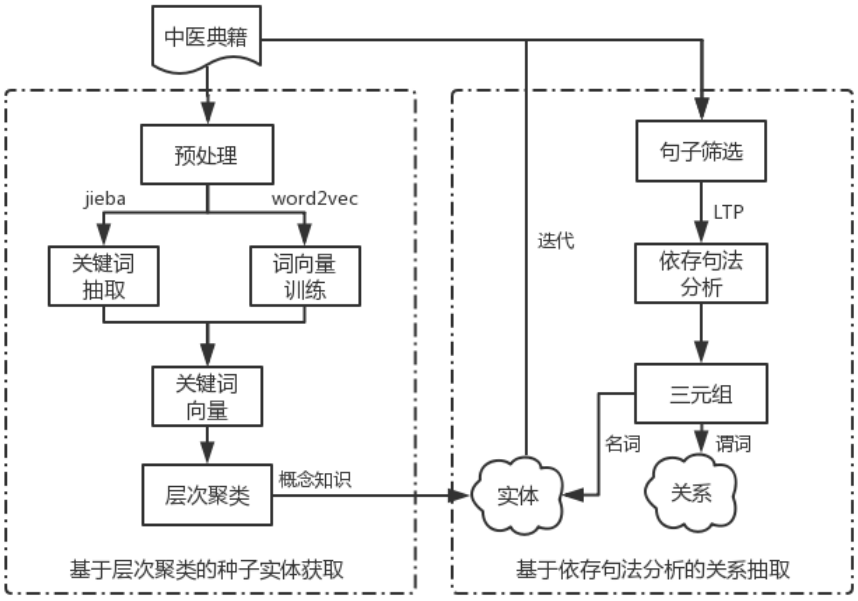

图 3-1　中医典籍先验知识获取流程图

以《黄帝内经》为例，利用该方法进行处理，最后得到了 2683 个事实三元组，下面展示的是部分结果。

表 3-7　中医典籍的修辞手法

动词	三元组	动词	三元组
受于	(阳,受入,六腑) (阴,受入,五脏)	知	(鼻,知,臭) (舌,知,五味)
入于	(辛,入于,胃) (卫气,入于,阴)	走	(厥气,走,喉) (辛,走,气)
走于	(胃,走于,阳明) (上,走于,息道)	治	(针石,治,外) (汤液,治,内)
至为	(太阴,至为,埃溽) (少阳,至为,炎暑)	病	(厥阴,病,阴痹) (太阳,病,骨痹)
通于	(天地,通于,肺) (风气,通于,肝)	有	(人,有,四经) (病,有,标本)
属于	(诸髓,属于,脑) (诸筋,属于,节)	应	(左胁,应,春分) (左手,应,立夏)
始于	(四气,始于,二刻) (春气,始于,下)	无	(刺筋,无,伤骨) (刺骨,无,伤髓)
起于	(少阴,起于,涌泉) (厥阴,起于,大敦)	为	(胃,为,仓廪官) (其色,为,苍)
名曰	(手阳明,名曰,偏历) (手少阳,名曰,外关)	为	(冬,为,飧泄) (火,为,阳)
留于	(热气,留于,小肠) (卫气,留于,腹中)	令	(肺疟,令,心寒) (心疟,令,烦心)
发于	(阴病,发于,骨) (阳病,发于,血)	禁	(心病,禁,咸) (脾病,禁,酸)
出于	(心,出于,中冲) (肺,出于,少商)	当	(黑,当,肾碱) (青,当,肝酸)
治以	(厥阴胜,治以,甘清) (热反胜,治以,苦)	藏	(脾,藏,肉) (肝,藏,血)
外合于	(厥阴,外合于,浑水) (太阳,外合于,淮水)	生	(暑燥,生,寒) (心,生,血)
受气于	(肝,受气于,心) (心,受气于,脾)	如	(夏脉,如,钩) (冬脉,如,营)

3.5 本章小结

本章节首先对中医典籍的语言特点进行深入分析，根据其特点手工构建了领域基础词表。此外提出了一种基于无监督学习的先验知识快速获取方法，利用关键词提取、依存句法分析和迭代思想扩充了领域词表，构建了以动词为核心的中医典籍事实三元组，利用基于词向量的层次聚类方法对实体进行分类，获取层次概念，用于特征提取，大大减少了深度学习训练集的人工标注工作量，减少了对领域知识的依赖，同时成为了分类的参考依据，与深度学习的实验结果相互验证。

4 基于 BiLSTM-CRF 的命名实体识别

命名实体识别是一个序列标注问题：给定一个句子，为句子序列中的每一个字做标注。2016 年，Guillaume[65] 等人首次将结合了字、词向量的双向长短期记忆神经网络和条件随机场(BiLSTM-CRF)模型应用于实体识别任务，并被验证具有更强的泛化性、更少依赖人工特征。Jagannatha[66] 等人实验对比了 CRF、BiLSTM、BiLSTM-CRF 三种模型对英文电子病历的实体识别效果，表明所有基于 LSTM 的模型都比 CRF 具有更好的效果，且 BiLSTM-CRF 模型能够进一步提高 2%至 5%的准确率。目前，中医命名实体识别任务中最主流的深度学习模型也是 BiLSTM-CRF 模型，但是大多仍停留在处理对电子病历和现代医案，识别的均为药材、病名等简单实体。

中医典籍的句子多为复合型长句，重要的信息可能出现在句子的任何位置，因此需要获取过去和未来的信息进行预测。BiLSTM-CRF 模型能获取过去和未来的上下文信息，非常适用于中医典籍的实体识别。根据深度神经网络特点和中医典籍的语言特点，本文创新性地将中医典籍字向量训练与深度学习相结合，提出了中医典籍中复合实体识别方案，得到了较好的实验效果。

4.1 模型与算法

双向长短期记忆神经网络和条件随机(BiLSTM-CRF)模型一共可分为四层，分别是输入层、Embedding 层、BiLSTM 层和 CRF 层。以输入“冬为飧泻”为例，最后输出每个字的预测标签“S-ZR”、“O”、“B-BL”、“E-BL”，模型的详细结构如图 4-1 所示：

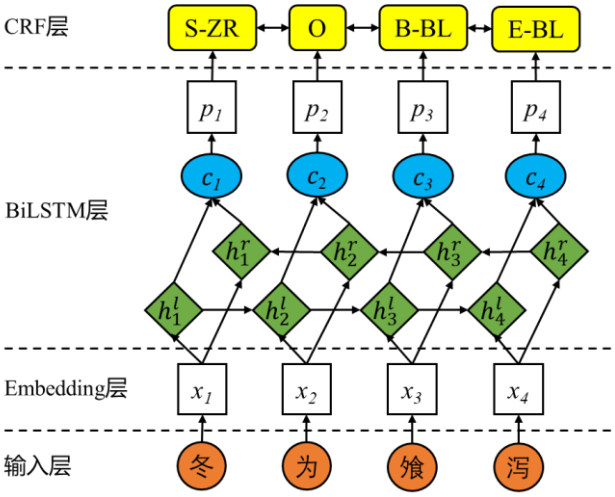

图 4-1　基于字的 BiLSTM-CRF 模型结构

第一层为输入层。目前大多模型的输入都是词语，这就导致实体识别的效果严重依赖于分词的效果。而第三章中已经分析了中医典籍的分词难点，没有明显表示词的边界标识符，且存在着大量通假字、生僻字，表述不统一，语言富有浪漫主义色彩，难以获取好的分词文本。中医典籍的语言非常精炼，而且有韵律感，如“法于阴阳，和于术数”，包含相当多的嵌套式结构，即有些中医命名实体又包含它的子实体，比如“阴阳”是一个实体，同时又可以进一步拆分为“阴”和“阳”。也就是说中医典籍中的“字”就包含了大量语言信息，因此该模型以“字”作为初始输入，规避了分词效果不佳带来的错误累积。将一个包含 n 个字的句子记作 W=(w_1,w_2,w_3,…,w_n)，构成一个字典，其中 w_i 是句子的第 i 个字在字典中的 id，进而可以得到每个字的 one-hot 向量，维数是字典大小，即字的个数。

第二层是 embedding 层，目的是将第一层的输入的字序列中各个字映射为相应的向量，传递给 BiLSTM。原理是利用预训练或随机初始化的 embedding 矩阵将句子中的每个字 w_i 由 one-hot 向量映射为低维稠密的字向量(character embedding)。为了解决深度神经网络模型中，要训练大量的参数，而训练数据集的大小有限的问题，我们采用基于无监督学习方法，即使用高质量的预训练结果进行参数的初始化，从而得到更好的效果。本文使用 Word2vec 对语料进行预训练得到字向量，作为 embedding 层的初始化参数。对于输入的 one-hot 字向量，经过 embedding 层转变得到了新的字向量 X = (x_1,x_2,x_3,…,x_n)，其中 x_i 就是预训练中维度指定为 d 的向量。为防止过拟合，加入了一层 dropout 正则化机制。

第三层是双向 LSTM 层，用于自动提取句子特征。双向的网络结构可以使当前时序的输出充分考虑过去和未来的上下文信息，得到当前字的上下文表征向量。本文中将 embedding 层获得的字向量(x_1, x_2, x_3, …, x_n)，作为双向 LSTM 各个时间步的输入，经过前向 LSTM 得到了左侧每个字的输出隐状态 $H^L = (h_1^l, h_2^l, h_3^l, ..., h_n^l)$。同理，经过后向 LSTM 得到了右侧的输出隐状态 $H^R = (h_1^r, h_2^r, h_3^r, ..., h_n^r)$。再将各个位置输出的隐状态按位置拼接得到第 i 个位置的序列 $c_i = [h_i^r, h_i^r]$，最终得到完整的隐状态序列 C =(c_1,c_2,c_3,…,c_n)。在加入 dropout 后利用一个全连接层（U,b），将隐状态向量映射到 k 维，k 是标注集的标签数，从而得到自动提取的句子特征，即 P=(p_1, p_2, ... , p_n)，p_i 的每一维 $p_{i,j}$ 都可以看成是将字 w_i 分类到第 j 个标签的分值。

最后一层为 CRF 层，用于进行句子级的序列标注，保证在全局上生成最优标注序列。BiLSTM 层输出的 p_i 之间是相互独立的，也就是说对于当前位置的标签预测，只与该位置的 p_i 有关，虽然我们可以使用分类算法，如

softmax 对每一个位置取分值最大值进行 k 类分类，但事实上前后的标注结果之间具有强依赖性，我们不能保证标签每次都是预测正确的，因此需要利用该位置的前后输出信息。利用 CRF 层可以再训练过程中，从训练数据里自动学习获得一些约束性的规则，譬如：句子中第一个词总是以标签“B-”、“S-”或“O”开始，而不可能是“I-”和“E”;“B-SL”(中医生理实体的起始)，接下来只会是“I-SL”或“E-SL”，不可能是“O”或者“S-SL”,也一定不可能是“I-ZF”(治则治法实体的中间内容)。因此，使用 CRF 对整个句子进行联合建模，非法序列出现的概率将会大大降低，提升了标签序列预测的准确率。

CRF 层的参数是一个 (k+2)×(k+2) 的状态转移矩阵 A，加 2 是因为考虑到在句子首部添加一个起始状态并且在句子尾部添加一个终止状态。$A_{i,j}$ 表示的是从第 i 个标签到第 j 个标签的转移得分。假设对于一个输入句子 W= (w_1,w_2,w_3,…,w_n)，得到一个预测标签序列 y = (y_1, y_2, ... , y_n)，那么定义这个预测的得分为：

$$s(W,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{4.1}$$

其中 P_{i,y_i} 为第 i 个位置 BiLSTM 输出为 y_i 的概率， $A_{y_i,y_{i+1}}$ 为从 y_i 到 y_{i+1} 的转移概率，可以看出整个序列的打分等于各位置的打分之和，每个位置的打分由 BiLSTM 输出的 p_i 和 CRF 的转移矩阵 A 共同决定。这样一来，我们不再单纯将各个位置输出的 p_i 中最大概率值对应的标签作为最终标签，还需要考虑符合输出规则，例如假设 BiLSTM 输出的最有可能序列为 BBEO，但是转移概率矩阵中 B->B 的概率很小甚至为负，s 得分就不会是最高的，那么就排除了这个不符合的预测序列。

对于每个训练样本 W，利用 Viterbi 算法求出所有可能的标注序列 y 的得分 s(W,y)，然后利用 softmax 层对所有得分实现归一化，最终得到序列 y 的概率为：

$$p(y\mid W)=\frac{e^{s(W,y)}}{\sum_{\bar{y}\in Y_W}e^{s(W,\bar{y})}}) \tag{4.2}$$

模型训练时，对于句子输入序列 X，损失函数设置为对目标真实标记序列 Y 的概率取对数。为了使真实标记序列对应的概率最大化，我们采用取负值然后最小化的方法，引入梯度下降算法来求解参数，最大化 log 似然函数：

$$\begin{aligned}&\log(p(Y\mid X))=s(X,Y)-\log(\textstyle\sum_{\bar{Y}\in Y_x}e^{s(X,\bar{Y})})\\&=s(X\mid Y)-\log(\textstyle\sum_{\bar{Y}\in Y_x}s(X,\overline{Y}))\end{aligned} \tag{4.3}$$

模型的训练过程伪代码如下：

算法 2：BiLSTM-CRF 模型训练算法

输入：字序列$W = (w_1,w_2,\cdots,w_n)$，目标标签序列$T = (t_1,t2,\cdots,t_n)$

```
function Train_BiLSTM-CRF(W,T)
    X = (x1,x2,⋯,xn) ← embedding(W)
    for each echo do
        for each batch do
        H^L = (h_1^l, h_2^l, h_3^l, ..., h_n^l) ← Left_LSTM(X)
        H^R = (h_1^r, h_2^r, h_3^r ,..., h_n^r) ← Right_LSTM(X)
            for h^l, h^r ∈ H^L, H^R do
                C.append([h^l, h^r])
            end for
        P = (p1, p2, ... , pn) ← f(U,b,C)
        loss ← CRF_Log_likelihood(P,T)
        loss_gradient ← Adam_Optimizer(loss)
        updata_by_gradient(loss_gradient)
        end for
    end for
end function
```

在预测过程时，根据训练好的参数求出所有可能的 y 序列对应的 s 得分，使用动态规划的 Viterbi 算法来求解最优路径，预测结果记为 Y^*:

$$Y^* = \arg\max_{\overline{Y}\in Y_X}(s(X,\overline{Y})) \tag{4.4}$$

算法 3：BiLSTM-CRF 模型预测算法

输入：字序列 $W = (w_1,w_2,\cdots,w_n)$

输出：预测序列 $Y^* = (y_1,y_2,\cdots,y_n)$

```
function Predict_BiLSTM-CRF(W,T)
    X = (x1,x2,⋯,xn) ← embedding(W)
    H^L = (h_1^l, h_2^l, h_3^l, ..., h_n^l) ← Left_LSTM(X)
    H^R = (h_1^r, h_2^r, h_3^r ,..., h_n^r) ← Right_LSTM(X)
    for h^l, h^r ∈ H^L, H^R do
        C.append([h^l, h^r])
    end for
    P = (p1, p2, ... , pn) ← f(U,b,C)
    transition_params ← crf_log_likelihood(P,T)
    Y* ← viterbi_decode(transition_params)
end function
```

4.2 实现与分析

4.2.1 实验数据

目前，深度神经网络 BiLSTM-CRF 模型在中医领域上的应用仍局限于现代医案、电子病历的现代文的文本，识别的实体也局限于药方、病症等便于分词的简单实体。事实上，中医知识博大精深，中医是一种“分类医学”[67]，通过对阴阳、五行、人体、自然认识，建立了一种独特的分类体系，包括对人体、病因、病机、病症、治法、方药的分类。

根据前期对中医典籍的理解和归纳，我们以张德政教授[19]提出的基于本体的中医知识体系为指导，将中医知识划分为中医认识方法、中医生理、中医病理、辨证论治四大部分。其中，中医认识方法包括阴阳、五行学说、天干地支、数字等概括总结形成的术语；中医生理包括脏腑、气血、津液、精神、形体、官窍、情志、经络穴位、脏腑生理功能、脏腑生理特性等概念；中医病理包括疾病、病因、病机、症状等概念，是中医生理的异常情况；辨证论治则包括辨证方法、证候、治则、治法、方药、性味归经等概念。

本文的目标是从中医典籍中抽取中医知识并自动构建中医知识图谱，这里我们选用从网络中爬取的《黄帝内经》作为实验语料，根据对《黄帝内经》的理解，添加了特有而重要的实体类别——中医自然，包括了季节、方位、时间、颜色、味道、动植物、人名等实体。最终将《黄帝内经》中的命名实体划分为 5 类，分别是中医认识方法、中医自然、中医生理、中医病理和治则治法。并且结合第三章中基于层次聚类得到种子实体，验证了上述分类方法的正确性。

1) 实验语料的获取

确定了语料中命名实体的分类后，开始对网络爬取的语料进行数据预处理，包括清除乱码、统一标点符号、将繁体字全部转换为简体字、统一通假字的转换，如“四支”统一为“四肢”、“五藏”统一为“五脏”等。同时考虑到中医典籍可能存在无标点符号短句和存在大量复合长句的情况，为了减小模型训练的计算压力，加快训练速度，实验中将《黄帝内经》中的长句进行了分割，设置参数将其分割成最长为 20 个字的句子。针对数据集，我们采用了 BIOES 标注方式，即 B 表示实体的开始部分，I 表示实体的中间部分，E 表示实体的结尾部分，S 表示单个字符的实体，而非实体部分用 O 表示。同时 FF 表示中医认识方法，ZR 表示中医自然，SL 表示中医生理，BL 表示中医病理，以及 ZF 表示治则治法。具体的标注表如下表所示：

表 4-1 《黄帝内经》实体标注 BIOES 标签表

实体类别	单字符实体	多字符实体		
	标记	开始标记	中间标记	结束标记
中医认识方法	S-FF	B-FF	I-FF	E-FF
中医自然	S-ZR	B-ZR	I-ZR	E-ZR
中医生理	S-SL	B-SL	I-SL	E-SL
中医病理	S-BL	B-BL	I-BL	E-BL
治则治法	S-ZF	B-ZF	I-ZF	E-ZF
非实体标记	O	O	O	O

为减少人工标注语料的工作量及难度，我们采用 3.3 章节中获得的结巴分词自定义词表，该词表融合了根据中医典籍特点人工整理的词语以及搜狗细胞词库、百度百科、中医网站中的中医术语，随后利用层次聚类将词表进行聚类，给每个词赋予了一个标签。采用词表匹配方式对数据集进行自动 BIOES 打标签，并辅以人工校对方式快速构建所需的实验语料。该数据集共标注了 27642 个样本，将其中的 60%作为训练集，用于网络训练生成模型，20%作为验证集来选择最优训练模型，20%作为测试集，查看当前模型的识别效果，防止过拟合。

表 4-2 《黄帝内经》实体标注实验语料表

数据集	标注数量	中医认识方法	中医自然	中医生理	中医病理	治则治法
训练集	16585	2515	2862	8449	2252	507
验证集	5428	859	903	2816	712	138
测试集	5629	818	1005	2817	791	198

2) 预训练语料的获取

上文曾介绍，在 BiLSTM-CRF 模型的 embedding 层中，我们为了获得更好的效果，使用了预训练的字向量来初始化参数。实验中，我们利用 3.3 章节中爬取得到的 700 本中医典籍语料，融合了实验室语料库中拥有的 30 万份名老中医医案和《中华历代名医医案全库》中 1 万多份古医案，对其进行合并以及数据预处理，包括繁简体转换、移除 non-utf8 字符、统一不同类型标点符号、空格处理等，随后全部拆分成字，得到了 3.84G 的预训练语料。实验中，我们使用 Google 开源的词向量生成工具 Word2vec 的 python 版本——gensium 工具包，进行字向量的训练。使用 Skip-Gram 模型，窗口大小设为 10，从而得到 200 维的字向量，用于预训练输入。

整个中医命名实体识别实验的主要流程如下图所示：

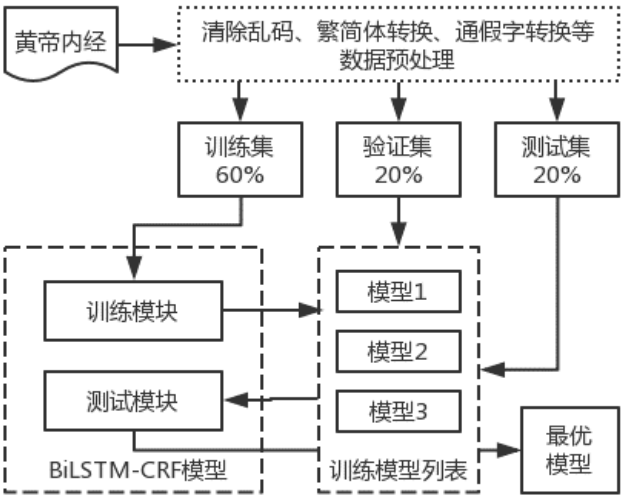

图 4-2 《黄帝内经》中医实体识别流程图

4.2.2 评价指标

由图 4-2 可知，实验是一个迭代寻找最优模型的过程，本实验使用最基本的 F_1 值作为算法性能的评价指标，而 F_1 值是精确率和召回率调和均值。首先我们定义四个参数 TP、FP、FN、TN，其含义如下表所示：

表 4-3 测试数据结果划分

	相关(Relevant),正类	无关(NonRelevant),负类
被检索到	true positives(TP 正类判定为正类数)	false positives(FP 负类判定为正类)
未被检索到	false negatives(FN 正类判定为负类)	true negatives(TN 负类判定为负类)

精确率(Precision, P)是正确被检索的实体数(TP)占所有实际被检索到的实体数(TP+FP)的比例，准确率计算公式为：

$$P(\%) = \frac{TP}{TP + FP} \times 100 = \frac{\text{正确的实体数}}{\text{抽取出的实体数}} \times 100 \tag{4.5}$$

召回率(Recall, R)是所有正确被检索实体数(TP)占所有应该被检索到的实体数(TP+FN)的比例。召回率计算公式为：

$$R(\%) = \frac{TP}{TP + FN} \times 100 = \frac{\text{正确的实体数}}{\text{语料中包含的实体数}} \times 100 \tag{4.6}$$

F1 值(F1)是对精确率和召回率的调和均值，它表示对精确率和召回率的综合考量。使用 F_α 能更好评估结果，F_α 值的原始公式为：

$$F_{\partial}(\%) = \frac{(\partial^2 + 1)P \bullet R}{\partial^2 P + R} \times 100 \tag{4.7}$$

其中，α用于调整精确率与召回率的权重，当α>1 时召回率更重要，当α<1 时，精确率更重要。通常情况下，我们将 α 默认为 1，即精确率与召回率的权重相同，即 F_1 值，计算公式为：

$$F_1(\%) = \frac{2P \bullet R}{P + R} \times 100 \tag{4.8}$$

4.2.3 参数设置

模型中，预训练的字向量维度设置为 200 维，由于考虑了到通常网络越复杂对训练数据的拟合效果越好，结合语料的规模，将 LSTM 的神经元数量设置为 50 个。为了解决梯度爆炸问题，采用 clip gradients 的方法，将权重控制在一定范围之内，使得最终的 loss 能下降到满意的结果，实验中将梯度阈值 clip 设置为 5。同时为了防止数据量过小或者说是网络结构复杂导致的模型过拟合现象，我们在训练过程中引入了 dropout 正则化机制。Dropout 的原理是对每一层的神经元以 P 概率随机剔除，用余下的神经元所构成的网络来训练迭代中的数据，这样也就减少了中间特征的数量，从而减少冗余，增加了每层各个特征之间的正交性，弱化了各个特征之间由于数据量太小导致产生的过多的相互作用，缓解了过拟合。Dropout 多用于上下层连接的时候，因此我们在 BiLSTM 层的输入和输出端都加入 dropout 层，设定 dropout 层的概率为 0.5。

模型采用反向传播算法拟合训练数据，针对每一个训练样例更新参数。为了改善模型训练方式，优化调整模型更新权重和偏差参数的方式，我们采用了 Adam 梯度下降算法，因为它的收敛速度更快，学习效果更好，而且可以纠正其他优化技术中存在的损失函数波动较大等问题。此外，还有一些超参数的设置，如：学习率、迭代次数、批大小等。

表 4-4　命名实体识别模型超参设置表

参数	值	参数	值
Character embedding size	200	Learn rate	0.001
Hidden units number	50	Batch_size	20
Clip gradients	5	Max_epoch	200
Dropout rate	0.5	Steps_check	50
梯度下降算法	Adam		

4.2.4 实验结果

本文的实验环境：(1) 操作系统为 Red Hat 4.8.5；(2) 内存 120G；(3) 硬盘 500G；(4) CPU 为至强 Xeon E5-2600V2，8 核；(5) 显卡为华硕 GTX1080TI，11G。以下结果为运行 3 次得到的最优模型各项分值的平均值。模型的精确率为 85.44%，召回率为 85.19%，F_1 值为 85.32%。

表 4-5 各个实体类别实验结果

实体	Precision(%)	Recall(%)	F1(%)
模型均值	85.44	85.19	85.32
中医病理（BL）	68.47	62.81	65.52
中医认识方法（FF）	93.86	86.99	90.30
中医生理（SL）	88.35	87.05	87.70
中医自然（ZR）	81.58	84.94	83.19
治则治法（ZF）	86.36	70.37	77.55

由实验结果可知，模型对中医认识方法实体识别效果最好，对中医病理的识别效果最差。结合标注的语料进行分析，可以得知中医认识方法实体的标注量较多，并且区分度最高，均为阴阳、数字等词，存在较少的歧义性；而中医生理训练语料中的标注的数量非常多，至少为其他类别标注数量的 3 倍，因此模型对其特征的学习效果也较好；中医自然的实体标注数量也较多，但是单字实体与其他类别的冲突较多，譬如“水”有的属于五行，有的属于自然，而属于五行的概率远高于自然，降低了对自然识别的准确率。治则治法虽然训练样本非常少，但是它的实体区分度很高，与其他类别的实体几乎没有冲突，因此也得到了较好的效果。然而中医病理，虽然也有较多的训练样本，但是它的词语大多边界不清，大部分是中医生理与病症的结合，譬如“肝痹”可以进一步拆分为“肝”和“痹”，“气血虚”可以进一步拆分为“心血”和“虚”，“气血”还可以进一步拆分为“气”和“血”；若按细粒度切分，则这个词属于多种实体组合而成，极大程度影响了实体识别的精度。

总的来说，中医典籍的命名实体识别准确率与两个因素有关：(1) 训练样本的数量；(2) 标注粒度。我们按照最大粒度和最细粒度两种方式进行标注，提高识别准确率，识别有层次概念的中医实体。

4.3 验证与对比

2015 年，百度研究员 Huang 等人[68] 便比较了 LSTM 网络、BiLSTM 网络、CRF 网络、LSTM-CRF 网络、BiLSTM-CRF 网络在自然语言处理任务上的性能，证明了 BiLSTM-CRF 网络有更好的效果。2017 年，Strubell[69] 等人[16]将 IDCNN-CRF 模型应用于实体识别，并说明该模型在训练速度和效果上优于 BiLSTM-CRF。因此，本章节通过不同模型效果、向量维度效果、组件参数效果的对比实验，验证了本文提出的模型和使用参数在中医典籍命名实体识别中的有效性。

4.3.1 不同模型效果对比

在传统机器学习中，CRF 模型被广泛用于序列标注问题，但是面向中医典籍的中文命名实体识别的研究相对匮乏，相似实验有：2015 年，孟洪宇[31]等人组合“字本身、词边界、词性、类别标签”的特征，用 CRF 进行《伤寒论》中的中医术语识别。因此本实验参考其方法，利用 CRF++工具包将 CRF 模型对中医典籍的识别结果作为 baseline。同时，为了充分利用 GPU 的处理性能，引入了 IDCNN-CRF 模型作为对比。各种深度神经网络模型的实验参数相同，字向量维度为 200，LSTM 中的隐单元数量和 IDCNN 中的过滤器数量一致。

表 4-6　实体识别模型不同模型实验结果

模型	参数组合	F1(%)
CRF	字符、词边界、类别标签	75.56
LSTM	字向量、pretrain、dropout	82.29
BiLSTM	字向量、pretrain、dropout	82.48
LSTM-CRF	字向量、pretrain、dropout	84.67
BiLSTM -CRF	字向量、pretrain、dropout	**85.32**
IDCNN-CRF	字向量、pretrain、dropout	85.01

由实验结果可知，CRF 模型的效果最差，F1 值仅有 75.56%。因为 CRF 模型采用窗口尺寸为 3 的上下文局部特征，特征提取能力有限，而重要的信息可能出现在句子中的任何位置，因此无法很好地处理长句的特征。而深度神经网络具有拟合非线性的能力，LSTM 可以很好地提取整个目标语句地特征，而 BiLSTM 相较于 LSTM，可以结合过去和未来的特征，F1 值又有了提

升。而加入 CRF 层后，F1 值提升了至少 2%，分析可知 CRF 层可以考虑标签之间的依赖关系。因此 BiLSTM-CRF 模型效果最好，F1 值达到了 85.32%，该模型可以有效利用过去和未来的特征标签预测当前的标签，提升了标记的准确性，减少了对词嵌入的依赖。而 IDCNN-CRF 的性能和 BiLSTM-CRF 仅相差 0.21%。总的来说，深度神经网络的特征提取能力远优于 CRF，而加入 CRF 层使深度神经网络可以利用句子级的标签，而 IDCNN 虽然网络结构较 LSTM 要复杂，参数较多，但是处理速度远高于 LSTM。

4.3.2 字向量维度效果对比

在本实验中，对比了不同字向量维度在实体抽取中的效果。本实验中模型使用 Word2vec 预训练的字向量，训练语料均相同，模型均为 BiLSTM-CRF。下表显示了不同维度的详细结果，以下数据均为运行 3 次求得的平均值。

表 4-7 实体识别模型不同字向量维度实验结果

维度	Precision(%)	Recall(%)	F1(%)
50	83.66	84.11	83.88
100	84.76	85.67	85.21
200	85.44	85.19	**85.32**
300	85.43	84.46	84.94
400	85.42	84.24	84.82

由实验结果可以看出，实验效果是一个先增后减的过程。50 维到 200 维 F_1 值逐步递增，而增速逐步放缓。到 300 维度之后，F1 值开始下降。分析可知，50 维、100 维时还不能对字进行充分的表征。随着字向量维度提高，需要训练的参数也成倍增多，模型训练过程中与实验数据的拟合情况逐渐变好，在 200 维时达到最优的平衡状态。此时维度继续增大，由于训练语料的限制，不再支持所需的参数得到充分训练，因此效果有所下降。因此在本文提到的实验中，均采用了 200 维的字向量。

4.3.3 不同参数效果对比

在 4.2.3 章节的实验中，我们在模型中可以有许多参数、组件的组合，例如选择预训练 pretrain、在 BiLSTM 输入端添加了一个 dropout 层，记作 L-dropout、在输出端添加了一个 dropout 层，记作 R-dropout，优化算法选择了 Adam。本实验中，我们比较不同参数、方法的组合对在实体抽取的效果影响。

实验结果如表 4.6 中所示，该表中的优化算法均选择的是 Adam。

表 4-8　实体识别模型不同参数组合实验结果

embedding 组件	dropout 组件	Precision(%)	Recall(%)	F1(%)
Pretrain	L-dropout+R-dropout	85.91	85.07	85.32
	L-dropout	85.37	86.04	85.64
	R-dropout	85.47	86.21	**85.75**
	-	84.76	85.29	85.02
Random	L-dropout+R-dropout	83.21	85.47	84.33
	L-dropout	84.66	85.28	84.96
	R-dropout	84.52	85.42	84.97

由上表中的结果可知，pretrain 字向量预训练的方法效果优于随机初始化矩阵，pretrain 可以更好的表征字的特征，更好地初始化 embedding 层的参数。而 dropout 也同样会影响实验效果，从实验结果看，加入两层 dropout 效果并非最优，两种 embedding 方式下均为单独使用 R-dropout 效果更好。其中，组合 Pretrain 和在输出端 dropout 层的效果最好，F1 值达到了 85.75。因此我们在后续的实验中选用 pretrain+R-dropout 组合。

在优化算法的对比实验中，我们在组合 pretrain、L-dropout 和 R-dropout 的情况下，对比验证了几个常用的优化算法，实验中参数均使用默认值。

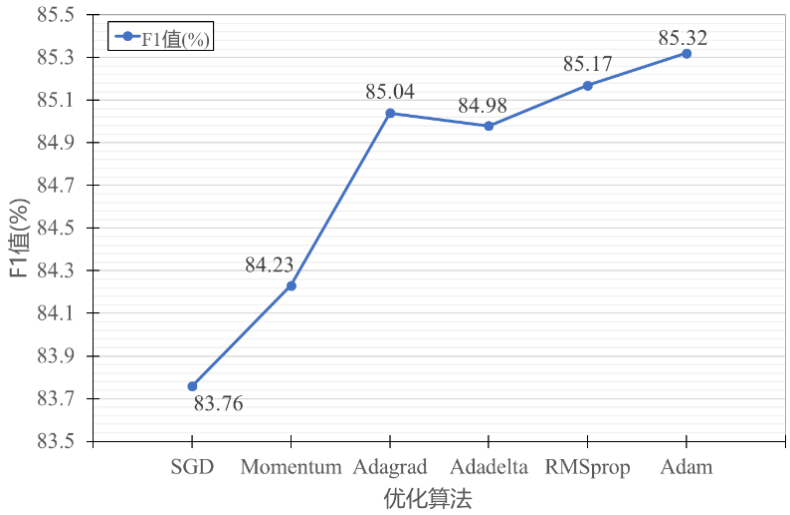

图 4-3　不同优化算法实验结果

可以看出，传统的 SDG 算法训练时间长，需要好的初始化和学习率调度方案。在没有人工调优的默认参数情况下，效果最差。Mementum 利用动量思想，在小的学习速率和中医典籍这种少量简单的数据集中效果不明显。

Adagrad、Adadelta 和 RMSprop 是比较相近的学习率自适应的优化算法，表现差不多，效果明显提升。在实际中最常用的调优方式为 Adma，也是一种学习率自适应的方法，减少了参数调优难度，在中医典籍中表现最好。因此，在实验中，我们也选用 Adam 算法作为梯度下降优化算法。

4.4 本章小结

本章针对中医典籍的文本特点与中医典籍实体抽取的难点，创新性地将字向量训练与深度学习的循环神经网络 BiLSTM-CRF 模型相结合，用于处理中医典籍实体识别任务，并对模型的结构进行了详细的介绍。模型以字向量作为输入，利用 BiLSTM 模型充分利用上下文的特征信息，最后利用 CRF 对输出的标签进行关联，赋予一些规则，提升了预测识别的准确率。同时以《黄帝内经》为实验语料，对模型进行了实验验证，详细介绍了语料准备方法以及评价指标，并且将该模型与其它相关模型进行比较，验证本模型的有效性。此外还设计实验对比了 embedding 维度、不同参数组合对实验效果的影响，选取效果最好的实验参数和组合。

5 基于 2Att-BiGRU 的实体关系抽取

关系抽取可以简单理解为一个分类问题：给定两个实体和两个实体共同出现的句子文本，判别两个实体之间的关系。目前主流方法是利用 CNN 或者双向 RNN 加注意力机制的深度学习方法。已有的文献大都是针对英文语料，使用词向量作为输入进行训练。本文尝试性地运用 LSTM 的简易变体网络——GRU，构建了一个加入字与句子的双重注意力机制的双向循环神经网络(2Att-BiGRU)模型，以天然适配中文特性的字向量作为输入，进行中文关系抽取，并验证了其在中医理论典籍中的有效性。

5.1 模型与算法

双向 GRU 加字级别 Attention 机制的模型想法来源于 Zhou[70] 等人 2016 年发表的文章，本文将原文的模型结构中的 LSTM 改为 GRU，且对句子中的每一个中文字符输入字向量，这样的模型对每一个句子输入做训练。加句子级别 Attention 的想法来源于 Lin[71] 等人 2016 年发表的文章，本文将其中对每个句子进行 encoding 的 CNN 模块换成了双向 GRU 模型，这样的模型对每一种类别的句子输入做共同训练。整个模型可以分为两大部分，如下图所示。

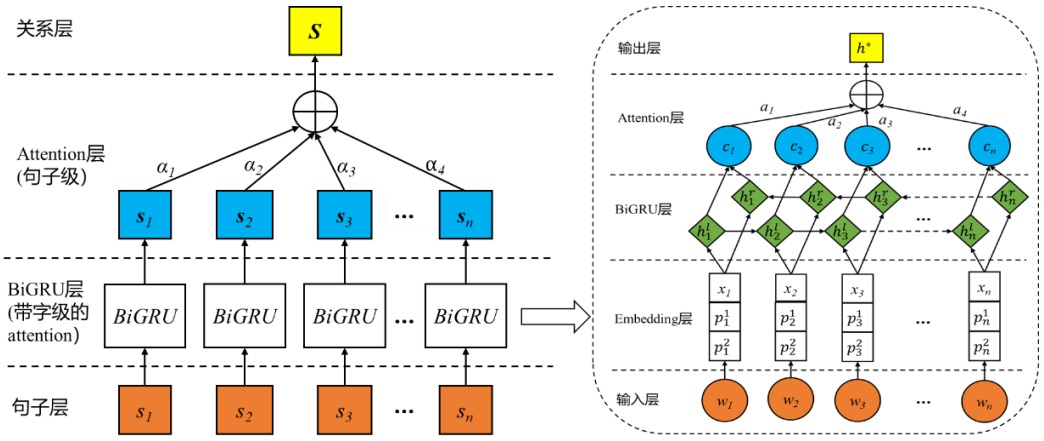

图 5-1 基于字、句双重 Attention 机制的 BiGRU 模型结构

首先，右侧的 BiGRU 模型是为了实现加图字级别的 Attention 机制，目的是为了赋予每一个单词不同的权重，衡量其对关系分类的重要诚度，最后得到完整的句子向量。该模型与上一章介绍的 BiLSTM 模型十分类似。

第一层是输入层，为了减少分词误差累积的错误问题，这里将每句话拆分为字，进行输入。假设一个句子 W 有 n 个字，则输入的句子集合为 W=(w_1,w_2,w_3,…,w_n)。

第二层是 Embedding 层，这里我们首先利用 Word2Vec 预训练的字向量，将输入的每个字映射为表征较好的字向量，记作 X = (x_1,x_2,x_3,…,x_n)。此外，由于中医语句中实体涵盖非常密集，往往包含了多个实体，比如“东方生风，风生木，木生酸，酸生肝，肝生筋。”如果将句子整体作为输入，无法体现不同实体对间的差异性，所以我们还引入了位置向量。针对每个字 w_i，它与句子中实体对 m_1 和 m_2 有两个位置信息 p_1 和 p_2。对于位置 p，若字在实体左侧则为负，在实体右侧则为正，若为实体则为 0。随机初始化两个距离映射矩阵，将 p_1 和 p_2 映射为位置向量p_i^1和p_i^2。最终，将字向量和位置向量进行拼接，得到句子每个字的表征向量，X = (e_1,e_2,e_3,…,e_n)，其中 e_i = [x_i, p_i^1, p_i^2]。

第三层是 BiGRU 层，与第 4 章命名实体识别中提到的作用一样，为了根据每个字上下文的向量，获取的高维度特征。利用 Embedding 层得到的向量，进行前向和后向的神经网络学习，得到 $H^L = (h_1^l, h_2^l, h_3^l, ..., h_n^l)$，$H^R = (h_1^r, h_2^r, h_3^r, ..., h_n^r)$，将各个位置输出的隐状态按位置拼接得到第 i 个位置的序列 $c_i = [h_i^r, h_i^r]$，最终得到每个字的输出向量 C =(c_1,c_2,c_3,…,c_n)。

第四层是 Attention 层，目的是为了生成权重向量，并将每一步的字级特征向量按照权重合并成句子级特征向量。过程为：首先将向量集合 C 中的每个字向量映射到[-1,1]的范围内。

$$M = \tanh(C) \tag{5.1}$$

计算各个字的权重向量 A = (a_1,a_2,⋯,a_n)，其中a_1+a_2+⋯+a_n =1。

$$a = soft\max(\omega^T M) \tag{5.2}$$

随后对各向量进行加权求和，增加特征的影响，得到句子向量 R。

$$h = Ca^T = a_1 \times c_1 + a_2 \times c_2 + ... + a_n \times c_n \tag{5.3}$$

最后对将每个 H 的值映射到[-1,1]范围内，得到最终的句子向量 h*。

$$h^* = \tanh(h) \tag{5.4}$$

其次，左侧模型的整体上添加了句子层级的 Attention 机制，字级别的 Attention 模型只在一个与关系 r 最有关的句子上进行训练，未充分利用了语料。而句子级别的 Attention 模型对包含实体对的所有有效句子计算句子权重，既充分利用了语料，又降低了错误标注句子所带来的影响。

第一层输入包含实体对 m_1 和 m_2 的 n 个句子的集合 S =(s_1,s_2,s_3,…,s_n)。

第二层利用加入字符级别 Attention 机制的 BiGRU 模型得到每个句子的向量 S =($\boldsymbol{s}_1$,$\boldsymbol{s}_2$,$\boldsymbol{s}_3$,…,$\boldsymbol{s}_n$)。

第三层为句子级别的 Attention 层，得到。我们先将关系 r 映射为向量 $\boldsymbol{r}$，再乘上该句子向量 $\boldsymbol{s}_i$ 和一个加权对角矩阵 A，得到了句子与该关系的匹配程

度 u_i，可以看出 u_i 取决于 $\boldsymbol{s}_i$ 在 $\boldsymbol{r}$ 上的映射的大小，那么与该实体关系更加密切的句子也就得到了更大的值。

$$u_i = \boldsymbol{s}_i A \boldsymbol{r} \tag{5.5}$$

再利用注意力机制给每个句子分配权重 α_i。

$$\alpha_i = \frac{\exp(u_i)}{\sum_k \exp(u_k)} \tag{5.6}$$

然后对句子集合 S 求得句子集合向量 $\boldsymbol{L}$，再通过一层网络得到最终句子集合向量 $\boldsymbol{S}$。

$$\boldsymbol{L} = \sum_i \alpha_i \boldsymbol{s}_i \tag{5.7}$$

$$\boldsymbol{S} = M\boldsymbol{L} + d \tag{5.8}$$

其中，M 是所有实体关系的向量所组成的矩阵，d 是偏置向量，使最后网络的输出向量维度等于关系的数目。

最后将向量 $\boldsymbol{S}$ 作为输入，利用一个 softmax 层计算出句子集合 S 属于各个关系 r 的概率值。

$$p(r|S,\theta) = \frac{exp(\boldsymbol{S}_r)}{\sum_{k=1}^{\boldsymbol{n}_r} exp(\boldsymbol{S}_k)} \tag{5.9}$$

其中，n_r 是数据集中所有包含的关系数目。

为了学习其中所有模型的参数θ，我们所有的训练句子集合和对应正确的关系记作训练集合(S_i, r_i)。在训练阶段，分类时需要最大化的是在网络参数下某实体关系的概率，用梯度下降算法最大化 log 似然函数，即损失函数为：

$$J(\theta) = \sum_{i=1}^{s} \log(p(r_i \mid S_i, \theta)) \tag{5.10}$$

模型训练过程的伪代码如下：

算法 4：字级别 Att-BiGRU 模型训练算法

输入：句子集合 S = (s_1,s_2,s_3,…,s_n),每句字序列 W = (w_1,w_2,⋯,e_1,⋯,e_2,⋯,w_n)，目标标签 t

1: ***function*** *Train_2ATT-BiGRU(S,W,t)*

2: ***for*** $s_i \in S$ *do*

3: $s_i = W_i = (w_1,w_2,\cdots,e_1,\cdots,e_2,\cdots,w_n)$

4: $X_i = (x_1,x_2,\cdots,x_n) \leftarrow embedding(W_i)$

5: $H^L = (h_1^l, h_2^l, h_3^l, \ldots, h_n^l) \leftarrow Left\_LSTM(X_i)$

6: $H^R = (h_1^r, h_2^r, h_3^r, \ldots, h_n^r) \leftarrow Right\_LSTM(X_i)$

7: ***for*** $h^l, h^r \in H^L, H^R$ *do*

算法 4：字级别 Att-BiGRU 模型训练算法（续）

C_i.append([h^l, h^r])
    end for
    $\boldsymbol{s}_i \leftarrow attention\_char(C_i)$
    $\boldsymbol{S}$.append($\boldsymbol{s}_i$)
  end for
  $S^* \leftarrow attention\_sentence(\boldsymbol{s}_i)$
  $loss \leftarrow softmax\_loss(S^*,t)$
  $Adam\_Optimizer(loss)$
  $loss\_gradient \leftarrow Adam\_Optimizer(loss)$
  $updata\_by\_gradient(loss\_gradient)$
end function
function $attention\_char(C_i)$
    $M \leftarrow tanh(C_i)$
    $A \leftarrow softmax(w^T M)$
    $h \leftarrow C_i\, a^T$
    $s_i \leftarrow tanh(h)$
    return s_i
end function
function $attention\_sentence(\boldsymbol{s}_i)$
    $u_i \leftarrow \boldsymbol{s}_i A \boldsymbol{r}$
    $\alpha_i \leftarrow \frac{exp(u_i)}{\sum_k exp(u_k)}$
    $L^* \leftarrow \sum_i \alpha_i \boldsymbol{s}_i$
    $S^* \leftarrow \mathrm{M}L^* + \mathrm{d}$
    return S^*
end function

在预测阶段，预测结果记为 y，则选取对每个关系预测求得的分值中最大值对应的标签：

$$y = \arg\max_{r \in R}(p(r \mid S)) \tag{5.11}$$

5.2 实现与分析

5.2.1 实验数据

中医领域关系繁多，大多数实验均按照本体概念层面分为病位关系、证症关系、证治关系、方治关系、药治关系、方证关系、药证关系、方症关系、

药症关系、症因关系、病证关系、药性关系、药味关系。然而中医典籍的关系层次更高，更为复杂。以《黄帝内经》作为实验语料，不单单如传统以往只抽取简单的通用语义关系，而要进行自定义语义关系的抽取。我们根据第三章得到的先验知识和提取的动词特征，结合网络中对其的注解，将《黄帝内经》中的实体关系归纳总结为 6 大类，分别是表征关系、概念关系、促进关系、抑制关系、因果关系和包含关系。此外还有一类为未知(unknown)关系，代表该实体对出现在同一句子中，存在某种关系，但不确定的情况。

由于中医典籍的句子中实体密集，而且存在大量共用主语的情况，因此不进行长句切分，避免造成原本有关系的两个实体被划分到不同语句中的情况。利用远程监督与人工校对相结合的方式，筛选保留拥有两个以上实体的句子，获得实体之间的动词，对照动词表给句子打上标签，从而获得了按照（实体 A 实体 B 关系 句子）的方式标注好的训练数据集。本实验共标注了 8094 条样本数据，将其中的 70%作为训练集，30%作为测试集，即训练集为 5666 条，测试集为 2428 条，详细情况如下表所示：

表 5-1 关系类别标注表

标注序列	关系类别	具体内容	标注数量
0	unknown	-	1579
1	表征	在天为、在地为、在脏为、在窍为、在声为、在体为、其色、其音、其味、其类、其谷、其畜、其华在、其充为	876
2	概念	所谓、此谓、是谓、此为、谓之、为、曰、名曰、病名、命曰、名为、病名曰	683
3	促进	主、当、荣、胜、治、生、宜、藏、应、合、欲、入、归	1163
4	抑制	克、伤、恶、禁、出、死、无、治之以、方药-病症关系	932
5	因果	则、成为、发为、移热于、厥在、因于、病在、俞在、过在、出焉、生于、通于、并于、藏于、根于、藏精于、开窍于、入通于、传之于、内会于	1386
6	包含	有、属于、方剂-草药关系、	1475

此外对于预训练的字向量获取，我们沿用了 4.2.1 章节中获取的向量，利用 Word2vec 将数据预处理后的综合语料训练为 200 维的字向量，因为命名实体识别的对比实验证明，200 维的字向量对中医典籍中的字特征表征最好。

实体关系抽取实验的模型评价标准为精确率 P 和召回率 R 的综合权衡参数 F1 值，与命名实体识别实验中一样，因此这里不再赘述。

5.2.2 参数设置

在关系抽取模型 2Att-BiGRU 中，输入层引用了字向量和位置向量的拼接，字向量维度设置为 200 维，位置向量设置为 5。为了方便迭代，将 BiGRU 的隐单元数量设置为 200 个。将训练或测试期间每批实体对的数量设置为 123 个，同时为了防止过拟合，我们引入了 dropout 正则化机制，在 BiGRU 层的输入端与输出端都添加一层 dropout，并将 dropout 的概率设定为 0.5。此外，模型采用反向传播算法拟合训练数据，选用了常用的、性能较好的 Adam 梯度下降算法，学习率设置为 0.01，共训练 20 轮，并将批尺寸设置为 20。由于语料规模并不是十分大，考虑到刚开始模型效果肯定不佳，为了提高性能，因此我们设定为运行了 500 步以后，每 100 步输出一次模型。

表 5-2 关系抽取模型超参设置表

参数	值	参数	值
Character embedding size	200	Learn rate	0.01
Hidden units number	200	Batch_size	20
Position dimension	5	Max_epoch	20
Dropout rate	0.5	Steps_check	50
梯度下降算法	Adam	Steps_save	500

5.2.3 实验结果

为了得到最优模型，依次加载训练得到的所有模型，利用测试集求得每个模型的关系抽取效果，利用 F1 值表示，运行 5 次得到的平均值。

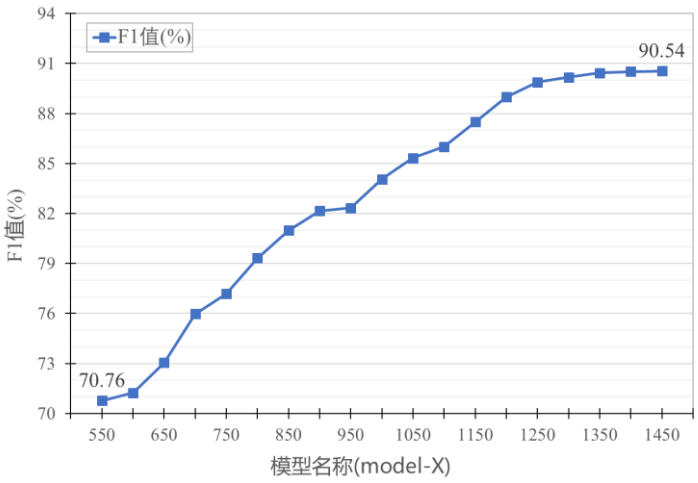

图 5-2 各模型性能测试结果

由实验结果可知，训练的最优模型为第 1450 步输出的模型，即最后一个模型—model_1450，F1 值达到了 90.54%。从整体曲线的走势可以看出，F1 值随着训练步数增加而增加，从 model_550 的 F1 值 70.76%开始上升了 19.78%，但是增速一直放缓，结果说明了即使是最优模型仍旧还未学到语料的所有特征，后续可以适当地调整参数，增加训练的次数。将最优的模型再次进行 5 次测试取平均值，求得对于每类关系的 F1 值进行分析。实验结果如下所示：模型的精确率为 91.69%，召回率为 89.28%，F1 值为 90.53%。

表 5-3 最优模型对各个关系类别的实验结果

标记序号	实体	Precision(%)	Recall(%)	F1(%)
	模型均值	91.69	89.28	90.53
0	unknown	88.36	85.37	86.36
1	表征	93.86	86.99	90.30
2	概念	92.35	88.05	90.70
3	促进	89.58	92.94	90.89
4	抑制	90.93	86.37	88.65
5	因果	92.47	90.81	91.64
6	包含	94.11	91.75	92.93

从实验结果可以看出，unknown 类关系的识别效果最差，F1 值为 86.36%，原因是该类关系涵盖的种类繁多，区分度不高，其中或许可能涵盖了别的种类关系，但是在标注时引入了误差。而表征关系和概念关系虽然标注的数量不多，但是效果尚佳，F1 值均在 90%以上，因为这两类关系的区分度十分明显，这也就导致了精准率明显高于召回率，模型的扩展性不好。而抑制关系虽然标注数量稍微多于表征和概念，但是 F1 值却较低，原因是该类动词的多义性引入了一部分噪声，譬如“勿禁”和“禁”是两类关系，但是筛选句子标注时，可能丢失了前半个重要信息。其中，包含关系 F1 值最高达到了 92.93%，因为包含和因果关系标注样本均较多，此外区分较明确，包含关系大部分为空间距离和所属类的关系，而因果关系大部分为病因、治法类的关系，实体类别明确，歧义性较少，因此关系分类也十分明确。

5.3 验证与对比

很多实验都验证了添加 Attention 机制的深度神经网络的效果最优。本章节对不同的权重求和方式、不同的模型、字与词向量的维度、数据集的规模、

组件参数效果等方面进行了对比实验，验证了本文提出的模型和使用参数在中医典籍自定义语义关系中的有效性。

5.3.1 不同模型效果对比

首先，验证句子级别 Attention 机制的有效性，我们利用带字级别 Attention 的 BiGRU 模型，选用了不同的注意力机制。第一种是 ONE 方式，即对于每个实体对，选取最有可能的一个句子进行训练和预测；第二种是 AVE 方式，即对每个句子向量求和，再取平均值，代表每个句子的权重相同；第三种就是本文模型中的 Attention 机制，赋予句子不同的权重。实验结果的 PR 图如下所示。

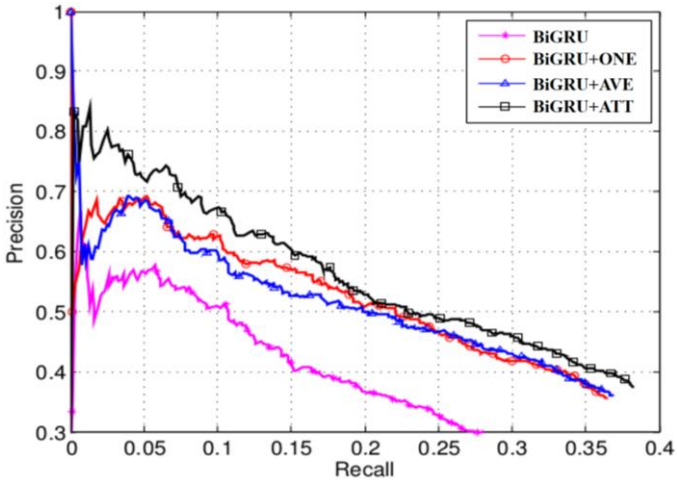

图 5-3　注意力机制效果对比图

由实验结果可知，引入注意力机制的效果都优于不加注意力的 BiGRU 模型。ONE 方式说明原数据集中包含很多噪声句子，整体的分类效果肯定不如只选择影响力最强的那个句子分类效果好。AVE 方式说明句子间信息的互补可以提高性能，就是说它虽然引入了噪声，但是同一实体对多个句子之间信息的一个相互补充对提高效果还是有帮助的。而 AVE 方式和 ONE 方式的效果很接近，说明这两种方法对噪声语料的处理都不是十分恰当。而 ATT 方式的效果最好，因为它对不同的句子赋予了不同的权重，可以有效降低噪声句子的影响，提高正样本句子的影响度，完全利用了语料，用到了多个句子的信息。

其次，验证模型的网络类型和 Attention 层数的影响，因为关系抽取本质上是分类问题，CNN 模型在分类问题上有特有的优势；而上文中我们验证了双向 LSTM 模型在原理和实践中，效果均优于 LSTM，GRU 是标准 LSTM 的

一种变体。这里，我们只比较 CNN 和在标注问题上表现更优异的 BiLSTM 的效果，同时对比了不同层级加入 Attention 的效果。

表 5-4 不同网络和注意力层次的模型效果对比

编号	模型	Attention	Precision(%)	Recall(%)	F1(%)
1	CNN	-	82.87	77.51	80.14
2		句子级	89.23	87.87	88.55
3	BiLSTM	-	78.36	73.10	75.72
4		字符级	90.93	86.37	88.64
5		句子级	88.94	85.78	87.36
6		字符级+句子级	92.35	88.77	90.56
7	BiGRU	-	78.24	72.74	75.49
8		字符级	89.60	86.36	87.98
9		句子级	88.53	86.91	87.72
10		字符级+句子级	91.69	89.28	90.53

对比 1、3、7 组实验，不加入 Attention 机制的情况下，CNN 的效果远远优于 RNN 结构的模型，这证明了 CNN 的模型结构确实对于分类问题有优势，而 BiLSTM 和 BiGRU 的效果相近，但是 GRU 作为标准 LSTM 的简易结构，在运行效率上远高于 LSTM。对比 2、5、9 组实验，均加入了句子级别的 Attention 机制，Att-CNN 的效果仍旧最好，但此时三个模型的效果已经非常接近了，BiLSTM 和 BiGRU 的提升效果非常明显，说明 LSTM 模型对 Attention 机制敏感。对比 4 和 5、8 和 9 两组实验，验证了针对 BiLSTM 和 BiGRU 均是字符级 Attention 的效果要优于句子级 Attention 机制。当然，6 和 10 两组实验证明了加入双层 Attention 机制的效果要高于单个 Attention 机制。从结果看，BiLSTM 模型效果均优于 BiGRU，但是两者效率十分接近，在 GRU 的运行效率更优的情况下，综合考虑我们选用 2Att-BiGRU 模型进行关系抽取。

5.3.2 数据规模效果对比

本实验是为了验证训练集、测试集的数据规模对分类效果的影响，从而验证模型的效率。语料选用《黄帝内经》标注好的 5000 多条训练集，规模从 500 条开始，每次增加 500 条，每次保留 2Att-BiGRU 模型训练后最后一个输出的模型为最优模型。测试集也使用了原有标注好的，随机选取 1 句、2 句以及 200 句进行测试，每组实验运行 3 次求平均值，获取最优模型的 F1 值。

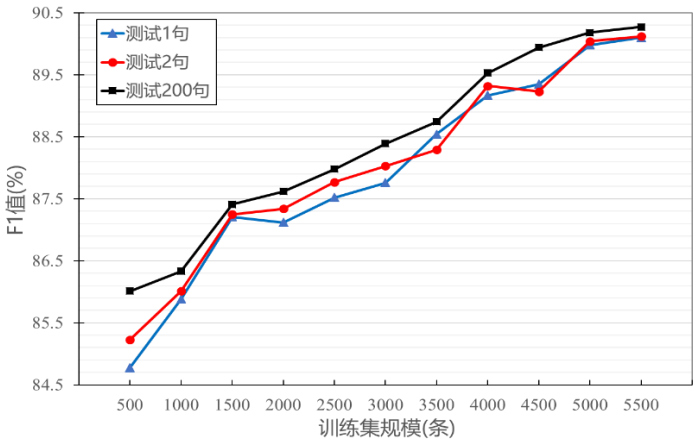

图 5-4　数据规模实验效果图

从三条曲线的整体走势可以看出，模型 F1 值随着训练集数量的增大而增大，说明训练集的数量越多越好。我们的样本量不充足，后续实验可以增大训练集的标注数量。而将测试集的选取数量可对比看出，测试 200 句的效果要优于随机选取 2 句和 1 句的情况，说明测试集的数量增多，也可以提升效果。结果中，选取 200 条句子的测试效果未达到前期实验的最优效果，说明实验的数据集偏小。结果中测试 1 句的某几个点效果要优于 2 句，但由于数据规模小，并且我们直接用的是最后一个输出模型，因此不排除特殊情况。

5.3.3 向量维度效果对比

在本实验中，我们对比字、词向量的选取以及不同字向量维度对关系抽取效果的影响。沿用第四章中获取的处理后的语料，利用添加了自定义词典的结巴分词将其进行分词。然后用 Word2vec 训练成相同维度的字和词向量作为输入，其他模型参数保持一致。以下数据均为运行 3 次求得的平均值。

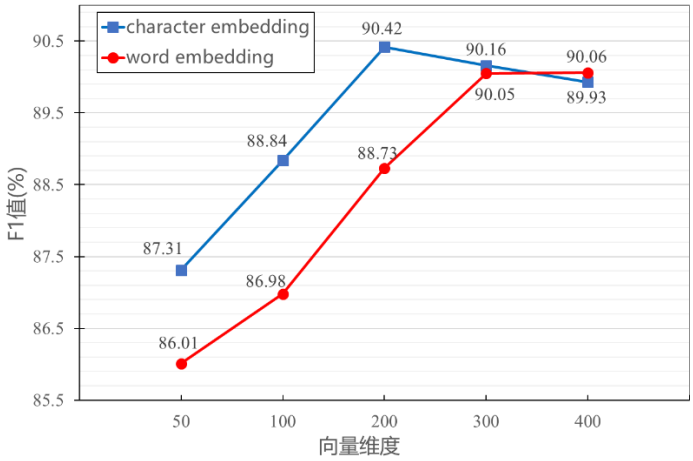

图 5-5　不同向量类型和维度效果对比

可以看出在相同维度下，使用字向量效果优于词向量。单从字向量来说，50 维到 200 维时，模型效果逐渐递增，200 维时对字的特征表征得最好，因此它的 F1 值最高，达到了 90.42%。但从 300 维开始效果下降。而词向量的效果随着维度增加一直在增加，在 400 维时基本达到了最优，但是词向量的最优效果仍旧差于字向量的效果，说明我们使用字向量作为输入具有有效性。鉴于此，在之后的实验中我们均选用 200 维的字向量。

5.3.4 不同参数效果对比

在关系抽取模型中，我们也使用了预训练初始化参数，引入 dropout 机制防止过拟合。每个 BiGRU 模型输入端添加了一个 L-dropout，在输出端添加了一个 R-dropout。本实验中，我们将验证预训练的优势，并且比较不同组件组合对关系抽取的效果影响。

表 5-5 关系抽取模型不同参数组合实验结果

embedding 组件	dropout 组件	Precision(%)	Recall(%)	F1(%)
Pretrain	L-dropout+R-dropout	91.78	89.28	90.53
	L-dropout	89.7	90.54	90.12
	R-dropout	90.89	88.97	89.93
	-	90.96	89.06	90.01
Random	L-dropout+R-dropout	91.27	88.41	89.84
	L-dropout	88.09	90.37	89.23
	R-dropout	88.37	90.45	89.41

由上表中结果可知，运用 pretrain 预训练的效果均优于随机生成映射矩阵的方式，因为使用 pretrain 方法初始化 embedding 层，初始化时字向量已包含语义信息，在模型训练时只需微调即可拟合分类任务。并且在使用 pretrain 的情况下，添加 L-dropout 和 R-dropout 的效果最优，F1 值达到了 90.53%。原因分析是，关系运用动词作为分类依据，并且实验语料仅为《黄帝内经》，相似句式太多、数据结构较为固定，模型容易发生过拟合，因此在 BiGRU 的输入层和输出层均加入 dropout 的方式效果更好。

5.4 本章小结

本章针对中医典籍的关系抽取任务，提出了中医典籍字向量和双重注意力机制、双向长短期记忆神经网络相结合的关系抽取模型(2Att-BiGRU)，并

且介绍了模型的结构和训练过程。该模型引入预训练的字向量和位置向量，利用字符级别的 Attention 机制赋予字向量不同的权重，拼接得到句子向量。利用句子级别的 Attention 机制赋予每个句子不同的权重，充分利用语料，去除噪声影响，得到句子集合的向量。最后将向量传入 softmax 分类器，进行关系分类。

此外，本章还对模型进行实验验证，利用第三章中获取的知识，对中医关系进行分类，并以《黄帝内经》为语料，抽取了其中蕴含的语义关系。并且将该模型与其它相关模型进行比较，同时实验对比了向量类型、向量维度、数据规模以及不同组件的不同效果，验证本工作模型的有效性，从而选取效果最好的模型结构。

6 中医典籍知识图谱自动构建方案设计

本章主要总结介绍完整的中医典籍知识图谱的自动构建方案流程，针对任意中医理论典籍，均能快速、自动地生成效果较好的中医知识粗图谱，其步骤主要包括知识获取、知识表示和知识可视化。其中，知识获取的流程可分为种子知识获取、实体识别和关系抽取三个模块，其详细内容已在第 3、4、5 章中进行了介绍。

6.1 方案总体架构

中医理论典籍知识图谱自动构建的完整方案主要可以细分为以下几个模块：先验知识获取模块、实体识别模块、关系抽取模块、知识表示模块、知识可视化模块，总体架构图如下所示。

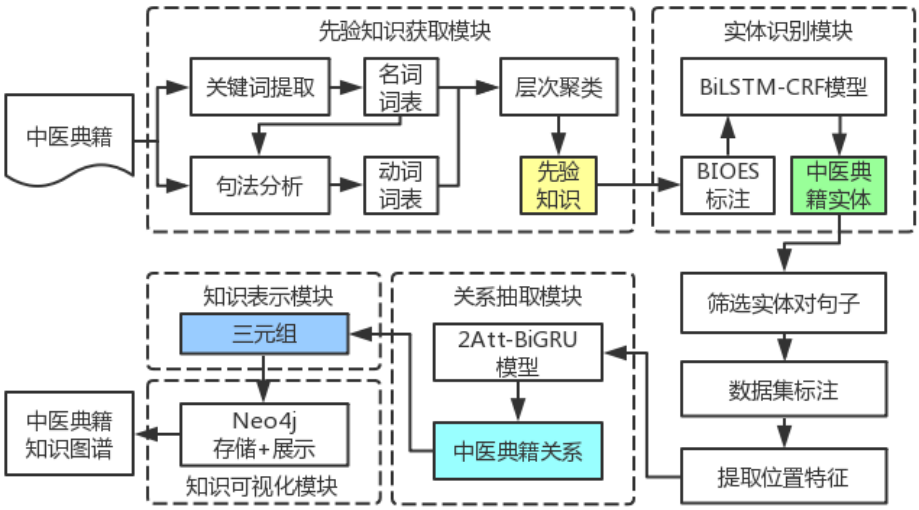

图 6-1　方案总体架构图

本方案采用了模块的串行处理方式，以中医理论典籍作为数据源，如经典的《黄帝内经》、《难经》、《神农本草经》等。先验知识获取模块首先提取关键词，构建领域基础词表；随后利用 Bootstrapping 的思想，不断扩充种子实体表和动词表，以依存句法分析方法得到以动词为核心的三元组，最后利用词向量进行层次聚类，得到拥有概念层次的种子实体和关系类别，进行特征提取。若语料类别相似，也可以直接跳过先验知识获取模块，共用已有的先验知识。

随后，进入实体识别模块，以具有层次概念和分类的先验知识为指导，进行实体类别的划分，以及采用 BIOES 标注方式对训练集自动打标签。利用基于 BiLSTM-CRF 的实体识别模型，自动获取中医典籍的命名实体。再利用

获得的命名实体集对语料中的句子进行筛选，保留拥有两个实体以上的句子，并根据先验知识对关系的分类进行训练集标注，基于双重注意力机制的双向长短期记忆神经网络 2Att-BiGRU 模型进行中医典籍的关系抽取，最终从中医典籍中得到中医知识数据三元组。

知识表示模块将先验知识获取模块得到的三元组与深度学习方法抽取出的三元组进行整合，统一用三元组形式进行知识表示。知识可视化模块中利用 Neo4j 图数据库来存储中医理论典籍的知识数据，最后对实体-关系进行可视化展示，完成知识图谱的构建。

6.2 知识表示

本文基于本体表示法，以(E1，R，E2)的三元组形式表示中医典籍的知识。因为这种知识表示方法具有结构性、联想性和自然性，能够把事务的属性和关系以自然语言的形式非常直观地展示出来，能够有效地表达医学知识，提高推理和检索的效率。因为，本体描述语言 RDF 和 RDFS 定义了标准化的数据结构，使得数据能够互联互通、存储、查询以及推理。

本方案已通过实体识别模块得到了 31084 个中医实体，每个实体均有一个类别标签，即中医认识方法、中医自然、中医生理、中医病理和治则治法 5 种，能够有效对具有多种含义的实体进行区分，譬如“水”有的属于中医自然(ZR)，有的属于中医认识方法(FF)。同时通过关系抽取模块获取了 24952 对实体间的关系，关系共有未知、表征、概念、促进、因果和包含 6 类。为了更直观且容易理解，最终我们将关系抽取模块获得的实体对和关系，以及先验知识获取模块获得的以动词为核心的先验知识均用三元组表示，则《黄帝内经》知识的部分表示如下表所示：

表 6-6 《黄帝内经》知识表示示例表

关系	三元组
unknown	(肺,unknown,寒饮食)、(软而散,unknown,灌汗) 、(形乐志乐,unknown, 针石)
表征	(东,表征,筋)、(南,表征,徵)、(西,表征,皮毛)、(北,表征,羽)、(中央,表征,脾)
概念	(肝,概念,五脏)、(弦,概念,五脉)、(胸胁,概念,病症)、(鼽衄,概念,病症)
促进	(金,促进,辛)、(辛,促进,肺)、(肺,促进,皮毛)、(燥,促进,风)、(北,促进,寒)
抑制	(怒,抑制,肝)、(风,抑制,筋)、(湿,抑制,肉)、(淫气,抑制,心)、(肝病,抑制,辛)
因果	(肺脉,因果,唾血)、(肾脉,因果,折腰)、(热气,因果,针热)、(真气去,因果,偏枯)
包含	(东风,包含,春)、(脾,包含,括蒌实)、(赤色,包含,心)、(中,包含,神机)

6.3 知识可视化

本文采用开源的图数据库 Neo4j 实现中医典籍领域知识的存储和可视化展示，具体过程主要包括：知识获取、知识导入和图形绘制，如图 6-2 所示。

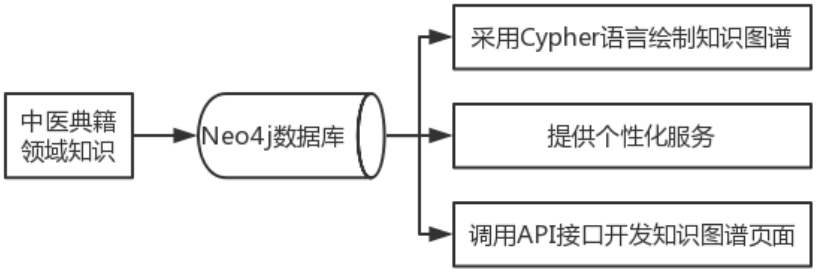

图 6-2 知识图谱绘制过程

上文中已完成了知识获取和知识表示，针对知识导入环节，Neo4j 支持通过 Cypher 语句手动创建、csv 或 excel 等文件自动导入等多种方式。依托实验室对 Neo4j 前期探索和累积的经验，这里选择利用 Java API 从 excel 中自动导入知识，创建节点和关系。

(1) 唯一性约束。知识获取得到的 31084 个实体中存在着大量重复的情况，为了确保同一个概念下的实体均唯一，我们首先对定义的 5 类中医典籍命名实体类别创建节点的唯一性约束：

Create Constraint On (f: FF) Assert f.name IS UNIQUE

Create Constraint On (s: SL) Assert s.name IS UNIQUE

Create Constraint On (z: ZR) Assert z.name IS UNIQUE

Create Constraint On (b: BL) Assert b.name IS UNIQUE

Create Constraint On (t: ZF) Assert t.name IS UNIQUE

(2) 创建节点。因节点数量庞大，本文将抽取出的实体经处理后保存在 excel 文件中，每个实体有三个属性：唯一的标识 id、实体的内容 name 以及实体所属的标签类别 label，例如“gn_1-五脏-概念”、“gn_23-五谷-概念”、“gn_5-五体-概念”、“gn_wg_1-麦-五谷”、“gn_wt_3-血脉-五体”、“gn_wt_xm_1-血-血脉”、“gn_wt_xm_2-脉-血脉”等，其含义是五脏、五谷、五体为概念，麦为五谷概念下的子概念，血脉为五体下的子概念，而血脉之下又可细分为实体血和实体脉。最后利用 Java API 按行循环读取 txt 文件中的数据，连接 Neo4j 数据库“hdnj.db”，id 和 name 存入节点的属性，label 设置为该节点的标签，生成实体节点。

(3) 创建关系。生成所有的实体节点之后，需要导入实体之间的关系。我们将实体间的关系保存在另一个 excel 中，以类似于三元组的形式对实体关

系进行分类，每一行代表一个三元组，共有至少 5 列：实体 A 的标签 labelA、实体 A 的内容 nameA、实体之间的关系 relation、实体 B 的标签 labelB、实体 B 的内容 nameB。例如“概念-五脏-概念-五脏-心”、“五方-中-表征-五体-肉”、“五志-怒-抑制-五脏-肝”、“五味-辛-促进-五味-酸”、“四时-冬-因果-六气-寒”。随后利用 Cypher 语言 create (a)-[r:" realtion+"{weight:1}]->(b) 创建实体 A 与实体 B 之间的关系，这里每个关系权重都为 1，并且不存在去重的效果，即一对实体之间可能存在多个关系，最终完成了知识在 Neo4j 中的存储。

(4) 图形绘制。接下来，我们利用 Neo4j 的可执行文件打开 hdnj.db，提供图形窗口。采用 Cypher 的查询语句 Match 就能将满足条件的领域实体和关系在图形界面上进行展示，如“match (n:概念) return n limit 5”即返回标签为概念的 5 个实体。使用“match n return n”获得中医典籍《黄帝内经》知识图谱的全貌，由于窗口限制，部分结果如下图所示。

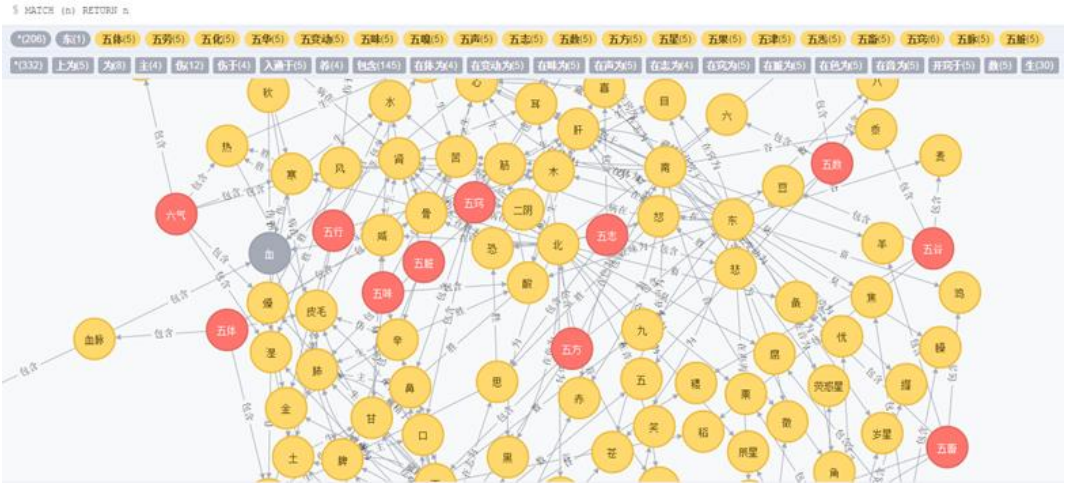

图 6-3 部分中医典籍知识图谱展示

事实上，领域知识图谱的构建过程十分漫长，经常伴随着新的实体关系的不断出现。针对这一情况，Neo4j 支持采用 Cypher 语句添加新的领域知识，实现对现有知识图谱的扩展。当中医典籍知识图谱构建完成后，利用 Neo4j 数据库可以对知识图谱进行搜索，提供个性化的服务。

6.4 本章小结

本章主要提出并介绍了中医理论典籍知识图谱构建方案的整体框架与流程，详细介绍了以三元组为核心的知识表示模块及其特点。同时介绍了利用 Neo4j 图数据库进行知识存储，并最终进行可视化展示的流程，实现了中医典籍知识图谱的自动构建。

7 总结与展望

本文首先调研了目前知识图谱，特别是中医领域知识图谱构建的在国内外研究现状，以及构建中医典籍知识图谱的必要性和先进性。随后介绍了构建知识图谱的关键任务和相关技术，深入研究中医典籍语言特点，明确研究重点、难点和实现技术原理上的适用性。针对深度学习循环神经网络的结构特点，提出了一种中医典籍知识的快速自动获取方案，首先结合无监督学习快速获取的先验知识进行特征构建的方法，其次基于循环神经网络模型进行实体和关系串行抽取，并进行了大量的对比分析实验，验证模型方法的有效性。最后设计总结了完整的中医典籍知识图谱自动化构建方案，以《黄帝内经》为例，将抽取出的知识以三元组的形式进行表示，并利用 Neo4j 图数据库实现了知识的存储和可视化。

本文的主要贡献如下：

(1) 深入分析中医典籍的语言特色，确定中医基础理论体系。提出了一种中医典籍先验知识快速获取方法，基于无监督学习方法和迭代思想构建并领域词表，结合词向量的层次聚类和自然语言处理技术获取层次概念和先验知识，从而进行提取特征，规范中医实体的类别和实体之间的关系类别，减少人工干预和对专业知识的依赖，同时减少后续深度学习标注工作量。

(2) 分析了当前主流的深度学习模型特点，创新性地将中医典籍字向量训练与深度学习相结合，提出基于双向长短期神经网络串行进行实体抽取与关系抽取的知识获取方法，减少了分词累计的效果误差，明显提高对中医典籍中多实体和关系的识别准确率，在一定程度上解决了语料不足的限制。

(3) 采用不同的深度学习模型处理中医典籍实体识别和关系抽取的任务，利用 BiLSTM-CRF 模型进行实体识别，重复利用中医典籍句子的上下文信息，同时结合 CRF 添加一定的规则，提升了识别准确率；尝试性地将 2Att-BiGRU 模型处理关系抽取任务，大大降低了噪声数据的影响。同时针对模型的结构、参数等进行了大量的对比分析实验，验证了该模型的有效性，为不同任务匹配了恰当的模型和方法。

(4) 提出了一种中医典籍知识图谱自动化构建的完整方案，明确知识获取、知识表示和知识存储及可视化模块的实现方法，实现《黄帝内经》复杂全面的粗图谱构建，补充了中医药大数据知识图谱在中医典籍方面的空缺，为中医智能诊疗提供服务。

本文的下一步工作计划如下：

(1) 经大量实验验证，深度神经网络加入 Attention 机制在各领域均有效

果的提升，因此下一步可以尝试在命名实体识别的 BiLSTM-CRF 模型中添加字级别的 Attention 机制，用于提升模型的识别效果。

(2) 本文根据中医语料特点将模型的输入以字符为单位，验证了字向量对特征有更好的表征。杜琬晴[72] 等人根据字的古意字形，将字进一步拆分为部首，譬如“朝”拆分为“十、日、月”训练成向量用于表示特征，效果有进一步提升。该思想原理上很适用于表征中医典籍中的字，因此下一步尝试将输入改为更细粒度的部首向量。

(3) 针对实体和关系的分类，本文目前主要是根据词向量聚类和领域经验知识相结合的方法进行划分，类别较少实体 5 类，关系 6 类。但是中医典籍的实体与关系类型相当复杂，特别是中医典籍的关系，少有明确表示的层次和语义关系，更多的是隐晦表达的关系，没有谓词连接。因此后续需要对实体与关系种类的进一步拓展，将其进行更合理的分类，从而增加知识图谱的信息多样性与领域知识完备性。

(4) 目前本文该方法只针对中医典籍，但是典籍的篇幅较短，数量也有限，后期可以尝试将此方法迁移至对中医古医案、现代医案的处理，验证方法的可移植性和通用性。

(5) 本文的数据规模仍旧较小，制约了模型的识别效果，后续需要扩充训练数据的标注样本量。并且为了解决一字多义性，如“水”可能属于中医认识方法中的五行，也可能属于中医自然，需要下一步需要将识别出的实体进行实体对齐，消除影响；此外还可利用知识推理扩充实体间的关系。

(6) 由于知识图谱主要采用图数据库进行存储，在受益于图数据库带来的查询效率的同时，也失去了关系型数据库的优点，如 SQL 语言支持和集合查询效率等。在查询方面，如何处理自然语言查询，对其进行分析推理，翻译成知识图谱可理解的查询表达式以及等价表达式等也都是知识图谱应用需解决的关键问题。

参考文献

[1] 袁锋. 中医医案文本挖掘的若干关键技术研究[D]. 山东师范大学, 2016.

[2] 何裕民. 中医学导论[M]. 北京：中国协和医科大学出版社，2004: 1-6, 84-86, 49-55.

[3] 王鑫, 王丁, 李向宏. 基于汉语分词的信息抽取技术[J]. 信息技术, 2003(04): 101-104.

[4] Singhal A. Introducing the Knowledge Graph: Things, Not Strings[J]. Offical Google Blog, 2012.

[5] Auer S，Bizer C，Kobilarov G，et al. DBpedia: A nucleus for a Web of open data[C]. Proc of thr 6th Int Semantic Web and 2nd Asian Conf on Asian Semantic Web Conf. Piscataway, NJ：IEEE, 2007: 722-735.

[6] Bollacker K, Cook R, Tufts P. Freebase: A Shared Database of Structured General Human Knowledge[C]. AAAI Conference on Artificial Intelligence, Vancouver, British Columbia, Canada. DBLP, 2007: 1962-1963.

[7] Amarilli A, Galárraga L, Preda N, et al. Recent Topics of Research around the YAGO Knowledge Base[M]. Web Technologies and Applications. Springer International Publishing, 2014: 1-12.

[8] Philpot A, Hovy E, Pantel P. The Omega Ontology[J]. Prep, 2005: 59-66.

[9] Ponzetto S P, Navigli R. Large-scale taxonomy mapping for restructuring and integrating wikipedia[C]. International Jont Conference on Artifical Intelligence. Morgan Kaufmann Publishers Inc, 2009: 2083-2088.

[10] Etzioni O, Cafarella M, Downey D, et al. Web-scale information extraction in knowitall:(preliminary results)[C]. 2004: 100-110.

[11] Banko M, Cafarella M J, Soderland S, et al. Open information extraction from the web[C]. International Joint Conference on Artifical Intelligence. Morgan Kaufmann Publishers Inc, 2007: 2670-2676.

[12] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning[C]. Twenty-Fourth AAAI Conference on Artificial Intelligence. AAAI Press, 2010: 1306-1313.

[13] Wu W, Li H, Wang H, et al. Probase:a probabilistic taxonomy for text understanding[C]. ACM, 2012: 481-492.

[14] CHENG Xueqi, JIN Xiaolong, WANG Yuanzhuo, et al. Survey on Big Data System and Analytic Technology[J]. Journal of Software, 2014.

[15] 李兵, 裘俭, 张华敏. 中医药领域本体研究概述[J]. 中国中医药信息杂

志, 2010, 17(3): 100-101.

[16] 阮彤, 孙程琳, 王昊奋, 等. 中医药知识图谱构建与应用[J]. 医学信息学杂志, 2016, 37(4): 8-13.

[17] 于彤, 刘静, 贾李蓉, 等. 大型中医药知识图谱构建研究[J]. 中国数字医学, 2015(3): 80-82.

[18] Li Y, Zhang M, Du K, et al. TCM Ontology and Brain Diseases[J]. World Science and Technology-Modernization of Traditional Chinese Medicine and Materia Medica, 2007.

[19] 张德政, 谢永红, 李曼, 石川. 基于本体的中医知识图谱构建[J]. 情报工程, 2017, 3(01): 35-42.

[20] 郝伟学. 中医健康知识图谱的构建研究[D]. 北京交通大学, 2017.

[21] Sundheim B, Sundheim B. Message Understanding Conference-6: a brief history[C]// Conference on Computational Linguistics. Association for Computational Linguistics, 1996: 466-471.

[22] 胡双, 陆涛, 胡建华. 文本挖掘技术在药物研究中的应用[J]. 医学信息学杂志, 2013, 34(8): 49.

[23] 范岩. 基于条件随机场模型的中医文献知识发现方法研究[D]. 北京交通大学, 2009.

[24] Fukuda K, Tsunoda T, Tamura A，et al. Towards information extraction：identifying protein names from biological papers[C]. Proc. Pacific Symposium on Biocomputing, 1998, 3.707-718.

[25] Bikel D M, Miller S, Schwartz R, et al. Nymble: a High-Performance Learning Name-finder[J]. Anlp, 1998: 194-201.

[26] Mccallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]. Conference on Natural Language Learning at Hlt-Naacl. Association for Computational Linguistics, 2003: 188-191.

[27] Asahara M, Matsumoto Y. Japanese Named Entity extraction with redundant morphological analysis[C]. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, 2003: 8-15.

[28] Okanohara D, Miyao Y, Tsuruoka Y, et al. Improving the scalability of semi-Markov conditional random fields for named entity recognition[C]. International Conference on Computational Linguistics. Association for Computational Linguistics, 2006: 465-472.

[29] 王世昆, 李绍滋，陈彤生. 基于条件随机场的中医命名实体识别[J]. 厦门大学学报(自然版), 2009, 26(3):359-364.

[30] 张五辈, 白宇, 王裴岩, 等. 一种中医名词术语自动抽取方法[J]. 沈阳航空航天大学学报, 2011, 28(1): 72-75.

[31] 孟洪宇, 谢晴宇, 常虹, 等. 基于条件随机场的《伤寒论》中医术语自动识别[J]. 北京中医药大学学报, 2015, 38(9):587-590.

[32] Hiton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 2006, 313(5786)：504-507.

[33] Wu Y, Jiang M, Lei J, et al. Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network[J]. Stud Health Technol Inform, 2015, 216: 624-628.

[34] Zhiheng Huang, Wei X, Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv, 2015, 1508.01991.

[35] 张帆, 王敏. 基于深度学习的医疗命名实体识别[J]. 计算技术与自动化，2017, 36(1): 123-127.

[36] 薛天竹. 面向医疗领域的中文命名实体识别[D]. 哈尔滨工业大学, 2017.

[37] 步君昭. 生物医学文献中的药物名抽取方法研究[D]. 哈尔滨工业大学, 2016.

[38] Sumida A, Torisawa K. Hacking Wikipedia for Hyponyiny Relation Acquisition[C], IJCNLP, 2008, 8：883-888.

[39] 刘磊, 曹存根, 王海涛. 一种基于"是一个"模式的下位概念获取方法[J]. 计算机科学, 2006, 33(9): 146-151.

[40] Fang N M, Non-Member C Y, Member F R. Hyponym extraction from the web by bootstrapping[J]. IEEJ Transactions on Electrical & Electronic Engineering, 2011, 7(1): 62-68.

[41] 雷春雅, 郭剑毅, 余正涛, 等. 基于自扩展与最大熵的领域实体关系自动抽取[J]. 山东大学学报(工学版), 2010, 40(5): 141-145.

[42] Guodong Z, Jian S, Jie Z, et al. Exploring Various Knowledge in Relation Extraction.[C]. Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2002: 419-444.

[43] 董静, 孙乐, 冯元勇,等. 中文实体关系抽取中的特征选择研究[J]. 中文信息学报, 2007, 21(4): 80-91.

[44] 刘克彬, 李芳, 刘磊,等. 基于核函数中文关系自动抽取系统的实现[J]. 计算机研究与发展, 2007, 44(8): 1406-1411.

[45] Zeng D, Liu K, LaiS, et al. Relation classification via convolutional deep neural network[C]. Proceedings of COLING, 2014: 2335-2344.

[46] 陈宇, 郑德权, 赵铁军. 基于 Deep Belief Nets 的中文名实体关系抽取[J]. 软件学报, 2012, 23(10): 2572-2585.

[47] Liu C Y, Sun W B, Chao W H, et al. Convolution Neural Network for Relation Extraction[C], International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, 2013: 231-242.

[48] 曾东火. 基于深度学习的药名实体关系抽取[D]. 哈尔滨工业大学, 2017.

[49] 冯钦林. 基于半监督和深度学习的生物实体关系抽取[D]. 大连理工大学, 2016.

[50] 蒋振超. 基于词表示和深度学习的生物医学关系抽取[D]. 大连理工大学, 2016.

[51] 郑洁琼. 生物医学文本中实体关系抽取的研究[D]. 大连理工大学, 2017.

[52] 杨晨浩. 基于深度学习的中文电子病历实体修饰与关系抽取研究及算法平台开发[D]. 哈尔滨工业大学, 2016.

[53] Lafferty J D, Mccallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc, 2001:282-289.

[54] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011.

[55] Shen Yelong, He Xiaodong, et al. Learning semantic representations using convolutional neural networks for Web search[C]. Proc of the 23rd International Conference on World Wide Web. New York: ACM Press, 2014.

[56] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[C]. Proc of the 52nd Annual Meeting of the Association for Computational, 2014: 655665.

[57] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[58] Fisher Yu, Vladlen Koltun. Multi-scale context aggregation by dilated convolution, 2016.

[59] N Kalchbrenner, L Espeholt, K Simonyan, et al. Neural Machine Translation in Linear Time, 2016.

[60] E Strubell, P Verga, D Belanger, et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions, 2017.

[61] Mikolov T, Kombrink S, Burget L, et al. Extensions of recurrent neural network language model[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2011:5528-5531.

[62] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[63] Rush A M, Chopra S, Weston J. A Neural Attention Model for Abstractive

Sentence Summarization[J]. arXiv, 2015, 1509.G0685v2.

[64] 张冲. 基于 Attention-Based LSTM 模型的文本分类技术的研究[D]. 南京大学, 2016.

[65] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[J]. 2016:260-270.

[66] Jagannatha A, Yu Hong. Structured prediction models for RNN based sequence labeling in clinical text[C], Proc of Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 856-865.

[67] 李崇超. 论中医知识体系中的分类[J]. 中医杂志, 2015, 56(21):1804-1807.

[68] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.

[69] Strubell E, Verga P, Belanger D, et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions[J]. 2017.

[70] Zhou P, Shi W, Tian J, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification[C]. Meeting of the Association for Computational Linguistics, 2016:207-212.

[71] Lin Y, Shen S, Liu Z, et al. Neural Relation Extraction with Selective Attention over Instances[C]. Meeting of the Association for Computational Linguistics, 2016:2124-2133.

[72] 杜琬晴, 闫家馨, 王一, 等.古汉字构形释义法——理解《黄帝内经》术语的新思路[J]. 北京中医药大学学报, 2017, 40(08): 626-629.

附录A

1) 《黄帝内经》名词词表

中医认识方法：五脏、五华、五荣、五充、五合、五体、五方、五色、五窍、五味、五行、五畜、五谷、五星、五音、五数、五臭、五液、五声、五变动、五志、五菜、五果、五劳、五脉、五气、五性、五德、五用、五化、五令、五虫、五政、五变、五眚、五时、木、火、土、金、水、天干、甲乙、丙丁、戊己、庚辛、壬癸、五炁、五神、五实、五虚、五卫、六腑、六气、八风、三阴三阳、十二官、十二经、太阳、少阳、阳明、厥阴、少阴、太阴、十二经脉、十二皮部、奇经八脉、十五络穴、阴、阳、十三方剂、九针、九刺、九变、十二刺、十二经、五刺、十二经水、君主之官、相傅之官、将军之官、中正之官、臣使之官、食廩之官、传道之官、受盛之官、作强之官、决渎之官、州都之官、阴阳、五夺、五逆、五变、五过、三失、九变、四过、三常、三失、五禁、五宜、五运、五诊、五别、五决、八虚、十一焦、十四椎、十二节、十二从、十二经络脉、十二痈、十二邪、十二脉、十二脏、十二分、十二俞、十二疟、十二部、十二藏、十二原、二阳、二阴、二脏、二穴、二火、二痈、二之气、二十五人、二十五输、二十五阳、二十五俞、二十八宿、二十八脉、二十八会、二十八星、三脏、三品、三水、三痈、三经、三焦、三候、三椎、三针、三部、三之气、三气之纪、三十六输、三百六十五节气、三百六十五节、三百六十五穴、三百六十五脉、三百六十五络、十五痈、十五络、五十营、五十七穴、五十九俞、五十九痈、四时、四气、四脏、四藏、四季、四淫、四经、四街、四极、四椎、四傍、四肢、四支、四之气、四难、四塞、四德、四关、四海、四厥、四野、五阳、五藏、五风、五经、五焦、五痹、五形志、五肤俞、五之气、五丸、五趾、五络、五位、五类、五中、五火、五官、五阅、五使、五输、五俞、五理、六府、六节、六经、六寸、六元、六椎、六经脉、六律、六俞、六之气、六分、六纪、六合、六痈、六脉、七损、七诊、七窍、七椎、七星、七疝、七节、七焦、八远、八风、八益、八节、八纪、八溪、八正、八俞、八痈、九州、九窍、九脏、九候、九气、九节、九野、九焦、九分、九星、九谊、九宫

中医生理：肝、心、脾、肺、肾、爪、面、色、唇四白、唇、毛、发、筋、血脉、脉、血、肌、肉、肌肉、皮、皮毛、骨、髓、骨髓、目、耳、口、鼻、二阴、舌、泪、泣、涕、汗、唾、涎、液、呼、笑、歌、哭、呻、握、忧、哕、咳、栗、怒、喜、忧、思、恐、行、视、坐、卧、立、弦、石、钩、

代、毛、魂、神、志、魄、精、意、胆、小肠、胃、大肠、膀胱、三焦、足太阳、足少阳、足阳明、足厥阴、足少阴、足太阴、手太阳、手少阳、手阳明、手厥阴、手少阴、手太阴、肺手太阴、大肠手阳明、胃足阳明、脾足太阴、心手少阴、小肠手太阳、膀胱足太阳、肾足少阴、三焦手少阳、胆足少阳、肝足厥阴、心主手厥阴心包络、关枢、枢持、害蜚、枢儒、关蛰、害肩、督脉、任脉、冲脉、带脉、阴维、阳维、阴蹻、阳蹻、脉络、居阴之脉、同阴之脉、衡络之脉、会阴之脉、飞阳之脉、昌阳之脉、肉里之脉、散脉、解脉、心主、经络、经俞、孙络、直肠之脉、孙脉、络脉、经脉、列缺、偏厉、丰隆、公孙、通里、飞扬、大钟、内关、外关、光明、鸠尾、长强、大包、支正、蠡沟、穴位、天突、人迎、扶突、天窗、天容、天牖、天柱、风府、天府、天池、大迎、命门、井穴、荥穴、输穴、俞穴、经穴、合穴、原穴、少商、鱼际、太渊、经渠、尺泽、太渊、商阳、二间、三间、阳溪、阳谿、曲池、合谷、厉兑、内庭、陷谷、解溪、解谿、足三里、冲阳、隐白、大都、太白、商丘、阴陵泉、太白、少冲、少府、神门、灵道、少海、神门、少泽、前谷、后溪、后谿、阳谷、小海、腕谷、至阴、足通谷、束骨、昆仑、委中、京骨、涌泉、然谷、太溪、太谿、复溜、阴谷、太溪、中冲、劳宫、大陵、间使、曲泽、大陵、关冲、液门、中渚、支沟、天井、阳池、窍阴、侠溪、侠谿、足临泣、阳辅、阳陵泉、丘墟、大敦、行间、太冲、中封、曲泉、太冲、胸中、膈中、女子胞、脑、气口、寸口、心中、跟中、散俞、络俞、季胁、腹中、鬲、膻中、踹内、腨内、手臂、头首、下牙车、颧后、颊上、眉上、腰中、目下、鼻上、肌上、颈项、胸胁、胸肋、肩背、背、腨外、腰股、脊、头、足、舌本、鼻之交頞中、腘中、鼻外、上齿中、耳前、心系、小指之端、手外侧、踝中、肘内侧两筋之间、臂骨下廉、膑外后廉、髀枢、髀外、小趾之下、发际、额颅、喉咙、缺盆、膈、脐、胫外廉、足心、目锐眦、毛中、阴器、小腹、肘、肩髃、臂、臀、尻、额角、颃、腕、膑外廉、肘外廉、膑外后廉、肩解、肩胛、膂骨、循膂、胛、血气、精气、脉气、经气、心气、肝气、脾气、肺气、肾气、胃气、天葵、月事、真牙、发鬓

中医自然：东方、南方、中央、西方、北方、青色、苍、赤色、赤、黄色、黄、白色、白、黑色、黑、酸、苦、甘、辛、咸、鸡、羊、牛、马、彘、犬、猪、麦、黍、黄黍、稷、稻、豆、大豆、麻、糠米、岁星、荧惑星、镇星、太白星、辰星、角、徵、征、宫、商、羽、八、七、五、九、六、臊、焦、香、腥、腐、韭、葱、薤、藿、葵、李、杏、枣、桃、栗、宣发、郁蒸、云雨、雾露、霰雪、摧拉、炎烁、动注、肃杀、凝冽、为陨、燔(火芮)、淫溃、苍落、冰雹 、春、夏、长夏、季夏、秋、冬、生、长、化、收、藏、大弱风、

谋风、刚风、折风、大刚风、凶风、婴兀风、弱风、风、热、暑、湿、燥、寒、火、渭水、海水、湖水、汝水、浑水、淮水、漯水、江水、河水、济水、漳水、平旦、日中、黄昏、合夜、鸡鸣、日西、春三月、夏三月、秋三月、冬三月、春分、夏分、秋分、冬分、东风、西风、南风、北风、雷气、春气、夏气、秋气、冬气、土气、地气、阳气、天气、真气、食气、淫气、浊气、阴气、湿气、风气、寒气、厥气、热气、燥气、清气、正气、邪气、人气、雨气、天地、日月、星辰、内外、表里、真人、圣人、至人、贤人

中医病理：脉盛、皮热、腹胀、前后不通、闷瞀、脉细、皮寒、气少、泄利前后、饮食不入、脉沉而横、脉浮而盛、脉中手长、脉小实而坚、尺热、尺寒脉细、尺脉缓涩、浮大而短、洪大以长、乍短乍长、乍数乍疏、中外急、坚而搏、尺涩脉滑、大而虚、搏而绝、脉沉、脉浮、脉悬绝、滑大、臂多青脉、脉急、脉缓、脉小、脉大、脉滑、脉涩、气上、气缓、气消、气下、气收、气泄、气乱、气耗、气结、夜行、堕恐、惊恐、渡水跌仆、饮食饱甚、惊而夺精、持重远行、疾走恐惧、摇体劳苦、饱食、大饮、强力、阳盛、阳胜、阴胜、阴盛、溢阴、溢阳、寒极、热极、格阳、关阴、息积、胎病、胃疸、黄疸、瘈、瘈瘲、妊子、暴痛、癫狂、狂、胀、脑烁、败疵、赤施、兔啮、走缓、四淫、噫、咳、语、吞、欠、噫、气逆、哕、恐、心掣、水、癃、伏梁、蛊、肠澼、大瘕、腹满、胁痛、肭髓病、肺消、涌水、惊衄、鬲消、柔痓、血菀、血枯、癃溺血、口糜、虚瘕、食亦、鼓胀、濡泻、骨癫、洞泄、实寒变、腨(疒肩)　、索泽、痿厥、痿躄、脉痿、筋痿、肉痿、骨痿、阳厥、暴厥、厥狂、厥逆、痹厥、风厥、热厥、寒厥、薄厥、痈疡、痈肿、猛疸、夭疸、疵痈、米疸、井疸、甘疸、股胫疸、锐疸、厉痈　、肺痹、肾痹、肝痹、骨痹、行痹、食痹、喉痹、痛痹、着痹、筋痹、脉痹、肌痹、皮痹、周痹、心痹、脾痹、肠痹、胞痹、挛痹、胆瘅、热中、胸中热、膈中热、肝热、脾热、肾热、腹中热、少腹热、痹热、心热、肺热、足下热、疠风、肝风、心风、脾风、肺风、肾风、偏风、胸风、目风、漏风、内风、首风、肠风、泄风、酒风、大风、寒疟、温疟、瘅疟、肺疟、心疟、肝疟、脾疟、脾疟、肾疟、胃疟、风疟、痎疟、厥疝、心疝、疝瘕、(疒颓)疝、狐疝　、皮槁、脉凝泣、爪枯、筋急、头痛巅疾、下虚上实、狂笑、四支解堕、喘息、忧思、遗弱、飧泄、乏竭、心痹引背、妄言骂詈、头项痛、腰脊强、身热目痛、鼻干、不得卧、血溢、胸胁痛、腹满、口燥舌干、烦满、囊缩、烦闷善呕、呕血、发咳上气、胁痛出食、肌绝、喉吩、筋脉相引、血凝泣、鼽衄、目冥、支膈胠胁、心烦头痛、咳嗽上气、消气、夺血、口干、咳嗽、解堕、善梦、欲卧不能眠、善渴、膑痛、眠而有见、时欲怒、耳聋、心中欲无言、百节皆纵、

目寰绝系、善呻、妄言、腹胀闭、善噫善呕、皮毛焦、善溺、舌卷、唇胗、中热溢干、气急、气胀、气衰、气少、心痛、颈肿、舌卷不能言、消环自己、唾血、灌汗、溢饮、折髀、膝腨肿痛、寒热、消中、巅疾、心痛引背、疬、泄、大腹水肿、骨痛、头痛、足胫痛、肩脊痛、疝瘕、少腹痛、胃脘痛、脱血、解(亻亦)、多汗、后泄、溺黄赤、面肿、溏瘕泄、足胫肿、目黄、善忘、少血、胸痛引背、两胁胠满、便血、气淖泽、肺虚、气热脉满、足寒、经虚络满、头痛耳鸣、腠理闭、胸胁满、胸胁支满、目下肿、月事不来、心腹满、少腹盛、痒搔、九窍不通、烦心、咳唾、气泄、浸淫、心悬、四支不举

治则治法：振埃、发蒙、去爪、彻衣、解惑、汤液醪醴、生铁洛饮、左角发酒、马膏膏法、泽泻饮、豕膏、小金丹、半夏秫米汤、兰草汤、鸡矢醴、翘饮、寒痹熨法、乌鲗骨藘茹丸、砭石、按蹻、导引按蹻、毒药、灸焫、微针、汤液、灸刺、针石、熨引、百药、按摩醪药、镵针、员针、提针、锋针、铍针、员利针、毫针、长针、大针、巾针、絮针、厘针、綦针、锋针、俞刺、远道刺、经刺、络刺、分刺、大泻刺、毛刺、巨刺、焠刺、偶刺、报刺、恢刺、齐刺、扬刺、直针刺、输针、短刺、浮刺、阴刺、傍针刺、赞刺、半刺、豹文刺、关刺、合谷刺、输刺

2) 《黄帝内经》动词词表

主、伤、食、当、刺、荣、走、胜、恶、出、应、治、藏、归、入、宜、候、禁、合、为、则、生、欲、曰、络于、生于、通于、发于、出于、客于、结于、注于、伤于、在于、属于、并于、入于、藏于、根于、溜于、因于、病在、俞在、过在、厥在、所谓、是谓、此谓、此为、谓之、发为、名曰、为上、成为、当病、病名、出焉、则梦、欲如、命曰、名为、其色、其音、其虫、其令、其变、其味、其类、其畜、其谷、其臭、其数、入通于、藏精于、传之于、合入于、内会于、开窍于、受气于、入舍于、禀气于、移热于、移寒于、治之以、病名曰、其华在、在色为、在音为、在志为、在脏为、在声为、在体为、在气为、其性为、其用为、其化为、在天为、在窍为、在地为、其充在、其政为、其志为、其德为、其眚为、其色为、在变动为

3) 《黄帝内经》三元组

概念：（水，为，阴）、（火，为，阳）、（脉涩，曰，痹）、（胃气，曰，逆）、（手阳明，名曰，偏历）、（手少阳，名曰，外关）、（太阴，为，埃溽）、（少阳，为，炎暑）、（阳明，为，清劲）、（太阳，为，寒氛）、（阳明，为，司杀府）、（太阳，为，寒府）、（厥阴，为，生化）、（太阴，为，濡化）、（少阳，为，茂化）、（阳明，为，坚化）、（冬，为，飧泄）、（春，为，痿厥）、（寒薄，为，皴）、（陷脉，为，瘘）、（中央，为，土）（脏，为，阴）、（腑，为，阳）、

（背，为，阳）、（腹，为，阴）、（水，为，阴）、（火，为，阳）、（阳，为，气）、（阴，为，为）、（六经，为，川）、（肠胃，为，海）、（太阳，为，开）、（阳明，为，阖）、（久风，为，飧泄）、（脾，为，吞）、（胃，为，气逆）、（胆，为，怒）、（肝，为，泪）、（黄赤，为，热）、（青黑，为，痛）、（胃，为，仓廪官）、（脾，为，谏议官）、（肾，为，作强官）、（肝，为，将军官）、（阴阳不应，病名曰，关格）、（汗出而烦满，病名曰，风厥）、（少腹盛，病名曰，伏梁）、（肝，者也，魂之居）、（心，者也，神之变）

表征：（东方，在体为，筋）、（南方，在体为，脉）、（中央，在体为，肉）、（西方，在体为，皮毛）、（北方，在体为，骨）、（东方，在音为，角）、（南方，在音为，徵）、（中央，在音为，宫）、（西方，在音为，商）、（北方，在音为，羽）、（东方，在天为，玄）、（南方，在天为，热）、（中央，在天为，湿）、（西方，在天为，燥）、（东方，其政为，散）、（南方，其政为，明）、（中央，其政为，谧）、（西方，其政为，劲）、（东方，其眚，陨）、（南方，其变，炎烁）、（中央，其性，静兼）、（西方，其令，雾露）、（北方，其眚，冰雹）、（神，在地为，木）、（神，在脏为，肝）、（心，荣，色）、（肺，荣，毛）、（肝，其华在，爪）、（肺，其充在，皮）、（肝，其味，酸）、（肝，其色，苍）、（肺，其华在，毛）、（肾，其华在，发）、（肾，其充在，骨）

促进：（金，生，辛）、（辛，生，肺）、（肺，生，皮毛）、（皮毛，生，肾）、（北，生，寒）、（悲，胜，怒）、（燥，胜，风）、（肾，主，水）、（肺，主，鼻）、（春，养，阳）、（夏，养，阴）、（春，胜，长夏）、（长夏，胜，冬）、（甘，走，肉）、（浊阴，走，五脏）、（厥气，走，喉）、（辛，走，气）、（酸，走，筋）、（鼻，知，臭）、（舌，知，五味）、（口，知，五谷）、（左足，应，立春）、（左胁，应，春分）、（左手，应，立夏）、（秋，为，痎疟）、（暑燥，生，寒）、（心，生，血）、（肾，生，骨髓）、（脾，藏，肉）、（肝，藏，血）、（恐，胜，喜）、（寒，胜，热）、（咸，胜，苦）、（道，生，智）、（玄，生，神）、（肺，主，鼻）、（肾，主，耳）、（心，主，肾）、（肝，主，肺）、（心，合，脉）、（肾，合，骨）、（脾，合，肉）、（黄色，宜，甘）、（青色，宜，酸）、（白色，宜，辛）、（脾病，宜，糠米饭）、（心病，宜，羊肉）、（肾病，宜，大豆）、（肺病，宜，黄黍）、（心，主，噫）、（肺，主，咳）、（肝，主，语）、（脾，主，吞）、（酸，入，肝）、（辛，入，肺）、（苦，入，心）

抑制：（怒，伤，肝）、（风，伤，筋）、（忧，伤，肺）、（热，伤，皮毛）、（恐，伤，肾）、（寒，伤，血）、（咸，伤，血）、（刺肉，无，伤筋）、（刺脉，无，伤筋）、（刺筋，无，伤骨）、（刺骨，无，伤髓）、（病生于脉，治以，灸刺）、（不仁，治以，按摩醪药）、（热反胜，治以，苦）、（厥阴胜，治以，甘

清）、（毒药，治，内）、（针石，治，其外）、（汤液，治，其内）、（病在心，禁，温食热衣）、（病在脾，禁，温食饱食）、（病在脾，禁，湿地濡衣）、（病在肺，禁，寒衣）、（气病，无，多食辛）、（血病，无，多食咸）、（肉病，无，多食甘）、（肝病，禁，辛）、（心病，禁，咸）、（脾病，禁，酸）、（肾病，禁，甘）、（肺病，禁，苦）、（肝，恶，风）、（心，恶，热）、（肺，恶，寒）、（肾，恶，燥）、（形东志苦，治之于，灸刺）、（形苦志东，治之以，熨引）、（形东志东，治之以，针石）、（形苦志苦，治之以，甘药）、（久视，伤，血）（久卧，伤，气）、（久坐，伤，肉）、（久立，伤，骨）、（久行，伤，筋）

因果：（阳气有余，则，外热）、（厥阴，病，阴痹）、（太阳，病，骨痹）、（阳，受入，六腑）、（阴，受入，五脏）、（辛，入于，胃）、（苦，入于， 胃）、（甘，入于，胃）、（胆，出于，窍阴）、（少阳，病，筋痹）、（肾者，作，强官）、（胃，走于，阳明）、（上，走于，息道）、（少阴，起于，涌泉）、（厥阴，起于，大敦）、（热气，留于，小肠）、（卫气，留于，腹中）、（经水，注于，海）、（道，在于，一）、（气，在于，肺）、（卫气，在于，身）、（病，在于，三阴）、（内舍，在于，脉）、（病，在，皮）、（足厥阴，外合于，浑水）、（手太阳，外合于，淮水）、（手阳明，外合于，江水）、（天地，通于，肺）、（风气，通于，肝）、（雨气，通于，肾）、（肝，受气于，心）、（心，受气于，脾）、（脾，受气于，肺）、（四气，始于，六十二刻六分）、（五气，始于，五十一刻）、（春气，始于，下）、（秋气，始于，上）、（夏气，始于，中）、（五谷，入于，胃）、（卫气，入于，阴）、（夏脉，如，钩）、（冬脉，如，营）、（病始手臂，取，手阳明）、（病始头首，取，项太阳）、（肺疟，令，心寒）、（心疟，令，烦心）、（心病，禁，咸）、（脾病，禁，酸）、（阴病，发于，骨）、（阳病，发于，血）、（黑，当，肾碱）、（青，当，肝酸）、（肺，出于，少商）、（心，出于，中冲）

包含：（心，为，五脏）、（肝，为，五脏）、（脾，为，五脏）、（肺，为，五脏）、（肾，为，五脏）、（弦，为，五脉）、（石，为，五脉）、（钩，为，五脉）、（代，为，五脉）、（毛，为，五脉）、（鼽衄，乃，病症）、（胸胁，乃，病症）、（风疟，乃，病症）、（痹厥，乃，病症）、（濡泻，乃，病症）、（地，有，五理）、（人，有，四经）、（脉，有，阴阳）、（病，有，标本）、（两臂内痛，为，浸淫）、（民病黄瘅，为，腑肿）、（炼白沙蜜，为，丸）、（诸脉，属于，目）、（诸髓，属于，脑）、（诸筋，属于，节）、（性，为，喧）、（用，为，动）、（色，为，苍）、（虫毛，为，散）、（眚，为，陨）、（味，为，酸）、（志，为，怒）、（枣，为，五果）、（薤，为，五菜）、（荧惑星，为，五星）、（哕，为，五变动）、（耳，为，五窍）

作者简历及在学研究成果

一、 作者入学前简历

起止年月	学习或工作单位	备注
2012 年 9 月至 2016 年 6 月	在北京科技大学信息安全专业攻读学士学位	

二、 在学期间从事的科研工作

2017.03—2017.09，最高人民法院“国家法官学院内网搜索引擎”项目，参与。

2017.10—2018.07，国家重点研发计划重点专项“大数据驱动的中医智能辅助诊断服务系统”（编号：2017YFB1002300）中的任务 4：“中医临床智能辅助诊断与决策推荐”，参与。

三、 在学期间所获的科研奖励

《Link Us 团队管理系统》. 计算机软件著作权. 登记号：2017SR170265. 中华人民共和国国家版权局，第一作者，2017.05.10

《iSports 健身服务软件》. 计算机软件著作权. 登记号：2017SR170270. 中华人民共和国国家版权局，第一作者，2017.05.10

《一种法律知识图谱自动构建方法》. 国家发明专利. 专利号：201710270508.7. 国家知识产权局，第六作者，2017.04.24，已受理

《一种法律文书案由分类器的自动构建方法》. 国家发明专利. 专利号：201710281403.1. 国家知识产权局，第一作者，2017.04.26，已受理

《一种中医理论典籍的知识图谱构建方法》. 国家发明专利. 专利号：201810910004.1. 国家知识产权局，第二作者，2018.08.11，已受理

《一种面向中医古文的知识库构建方法》. 国家发明专利. 专利号：201811174093.4. 国家知识产权局，第三作者，2018.10.09，已受理

四、 在学期间发表的论文

[1] Jin Pei ,Jia Qi. Design and development of an intelligent industrial production information system based on wisdom manufacturing[C]. Proceedings of the 5th Academic Conference of Geology Resource Management and Sustainable Development. 2017: 218-223. 已发表，国际会议，EI 检索，检索号：20183405716675

[2] Chen P, Xie YH, Jin P, Zhang DZ. A wireless sensor data-based coal mine gas monitoring algorithm with least squares support vector machines optimized by swarm intelligence techniques[J]. International Journal of Distributed Sensor Networks. 2018, 14(5). 已发表，SCI 期刊，SCI 检索，检索号：000432711600001

独创性说明

本人郑重声明：所呈交的论文是我个人在导师指导下进行的研究工作及取得研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得北京科技大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

签名：_____ 日期：_____

关于论文使用授权的说明

本人完全了解北京科技大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵循此规定）

签名：_____ 导师签名：_____ 日期：_____

学位论文数据集

关键词*	**密级***	**中图分类号***	**UDC**	**论文资助**
学位授予单位名称*		**学位授予单位代码***	**学位类别***	**学位级别***
北京科技大学		10008		
论文题名*		**并列题名**		**论文语种***
作者姓名*			学号*	
培养单位名称*		**培养单位代码***	**培养单位地址**	**邮编**
北京科技大学		10008	北京市海淀区学院路 30 号	100083
学科专业*		**研究方向***	**学制***	**学位授予年***
论文提交日期*				
导师姓名*			**职称***	
评阅人	**答辩委员会主席***		**答辩委员会成员**	
电子版论文提交格式 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ） **推荐格式：**application/msword；application/pdf				
电子版论文出版（发布）者		**电子版论文出版（发布）地**		**权限声明**
论文总页数*				
共 33 项，其中带*为必填数据，为 22 项。				