分类号 TP391.1　　密级 _____

U D C _____　　编号 10741

硕 士 学 位 论 文

论文题目 ：祁连山可持续发展知识图谱构建研究

研 究 生 姓 名：韩运龙

指导教师姓名、职称：尚庆生 教授

学 科、专 业 名 称：管理科学与工程

研 究 方 向：数据分析与信息处理

提 交 日 期：2024 年 5 月 31 日

独创性声明

本人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的研究成果。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过的研究成果。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作了明确的说明并表示了谢意。

学位论文作者签名：韩亚东 签字日期：2024.5.31

导师签名：尚庆生 签字日期：2024.5.31

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：韩亚东 签字日期：2024.5.31

导师签名：尚庆生 签字日期：2024.5.31

Research on the Knowledge Graph of Sustainable Development in Qilian Mountains

Candidate : Han Yunlong

Supervisor: Shang Qingsheng

摘 要

随着人口激增和经济的快速发展，气候变化和资源枯竭等全球性危机日益加深，人们对社会、经济、生态三个领域的可持续发展越来越重视，特别是在自然生态环境领域的可持续发展研究一直是学术研究的热门议题。同时祁连山作为我国西部重要的生态安全屏障和重要水源产地，祁连山地区的可持续发展对于我国西部乃至全国的生态文明建设至关重要。当前由于气候环境的变化和人类活动的影响，祁连山面临着日益严重的环境问题和可持续发展挑战，对祁连山可持续发展现状进行成体系的归纳整理和研究成为重要而紧迫的任务。当下自然语言处理、深度学习技术的飞速发展，为研究处理祁连山可持续发展信息提供了新的方向，知识图谱的出现也为相关生态环境学术研究和可持续发展领域提供了直观有效的工具。因此，本文利用知识图谱思想，构建祁连山可持续发展知识图谱，为祁连山可持续发展的智能研究、知识问答、知识推理等方面提供重要支撑。本文主要内容如下：

（1）建立祁连山可持续发展信息数据集。将中国知网数据库中公开发表的中文期刊论文作为主要数据源，检索祁连山可持续发展相关信息文献并进行处理，建立原始数据集用于祁连山可持续发展文献研究热点分析，确定好实体类型和关系类型后对文本数据进行标注，最后把文本转换成特定格式完成标注数据集的创建。标注数据集共包含 6 种实体类别和 6 种关系类别，为后续构建知识图谱提供数据支撑。

（2）提出了融合注意力机制的 ALBERT-BiLSTM-Attention-CRF 轻量化模型。模型在特征提取层 BiLSTM 的后面引入注意力层，解决了 BiLSTM 模型存在的问题，同时轻量化的 ALBERT 层使得模型在本文的命名实体语料规模相对较小的情况下，以其参数少的优势，能够取得更好的性能。经过与其他命名实体识别模型实验对比，本文模型的准确率、F1 值都有较高的提升，证明了模型的有效性与可行性。

（3）构建祁连山可持续发展知识图谱。利用确定好的实体和关系类型，通过 ALBERT-BiLSTM-Attention-CRF 模型进行实体识别，根据抽取的实体特征，基于规则模版完成实体关系抽取，构建<实体，关系，实体>三元组，最后采用

Neo4j 图数据库对知识图谱进行存储与可视化展示。

关键词：祁连山 可持续发展 知识图谱 实体识别

Abstract

With the rapid population growth and rapid economic development, global crises such as climate change and resource depletion are deepening, and people are paying more and more attention to sustainable development in the social, economic, and ecological fields. In particular, the research on sustainable development in the natural ecological environment has always been a hot topic in academic research. At the same time, as an important ecological security barrier and important water source in western China, the sustainable development of the Qilian Mountains is crucial for the construction of ecological civilization in western China and even the whole country. Currently, due to changes in climate environment and the impact of human activities, the Qilian Mountains are facing increasingly serious environmental problems and sustainable development challenges. It has become an important and urgent task to systematically summarize and study the current situation of sustainable development in the Qilian Mountains. The rapid development of natural language processing and deep learning technology provides a new direction for researching and processing sustainable development information in the Qilian Mountains, and the emergence of knowledge graphs also provides intuitive and effective tools for related ecological environment academic research and sustainable development fields. Therefore, this article uses the idea of knowledge graphs to construct a knowledge graph of sustainable development in the Qilian Mountains, providing important support for intelligent research, knowledge question answering, knowledge reasoning, etc. in sustainable development of the Qilian Mountains. The main contents of this article are as follows:

(1) Establish a sustainable development information dataset for the Qilian Mountains. Using the Chinese journal articles published in the

CNKI database as the main data source, retrieve and process the relevant information literature on sustainable development in the Qilian Mountains, establish an original data set for the analysis of research hotspots in sustainable development literature in the Qilian Mountains, and label the text data after determining the entity types and relationship types. Finally, convert the text into a specific format to complete the creation of labeled data sets. The labeled data set contains six types of entities and six types of relationships, providing data support for the subsequent construction of a knowledge graph.

(2) A lightweight model of ALBERT-BiLSTM-Attention-CRF with attention mechanism is proposed. The model introduces the attention layer behind the feature extraction layer BiLSTM, which solves the problems of BiLSTM model. At the same time, the lightweight ALBERT layer enables the model to achieve better performance with less parameters in the case of relatively small scale of the named entity corpus in this paper. Compared with other named entity recognition models, the accuracy of the proposed model The F1 value has a higher increase, which proves the effectiveness and feasibility of the model.

(3) Construct a knowledge graph for sustainable development in the Qilian Mountains. Using the determined entity types and relationship types, perform entity recognition through the ALBERT-BiLSTM-Attention-CRF model. Based on the extracted entity features, complete entity relationship extraction based on rule templates, construct a <entity, relationship, entity> triplet, and finally use the Neo4j graph database to store and visualize the knowledge graph.

Keywords: Qilian Mountains; sustainable development; knowledge graph; entity recognition

目 录

1 绪论

1.1 研究背景及意义

自 21 世纪伊始，全球所面临的气候变化与资源枯竭等危机不断加剧，全球人口急剧增长、经济飞速发展以及人类活动和对其自然资源需求的持续上升，对自然生态系统施加了前所未有的压力。尤其在重点生态保护区域，由于城市扩张、交通基础设施建设和过度放牧等高强度人类活动，造成了生态系统中草地的退化和湿地的减少，严重破坏了生态系统的结构和功能，从而导致生态系统服务质量显著下降，导致生物多样性丧失、水土流失以及土地沙漠化等生态环境问题不断凸显[1]。在过去的半个世纪中，全球 60%的生态系统遭受了人类活动的干扰与破坏，与此同时，全球三分之一的自然保护区正面临着严重的生态压力。这不仅对生态系统和其提供的生态系统服务产生了巨大的负面影响，还对人类的生存和可持续发展构成了严重的威胁[2]。因此，保护生态环境、推动全球可持续发展已成为人类共识。

祁连山作为我国西部重要的生态安全屏障和重要水源产地，拥有丰富而独特的生态环境和生物多样性，位于祁连山北坡的祁连山国家级自然保护区是黑河、疏勒河及石羊河三条主要内陆河的发源地，维系着河西走廊的生态平衡[3]，因此，祁连山的可持续发展对于我国西部乃至全国的社会-经济-生态文明建设至关重要。然而，近年来，随着气候环境的变化和人类活动的影响，祁连山面临着日益严重的环境问题和可持续发展挑战。为了保护和恢复祁连山的生态环境，实现其可持续发展，对祁连山可持续发展现状成体系的研究和分析成为重要而紧迫的任务。

知识图谱是 Google 公司在 2012 年提出的，现已被广泛应用于搜索、知识问答、知识推理等领域[4]。构建祁连山可持续发展知识图谱，能将祁连山可持续发展相关知识构建形成系统规范成体系的可持续发展知识库，为祁连山可持续发展智能研究、知识问答、知识推理等方面提供重要支撑。

知识图谱可以看作是由实体、实体和实体之间的关系所组成，可以表示为三

元组<Subject，Predicate，Object>，因此知识图谱的构建流程包括了实体和关系的信息抽取、知识融合、知识存储等步骤。其中信息抽取可以看作知识图谱构建的核心步骤[5]，信息抽取一般分为两个重要部分：一个是命名实体识别(Named Entity Recognition，NER)，其任务是识别出实体类别，一个是实体关系抽取(Relation Extraction，RE)，其主要用于识别两个实体之间的关联关系。通过信息抽取，从祁连山地理、空间、经济等相关信息中抽取大量的实体和实体之间的关系组成三元组，在此基础上可以构建祁连山可持续发展知识图谱。

综上所述，本文以祁连山为研究对象，利用知识图谱思想，将与祁连山可持续发展相关的知识有效地组织起来构建形成较为完善的祁连山可持续发展知识图谱，建立的规范成体系的知识库能够为祁连山可持续发展相关研究提供数据基础，对于祁连山地区社会、经济和环境的可持续发展有着重要的意义。

1.2 研究现状

1.2.1 可持续发展研究现状

可持续发展概念的起源可追溯至 20 世纪 60 年代。1962 年，瑞典斯德哥尔摩召开了第一届联合国人类环境会议，这次会议被认为是可持续发展思想的起点，会议提出了“人类环境宣言”，呼吁保护环境和促进经济发展之间的平衡。在 1972 年，罗马俱乐部的研究团队发布了一份名为《增长的极限》的报告，该报告深入探讨了未来的人口数量、食品供应、工业产值、环境污染以及不可再生的自然资源消耗等方面的相互关联[6]。1987 年，联合国在《我们共同的未来》的研究报告对可持续发展概念进行了较为全面的论述，并首次将可持续发展定义为：“既能满足当代人的需要，又不对后代人满足其需要的能力构成危害的发展”。从此，可持续发展成为国际政策制定和学术研究的核心议题。

可持续发展概念包含了三个核心维度：经济、社会和环境。在经济维度上，可持续发展追求经济增长和繁荣，同时避免对资源的过度消耗和破坏。社会维度强调了社会公正、包容和公民参与，促进人们的福祉和权利。环境维度则强调了对自然资源的合理利用和保护，减少对生态系统的负面影响。2015 年，在第 70

届联合国大会上，193 个成员国通过了包括 17 个可持续发展目标（Sustainable Development Goals，SDGs）和 169 项具体目标为核心内容的《变革我们的世界：2030 年可持续发展议程》[7]。这些目标覆盖了从贫困消除到清洁能源，从经济增长到气候行动的广泛议题，旨在以综合方式全面解决社会、经济和环境 3 个维度的发展问题，解决人类和地球面对的持续性问题和新兴挑战，从而使人类全面走向可持续发展的道路。

总之，可持续发展是一个综合性的理念，涉及到经济、社会和环境的多重考量。它代表了人类对未来的关注和责任，是一种追求整体性和长期性的发展理念。随着绿色技术的发展与社会意识的提升，可持续发展也带来了新的经济机遇和创新潜能。

1.2.2 祁连山可持续发展研究现状

祁连山位于中国西北地区，是中国最重要的山脉之一。这座山脉东起甘肃省，西至青海省，南北绵延约 800 公里，总面积约为 5.2 万平方千米。祁连山地区是我国西部重要的水源地和生物多样性优先保护区域，是中国重要的生态屏障和自然资源宝库。其独特的气候条件和地理位置造就了丰富而多样的生物圈，草地、森林、湿地等资源分布广泛，承载着水源涵养、冰川保护和野生动物保护、调节气候与供水等重要的生态功能。然而，长期以来，在气候变迁、超载放牧、人为破坏及保护措施落后等多个因素的共同影响下，祁连山地区逐渐出现冰川退缩、水源涵养效能减弱、植被退化严重以及水土流失加剧等一系列问题。特别是在最近 30 多年里，全球多变的气候因素对祁连山地区的生态环境造成了严重的威胁，进一步加剧了其脆弱性[8]。

祁连山可持续发展的相关研究已经积累了较多的成果，国内外对于祁连山地区可持续发展问题的研究主要涉及生态保护、水资源变化、植被动态变化、社会经济发展等方面。

在有关祁连山地区生态环境的研究中，王涛等[9]在初步掌握祁连山生态环境现状的基础上，从制度机制、生态监测、生态红线、生态补偿和生态修复等五个角度，对祁连山生态环境现状及面临的主要问题进行分析并提出对策建议。王有

恒等[8]通过研究气候变化对祁连山地区的水资源影响发现，2000 年以来，祁连山气候升温加剧，四季降水量均呈增加趋势，同时冰雪融水增加，石羊河、黑河和疏勒河出山径流均呈增加趋势。蒋强等[10]对祁连山及周边地区降水的时空分布特征进行分析，发现降水整体呈东多西少、南多北少的空间分布特征。盖迎春等[11-14]构建黑河流域中游灌溉管理系统、黑河流域水资源管理决策支持系统和黑河流域可持续发展决策支持等系统，通过定量模拟灌溉过程、水资源分配过程以为流域水资源管理和综合可持续发展提供了决策支持。张江蕾等[15]探究祁连山地区植被覆盖度的时空变化及不同地形条件下的分异特征，发现发现 2000—2020 年祁连山自然保护区的植被状况有所恢复，海拔 2500～4 000 m、坡度 5°～25°及半阴坡地区植被覆盖面积增加明显。杨欣等[16]为揭示祁连山植被覆盖变化，结合植被、地貌和气象数据发现祁连山西北地区植被覆盖度较低，东南地区植被覆盖度较高，在 1982—2022 年祁连山植被覆盖度明显提高，但近十年以来植被有退化的趋势。张强等[17]基于卫星影像资料开展了祁连山生态敏感性评估研究，发现不同区域间的生态敏感性有明显差别，西部地区较东部更为敏感，并针对敏感性不同的区域分别给出相应的应对措施。

在祁连山地区社会经济发展的相关研究中，汪慧玲等[18]对祁连山区域的社会、经济发展状况进行了分析，认为其存在的突出问题是：经济来源与产业结构单一，生态保护与生产生活相互制约，并提出应强化祁连山地区的产业融合，发展生态旅游和新型农牧业的建议。邸华等[19]的研究指出，祁连山的森林保护区在经济发展方面表现出明显的滞后现象，主要依赖于农业和畜牧业，整体经济水平相对较低。蒋志成等[20]研究表明，祁连山区域的生态旅游开发给当地居民带来了双重影响，其中正面影响主要体现在居民意识的提高和收入的增加，负面影响主要源于“追逐利益”，然而，正向影响的力量更为显著。张文昌等[21]对祁连山地区生态与经济可持续发展的路径进行研究，剖析祁连山地区的生态现状，从加强生态保护，健全生态补偿制度和发展生态经济等方面进行了探讨。李华芸等[22]以天祝藏族自治县为研究对象，研究祁连山保护地的生态状况，剖析该地区的经济发展面临的困难，探索了一种既能保护环境，又能促进社会经济发展的双赢路径。祁帜等[23]以肃南县研究对象，在坚定的理念加持下，提出了以生态环境保护为主，

并通过渐进的方式推动社会经济和产业的发展，以克服转型困境。王天雁等[24]在祁连山重要牧区展开调研，发现尽管祁连山草场保育工作取得了显著效果，但牧户的可持续收入却受到制约，为此，提出了综合运用法制化手段来帮助牧民实现增收致富的建议。

通过对以上文献进行梳理，发现祁连山区域在生态和社会经济方面，存在生态承载能力有限及经济来源比较单一的问题。祁连山地区东西部之间的生态状况有着明显的差异，西部地区生态环境遭到较为严重的破坏，传统的生产生活发展模式对祁连山生态环境也伤害较大，目前还存在着冰川消融、植被退化及生态敏感脆弱等现象。但随着社会经济的快速发展，人们对于环境保护和可持续发展的意识不断提高，相关政策和法规的完善为祁连山地区的可持续发展提供了新的挑战和机遇，生态旅游等新兴产业的发展也为祁连山地区的可持续发展提供了新的社会经济发展机会。

1.2.3 知识图谱研究现状

随着大数据及人工智能时代的到来，知识工程越来越受到学者们的重视，如何将海量的数据进行提炼，进而获取到有用的信息，是大数据分析的关键，而知识图谱（Knowledge Graph）技术正好提供了一种从海量文本和图像中提取结构化知识的重要手段[25]。知识图谱是一种独特的图形工具，它能够以直观的方式展现科学知识的内在联系以及科学进步的历程。这种方法巧妙地结合了传统文献计量法和现代文本挖掘技巧，同时融入了诸如复杂网络、统计学、数学和计算机科学等多个学科的理论和技术。2012 年 Google 公司提出了知识图谱的概念[26]，将现实世界中各种类型的数据转化为“节点-边-节点”三元组结构，其中节点代表物理世界中的实体或概念，边表示实体与实体的关系，现已被广泛应用于搜索、智能问答、知识推理等领域[27]。

根据所涵盖的知识点的广度，知识图谱可以被划分为两大类别：一是适用于所有领域的通用的知识图谱，二是针对特定领域的特定知识图谱。国内外对通用知识图谱的研究相当多，例如国外知名的知识图谱有 Freebase[28]、YAGO[29]、DBpeida[30]等；国内比较成熟的知识图谱有百度知心、搜狗知立方、CN-Probase[31]

等，这一类知识图谱主要应用在搜索领域，强调知识的广度。领域知识图谱不同于通用知识图谱，领域知识图谱对行业知识的准确性、权威性要求非常高，通常要求其能够实现领域辅助决策分析的功能。

近年来，通用知识图谱普及性更高，是现有知识图谱的基座，而领域知识图谱则是与各领域相结合，并在构建完成后并入通用知识图谱，为将来的深入研究提供数据支撑。目前领域知识图谱的研究主要集中于电子商务、医疗和金融等领域。Zheng[32]等在化学工业领域采用知识图谱的方法进行研究，以化工行业的企业信息为基础，结合化工企业的公开信息，建立化学工业的知识库。Rotmensch等[33]将 Logistic 回归与贝叶斯网络的随机建模相结合，自动搭建医学领域中的知识图谱。咎红英等[34]在收集和整理各类医疗文献资料的基础上，将医疗文献数据划分为 12 个实体类别以及 12 种关系类型，对医疗文献数据进行标注后，利用实体识别、关系抽取等方法，建立起相应的医疗文本知识图谱。Liu 等[35]以轨道交通安全事件为研究对象，通过收集轨道交通安全事件相关数据，建立基于事件之间相关联的知识图谱，基于知识图谱开展安全事故方法探究，以提供轨道交通安全的预警与防控策略的制订参考。在司法领域，陈彦光等[36]将涉毒案件刑事判决书作为数据来源，构建刑事案例知识图谱，并基于 Neo4j 图数据库完成对知识图谱的可视化存储。近年来在生态环境领域，许多研究人员也开始探索应用知识图谱的可能。常晋义等[37]提出了一个新的视角，即从生态监管的角度出发，建立一个生态环境监测信息的综合知识图谱数据库。徐超等[38]在生态红线研究的基础上，创建了一个专门用于生态红知识库的服务管理工具。陈兰鑫等[39]通过深入分析洞庭湖生态监测数据、监测技术以及监测目标之间的关联性，成功地构建了洞庭湖生态环境监测系统的知识图谱。张华等[40]基于生态环境领域文献文本构建生态环境领域知识图谱。这些研究为生态环境领域的知识体系构建提供了有价值的方法和技术支持。在国内外众多学者的深入研究之下，知识图谱相关的研究成果得以在医药、地理学、金融等多个领域广泛应用，为社会的进步与发展做出了显著的贡献。

1.3 研究内容与框架

1.3.1 研究内容

生态环境领域和地理领域的知识图谱已有学者开始探索研究，并取得不错的进展，自然语言处理、深度学习等技术也在生态、地理实体识别任务中也有着很好的应用场景，但是目前缺乏针对祁连山地区相关信息的深度学习训练数据集，同时也没有对祁连山地区构建知识图谱、建立知识库的研究。因此，基于上述问题，本文的主要研究内容如下：

（1）数据集获取与构建。针对目前祁连山地区相关信息数据集、知识库缺少的问题，从中国知网数据库中检索公开发表的中文期刊论文建立祁连山可持续发展信息原始数据集，并根据本文任务需求，确定实体类型、关系类型以及数据标注数据集。

（2）命名实体识别模型构建。在知识图谱构建中，生态环境领域相关实体类别繁多，难以归纳整理，而准确的实体识别对于构建知识图谱尤为重要。针对此问题，为提高祁连山可持续发展信息实体识别准确率，本文提出了融合注意力机制的 ALBERT-BiLSTM-Attention-CRF 轻量化模型。ALBERT 的参数较少能够提升模型的泛化能力，引入注意力机制能够使模型更好的关注文本序列位置信息。

（3）祁连山可持续发展知识图谱构建。将处理好的数据经过本文提出模型进行实体识别后，人工标注实体关系，构建“实体-关系-实体”三元组，最后通过 Neo4j 图数据库完成对知识图谱的存储与可视化展示。满足对祁连山可持续发展信息建立知识库，实现知识查询，信息搜索等应用领域的需求。

1.3.2 研究框架

本文根据研究内容分为五个章节，论文框架如图 1.1 所示，各章节的内容安排如下：

第一章绪论。绪论部分首先对本文的研究背景及意义进行详细阐述，然后介绍了可持续发展概念，祁连山可持续发展研究现状和知识图谱研究现状。最后提出了本文的研究内容和组织架构。

第二章相关理论与技术研究。本章节详细介绍了知识图谱构建过程涉及到的

相关理论与技术，包括构建流程、存储方式、命名实体识别技术、实体关系抽取技术、条件随机场、长短时记忆网络、注意力机制、Transformer 模型和 BERT 模型等。

第三章融合注意力机制的命名实体识别模型。本章节首先对 BERT-BiLSTM-CRF 模型的基本结构进行了介绍，接着对本文提出模型的 ALBERT 层和融合注意力机制的特征提取层进行了介绍，并自行构建祁连山可持续发展数据集，在数据集上对模型进行了对比实验，验证了本文模型的有效性。

第四章知识图谱构建与存储。本章节首先阐述了知识图谱的构建框架，其次完成知识图谱中的关系定义，然后通过 Neo4j 图数据库完成祁连山可持续发展知识图谱的构建和存储工作，基于 CiteSpace 软件对祁连山可持续发展文献数据进行了研究热点分析，最后对本文构建的知识图谱在祁连山综合集成研讨厅中的应用进行阐述。

第五章总结与展望。本章节对本文的研究内容进行总结，并提出下一步展望。

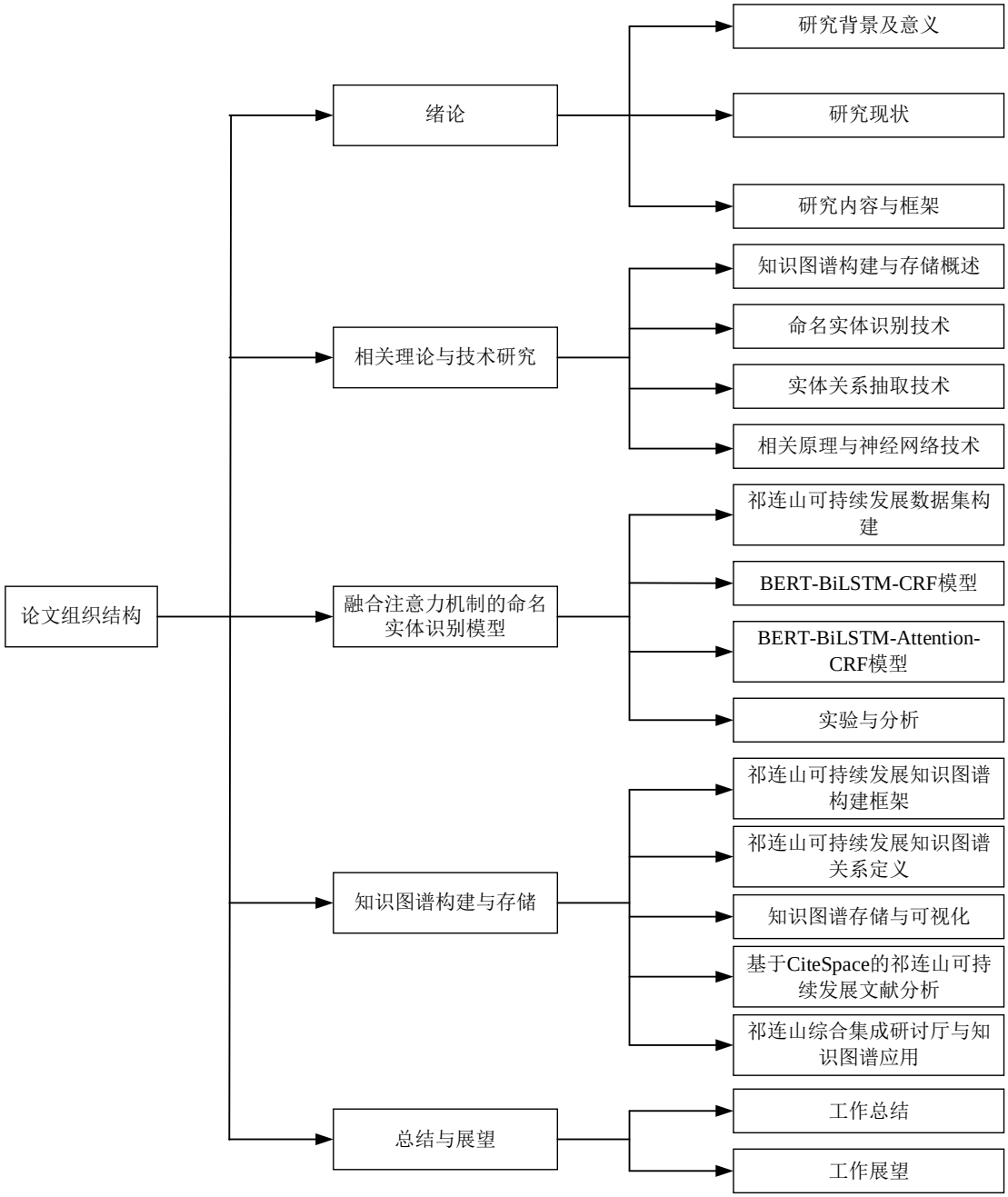

图 1.1 论文组织结构

1.4 本章小结

本章首先阐述了论文的研究背景及意义，接着总结了可持续发展概述和祁连山可持续发展的研究现状以及知识图谱的研究现状、最后介绍了主要研究内容和论文组织结构，为本文的下一步研究奠定了基础。

2 相关理论与技术研究

本章主要介绍知识图谱构建工作中用到的相关理论和技术，首先介绍了知识图谱的构建流程和储存方式，其次对知识图谱构建技术中的命名实体识别和实体关系抽取技术进行阐述，最后介绍在构建过程中需要用到的深度学习、神经网络模型。

2.1 知识图谱构建与存储概述

2.1.1 知识图谱构建概述

知识图谱的构建方法主要分为两大类：自顶向下和自底而上的方法。自顶向下的方法先为知识图谱定义好本体与数据模式，再抽取实体加入到知识库。该构建方式需要利用一些现有的结构化知识库作为其基础知识库，并且对数据质量要求较高[41]，这种方法适用于小规模的知识图谱构建，通常被应用于领域知识图谱的构建中。目前，应用最广泛的是自底而上的构建方式，这种方法借助机器学习、自然语言处理等技术，在开放的数据集中抽取出实体、关系、属性等信息，再经过知识融合等阶段最终构建成一个完整的知识图谱。

知识图谱的构建过程主要包含了数据获取、信息抽取、知识融合和知识存储四个阶段。其中数据获取是从学术论文数据库、政府公开数据、专业网站等渠道获取文本数据，经过机器学习、自然语言处理等技术进行数据清洗与处理；信息抽取过程通过命名实体识别和实体关系抽取技术从处理好的数据中提取出实体、属性以及实体间的相互关系；知识融合能够规范化整合不同来源的知识，在获得新知识之后，需要对其进行整合，以消除实体歧义，比如某些实体可能有多种表达，某个特定称谓也许对应于多个不同的实体等；知识存储过程是将处理好的三元组“实体-关系-实体”，通过 Neo4j 等图数据库进行保存。知识图谱构建流程如图 2.1 所示。

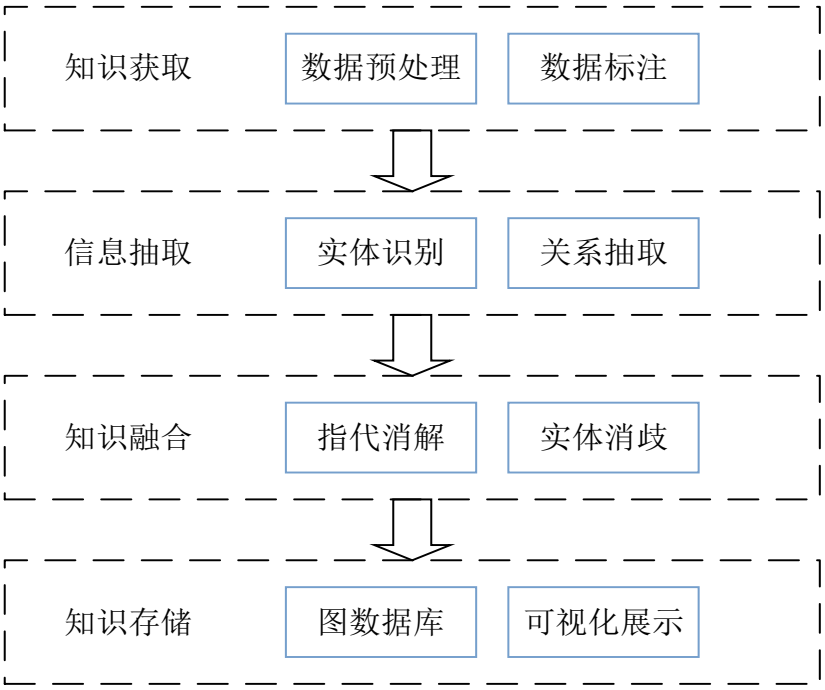

图 2.1 知识图谱构建流程图

2.1.2 知识图谱存储概述

知识图谱的存储方式主要有 RDF 存储和 Neo4j 等图数据库存储。资源描述框架(Resource Description Framework，RDF)[42]是用于描述网络资源 W3C 标准，RDF 存储将数据以三元组<Subject，Predicate，Object>的形式进行存储，通过三元组描述资源之间的关联。图数据库是以数学中的图论为理论根基的非关系型数据库，除了支持存储、分析、处理数据的功能之外，还更擅长海量数据之间的复杂关系分析，图数据库 Neo4j 是目前最为流行的知识图谱存储工具。

Neo4j[43]是一种由 Java 编程语言开发实现的开源非关系型图数据库，于 2007 年正式发布，除了支持 Java 语言开发的需求，还可以通过内在驱动 py2neo 进行 Python 语言的开发。Neo4j 的诞生就是专门为了图数据的存储，对于存储在其中的节点、边、属性等属性图结构部分都有专门的存储计划，Neo4j 能够简单清晰明了的描述纷杂的实体间关系。

在 Neo4j 图数据库中能够使用 Cypher 查询语言快速有效的查询知识图谱的相关信息，其可视化界面使得用户能够更加便捷、直观地对数据进行分析与处理。

2.2 命名实体识别技术

命名实体识别技术(Named Entity Recognition，NER)是信息抽取过程的一部分，其任务是识别出文本中具有特定意义的实体，如人名、地名、机构名、专有名词等，在命名实体识别任务中，技术早期使用基于规则和字典的方法，之后以传统机器学习方法为主，最后随着深度学习技术的迅速发展，逐渐转向基于深度学习的算法模型作实体识别任务的方法。

2.2.1 基于规则和词典的方法

基于规则和词典的方法是命名实体识别最早期的研究识别技术[44]，其依靠语言学家以专业的语言学知识，对特定领域的特征进行分析，人工制定独属于该领域的规则。但是这种方法难以迁移到别的领域，对于新的领域需要重新制定规则[45]。当词典的大小有限时，基于规则的方法可以达到很好的效果，而随着实体的增加，实体越来越丰富，面对大量的数据集想要好的识别效果，需要更加复杂的规则模板和词典集，这样就消耗大量的时间和人力[46]。因此这种方法并没有广泛的流行起来。

2.2.2 基于机器学习的方法

在 20 世纪末随着机器学习的发展，传统机器学习方法开始出现在命名实体识别任务中。研究者们将命名实体识别任务看做是一种序列标注任务，将要识别的实体看作是将要被标记的标注问题，利用词、上下文、 统计信息等特征训练命名实体识别模型来提取实体块，再将提取出来的实体块归纳整理，最终获得由若干词构成的命名实体及其类别。基于传统机器学习方法的模型主要有：支持向量机模型（Support Vector Machine，SVM）[47]，最大熵模型（Maximum Entropy，ME）[48]，隐马尔可夫模型（Hidden Markov Model，HMM）[49]和条件随机场模型（Conditional Random Field，CRF）[50]。在这些方法中，SVM 模型对文本序列进行逐字分类，判断实体的类别和边界。然而，这种方法在文本连续语义信息的提取上存在一定局限性，因此在实体识别方面表现不佳。ME 模型在设计上更加

紧凑，具有较强的普适性优势，但是该模型需要较长的训练周期，同时系统计算开销也较大。HMM 模型假设当前的转移状态只依赖前一刻的状态，假设当前的观测状态只依赖该时刻的状态。通过同时对转移状态和观测状态学习，有效地融合了文本的时间序列信息，它的训练和实体识别速度较 ME 模型速度要快上不少，但 HMM 模型的实体识别准确性有待提高。CRF 模型打破了隐马尔可夫的两个假设，能够更加全面地捕捉了文本中的全局依赖信息，从而有效地解决实体识别过程中的边界转移问题[51]。由于 CRF 模型的分类性能较好,目前多数作为融合模型中的一部分，在深度学习算法广泛应用。

2.2.3 基于深度学习的方法

随着计算机算力的提升，深度学习技术取得了迅猛发展，研究者们也开始将深度学习技术应用到命名实体识别任务中。相较于基于机器学习的方法，深度学习方法不需要人工提取特征，而是通过深度学习模型自动完成特征提取、 训练和预测任务。在命名实体识别任务中，常用的深度学习模型有卷积神经网络（Convolutional Neural Network，CNN）[52]、循环神经网络(Recurrent Neural Network，RNN)[53]、长短时记忆网络（Long Short-Term Memory，LSTM）[54]和 Transformer 模型等。这些模型能够有效地识别命名实体，并在不同的领域和语言中获得良好的性能。大量实验表明，在命名实体识别任务中，相比机器学习，深度学习技术可以用更少的资源而取得更好的效果。目前最主流的深度学习命名实体识别方法是基于双向长短期记忆网络(Bidirectional Long Short-Term Memory，BiLSTM)和条件随机场（CRF）的命名实体识别方法 BiLSTM- CRF[55]。在 BiLSTM-CRF 模型中，BiLSTM 层能够通过前向和后向状态来获取完整的上下文信息，CRF 层能够获取标签之间的顺序依赖信息，将 BiLSTM 层的输出作为 CRF 层的输入，可以得到最终的实体预测结果。同时由于注意力机制在处理文本序列时非常有效，Transformer 模型和 BERT 预训练模型也开始出现在命名实体识别任务中，并取得了很好的识别效果。

2.3 实体关系抽取技术

信息抽取过程中的又一个重要的研究是实体关系抽取技术（Relation Extraction，RE），其目的是识别实体之间特定的语义关系，最终构建“实体-关系-实体”这样的三元组。经过研究者们的多年探索，实体关系抽取的方法研究已经有较为丰富的成果，从早期的基于规则的抽取方法发展到现在主流的深度学习技术的流水线方法和联合抽取方法。

2.3.1 传统抽取方法

基于规则的方法同命名实体识别早期技术一样，依靠人工提前定义若干个基于词法、词义、词性的规则模式集合，利用这些规则模式集合去描述实体间关系。关系抽取时，将经过预处理的文本片段与提前规定好的规则模式进行匹配对应，而后进行实体关系判别，最后完成实体关系的抽取[56]。基于规则的方法主要分为基于触发词模式和基于依存关系两种方法。基于触发词模式方法将抽取关系变为抽取触发词，即抽取人工制定的指示关系存在的词语或短语，基于依存关系方法以动词为起点构建规则，对节点上的词性和边上的依存关系进行限定。基于规则的抽取方法，在特定领域和小规模数据集上取得了一定的成果，但需要大量的人工来构建所有可能的关系规则，语料人工标注成本过高，同时模型缺乏可移植性、召回率也比较低。

在基于传统机器学习的框架下，实体关系抽取被处理成一个实体关系分类的问题。这一方法开始于构建训练和测试样本，接着使用训练样本来训练一个关系抽取模型，再将该模型应用于测试样本以进行预测。基于机器学习的方法可以划分为有监督学习、半监督学习、无监督学习[57]。有监督学习方法将关系抽取问题视为多个关系分类问题的处理，全部使用的是标注过的数据集。有监督学习方法在多个关系抽取任务的数据集上都表现出良好的效果。然而，这一方法对标注过的数据依赖较高，增加了成本和工作复杂度。半监督学习方法在训练分类模型时，只使用有限的标注数据作为训练样本，并通过不断的迭代学习完成训练，从而有效地减少人工标注数据的需求。半监督学习的方法避免了大量的人工成本，但还存在着高度依赖训练样本质量以及迭代过程中可能引入噪声等问题。无监督学习在处理大规模语料库时，通过聚类技术自主地抽取实体关系，无需依赖人工标注

数据[58]。相较于有监督学习和半监督学习，无监督学习方法的优势在于其在大规模文本和无规则内容处理方面有着高度的适应性和可移植性。然而相对而言，无监督学习方法的准确性较低。

2.3.2 基于深度学习的抽取方法

相较与传统的机器学习方法，基于深度学习的关系抽取方法通过大量数据训练神经网络来学习语义特征，无需依赖人工特征选择，从而能够更好地进行关系抽取任务。因此，随着时间的推移，基于深度学习的方法逐渐成为实体关系抽取领域的主流方法。

基于深度学习的实体关系抽取方法，根据实体识别和关系抽取两个子任务完成的先后顺序，可分为两类方法：流水线方法和联合抽取方法[59]。早期的流水线的方法主要采用 CNN 和 RNN 模型，随着深度学习技术的方发展，改进和变形的网络模型结构 LSTM 和基于 Transformer 的双向编码表示（Bidirectional Encoding Representation from Transformer，BERT）相继出现，促使关系抽取模型性能得到更大的提升。基于 CNN 模型的方法[60]无需复杂的预处理，可以有效地进行句子的局部特征提取，将句子特征和语法特征结合进行关系分类。CNN 模型解决了从预处理系统中提取的特征可能会导致错误传播并阻碍系统性能的问题，但提取的特征被局限于卷积核内的信息，没有将文本距离较大的信息等提取到。RNN 是一种适用于处理时序型数据的神经网络，具有学习任意长度的各种短语和句子的组合向量表示的能力，基于 RNN 模型的抽取方法[61]将句子中的每个词语使用词向量表示后，通过双向 RNN 从两个方面来学习单词级特征，将单词级特征向量每个维度最大池化得到句子级特征向量，据此进行关系分类。虽然 RNN 能够捕捉到单词之间的位置关系和长短语的构成意义，但是较多的模型层数会导致出现梯度爆炸和梯度消失的问题。为了解决 RNN 存在的问题，研究者们将 RNN 的改进模型 LSTM 应用到了实体关系抽取任务中。LSTM 能够从语料中学习到长期依赖关系，捕捉到文本中较远距离的上下文信息，在关系抽取任务中有着较好的效果。LSTM 和结合 CNN、注意力机制等其他技术的关系抽取方法现已经成为实体关系抽取任务中的主流方法。基于 BERT 等预训练模型的方

法近年来在实体关系抽取任务中也取得了显著的进展。预训练模型能够通过大规模无监督的语料库学习通用的语言表示，从而捕捉到丰富的语义信息。这些模型广泛应用于实体关系抽取任务中，并通过微调或结合其他模型进行关系分类和抽取。

在关系抽取任务中，流水线方法的运用有效提升了性能。然而，在训练过程中，未充分考虑命名实体识别和实体关系抽取两个子任务之间的相关性，可能导致错误的传播，产生冗余信息[62]。为了解决这些问题，研究者人员提出了联合抽取方法[63]，将命名实体识别和实体关系抽取两个子任务进行整合，以期获得更好的结果。联合抽取模型通过单一模型同时处理实体及其关系抽取，该方法可以直接从非结构化文本中提取出<实体，关系，实体>三元组信息，实现了实体识别和关系抽取两个子任务间的隐性关联特征的有效整合，克服流水线方法具有的局限性。联合抽取模型方法根据其建模策略的不同，分为基于参数共享的方法和基于联合解码的方法。前者将实体和关系的建模过程分开处理，而后者则直接以三元组为基础进行模型构建。

基于参数共享的抽取方法，通过共享联合模型的编码层参数实现两个子任务的相互依赖，不断优化全局任务参数提高模型性能。在训练中，输入文本经过共享的编码层处理后，首先在解码层执行实体识别任务，然后根据实体识别的结果对有关联的实体对进行关系抽取，最终输出<实体-关系-实体>三元组。尽管基于参数共享的方法能在一定程度上缓解错误信息传递，但由于其对实体识别任务和关系抽取任务分别建模，可能会导致产生独立的实体信息，产生冗余。于是研究人员为了增强实体识别与关系抽取任务之间的依赖，提出了基于联合解码的方法[64]。联合解码是指在编码层采用统一的解码器，直接解码得到结构化的三元组，从而实现实体识别和关系抽取任务之间的深度融合，进一步减少错误传递，并提高关系抽取的准确性。然而，该方法因其具有较高的复杂性，导致设计一个有效的联合解码算法成为一个挑战难点。

2.4 相关原理与神经网络技术

2.4.1 条件随机场

条件随机场（Conditional Random Field，CRF）[65]是一种概率图模型，通过学习输入序列和输出标签之间的条件概率分布来进行预测，常用于序列标注任务，是命名实体识别任务中常用到的算法模型。CRF 首先给定输入序列 $X=\{x_1,x_2,...,x_n\}$，输出对应的实体标注系列为 $Y=\{y_1,y_2,...,y_n\}$。CRF 用 BIO 标记法表示实体的模型结构如图 2.2 所示，其中 B 表示实体开始，I 表示实体内部，O 对应非实体外部。

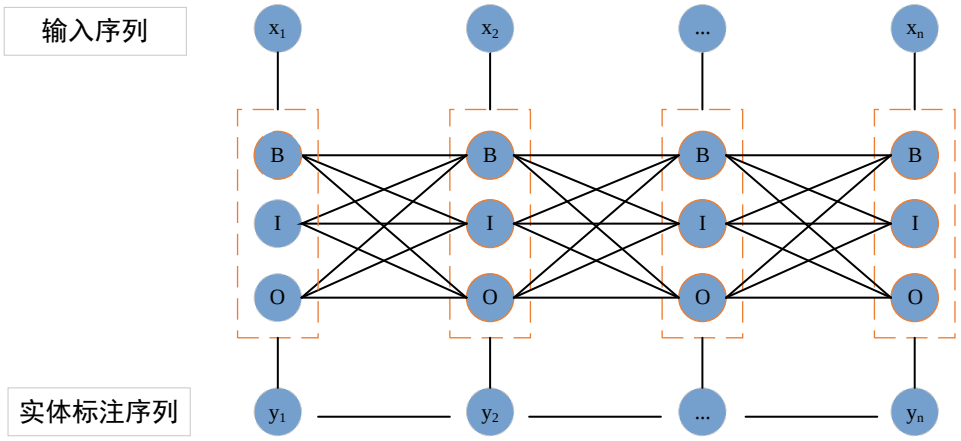

图 2.2 CRF 命名实体识别模型

CRF 模型的核心主要包含两类矩阵，即发射矩阵和转移矩阵。发射矩阵计算每个字符在特定标签下的概率分布，而转移矩阵则计算不同标签之间的转换概率。CRF 根据这两种矩阵计算文本对应的最大分数，此时预测的标签序列是与文本实际标签序列最接近的。CRF 假设条件概率分布函数 $P(y|x)$，条件概率计算公式为：

$$P(y|x)=\frac{1}{Z(x)}\exp(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)) \tag{2.1}$$

其中 t_k 是定义在边上的局部特征函数，即 $P(Y_i|Y_{i-1},Y_{i+1})$，又称之为转移函数，它同时依赖于当前位置和上一个位置。s_l 是定义在节点上的节点特征函数，即

$P(Y_i \mid X)$，又称之为发射函数。λ_k，μ_l是两个特征函数的权重。$Z(x)$是规范化因子，$Z(x)$公式为：

$$Z(x)=\sum_{y}\exp(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)) \tag{2.2}$$

2.4.2 长短时记忆网络

长短时记忆网络（Long Short Term Memory，LSTM）[66]基于循环神经网络进行优化，能够有效地避免梯度消失或梯度爆炸问题。LSTM 在 RNN 的基础上添加了一个记忆细胞，在隐藏层加入了三个门控机制：遗忘门、输入门和输出门。其中记忆细胞负责保存某一时刻的信息，遗忘门用来决定上一时刻保留到当前时刻的信息部分，输入门负责决定新信息是否存储到记忆细胞中，输出门确定当前时刻记忆细胞中需要输出的信息。LSTM 结构图如图 2.3 所示。

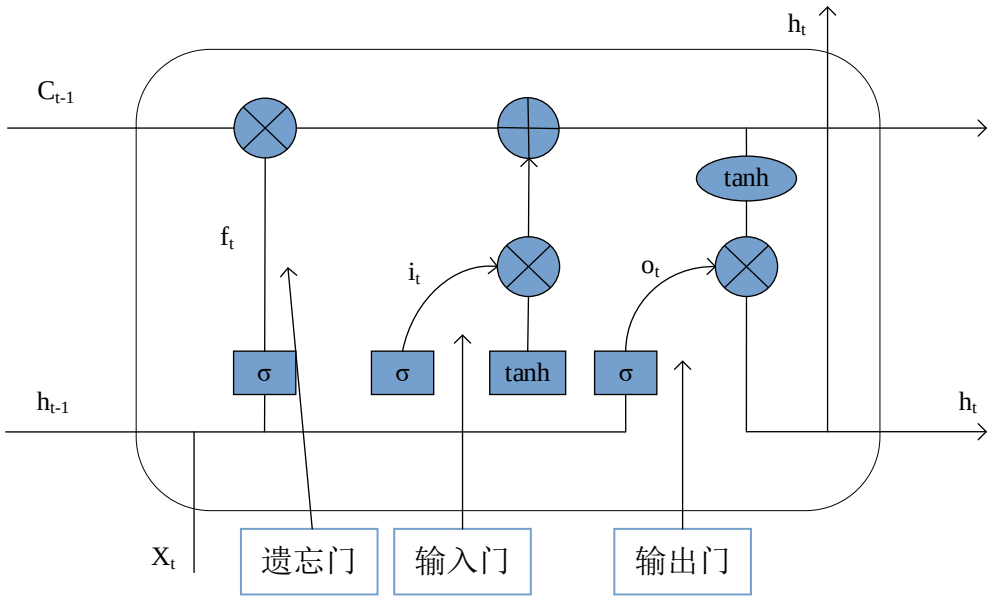

图 2.3 LSTM 模型结构图

遗忘门获取上一单元的输出h_{t-1}和当前输入x_t，通过 sigmoid 函数计算出遗忘概率f_t，f_t范围为 0-1，0 代表全部遗忘，1 代表全部保留，遗忘门的公式为：

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2.3}$$

其中 σ 表示 sigmoid 函数，W_f，b_f 是遗忘门的权重和偏置。

输入门先将获取到的 h_{t-1} 和 x_t 通过 sigmoid 函数计算出要保留的信息比例 i_t，再利用 tanh 函数计算出候选记忆细胞 $\tilde{C}_t$。最后通过遗忘概率 f_t、i_t、上一单元记忆细胞 C_{t-1} 和当前 $\tilde{C}_t$ 更新本单元记忆细胞 C_t。输入门公式为：

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2.4}$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{2.5}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.6}$$

其中 W_i，b_i 是输入门的权重和偏置，W_c，b_c 是计算候选记忆细胞的权重和偏置。

输出门先计算出信息的输出程度 O_t，再利用 tanh 函数计算输出值 h_t。输出门公式为：

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{2.7}$$

$$h_t = o_t * \tanh(C_t) \tag{2.8}$$

其中 W_o，b_o 是输出门的权重和偏置。

通过门控机制，LSTM 模型能够更好地处理长序列数据，在词性标注、命名实体识别等任务中取得了很好的效果。此外，LSTM 模型还可以增加一个反向 LSTM，或者与条件随机场、注意力机制等方法相结合，进一步提高模型的性能。

2.4.3 注意力机制

人类的大脑在处理信息时具有一定的限制，这主要体现在注意力上。当我们面临一段文字或一幅图像时，我们的大脑只会将更多的精力花费在高价值信息部分，同时过滤掉一些无关紧要的信息。这种现象源于大脑的自然反应，即集中资源在那些具有更高价值的信息上。注意力机制使得我们能够在海量信息中迅速找

到关键点，这也是我们在漫长的进化历程中所发展出的适应性策略。在深度学习领域，注意力机制被视为一种模仿人类视觉注意力的方法[67]。其核心思想是对输入的信息赋予不同的权重，过滤无关内容的同时提升对关键内容的关注度，使得模型注意权重较高的信息，给予这些信息更多的处理资源，从而加速处理、提高任务准确性，并且可以根据情况动态调整权重。近年来，越来越多的研究者开始探索如何将注意力机制与神经网络相结合，以期在各种任务中实现更高效的数据处理。注意力机制在任务建模中扮演着重要角色，它可以根据不同的原始数据赋予不同的权重，从而生成更为精确和有代表性的数据。在构建知识图谱的过程中，需要处理大量的文本信息，因此可以引入注意力机制提高模型的学习能力，从而提高对实体分类的准确性。注意力机制模块如图 2.4 所示。

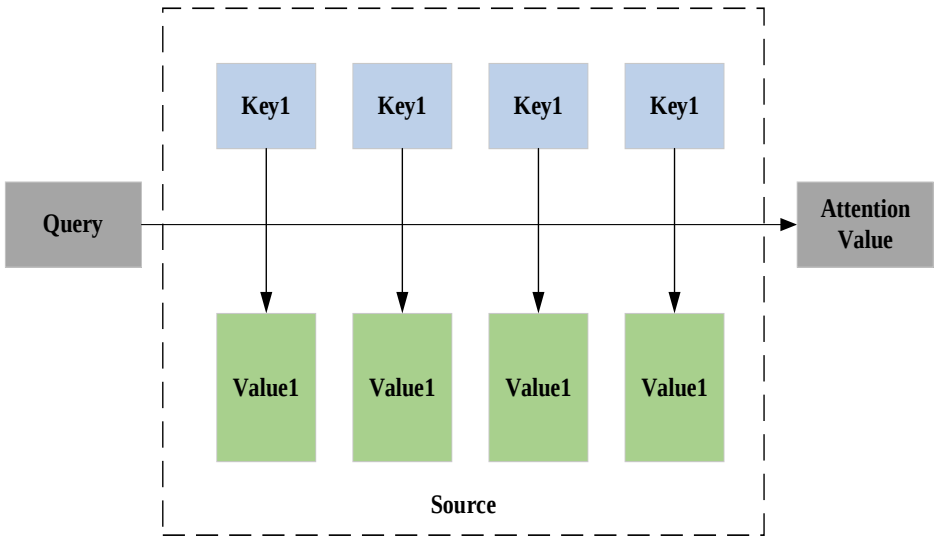

图 2.4 注意力机制模块

从模块图中可知，注意力机制包含三个向量，分别是：查询向量 Query、键向量 Key 和值向量 Value。注意力机制对于每一个 Query 向量，通过注意力打分函数计算其与 Key 向量的相似度，再经过归一化处理得到 Value 向量权重系数，最后加权求和得到最终的 Attention 输出。常用的注意力打分函数有点积、加性模型、感知网络模型等，本文使用的注意力得分通过点积计算，计算公式为：

$$s(q,k_i) = similiarity(Query, Key_i) = Query \cdot Key_i \tag{2.9}$$

通过 *soft* max 函数对得到的注意力得分进行归一化处理，计算权重系数的公式为：

$$a_i = soft\max(s(q,k_i)) = \frac{\exp(s(q,k_i))}{\sum_{j}^{n}\exp s(q,k_j)} \tag{2.10}$$

最后根据权重系数对 Value 值进行加权处理，得到 Attention 值，计算 Attention 值的公式为：

$$Attention(Query,Key,Value) = \sum_{i=1}^{n} a_i \cdot Value_i \tag{2.11}$$

2.4.4 Transformer 模型

Transformer 是在 2017 年由 Google 公司团队提出的[68]，面向机器翻译、文本分类等自然语言处理领域的神经网络模型。Transformer 的核心是它本身具有的自注意力机制，自注意力机制是一种常用于序列建模的注意力机制，能够在处理序列数据的同时考虑序列中不同位置的信息。与传统的注意力机制不同，自注意力机制只涉及到序列内部的信息交互，不需要依赖外部的信息。相比于传统的 RNN 和 LSTM 需要按顺序逐步处理序列数据，Transformer 不涉及任何序列顺序的处理，这使得 Transformer 具有良好的并行性和出色的性能。当下，Transformer 模型已经成为了一种处理序列数据任务的重要且有效的神经网络架构。Transformer 模型结构图如图 2.5 所示，模型主要由编码器和解码器组成。

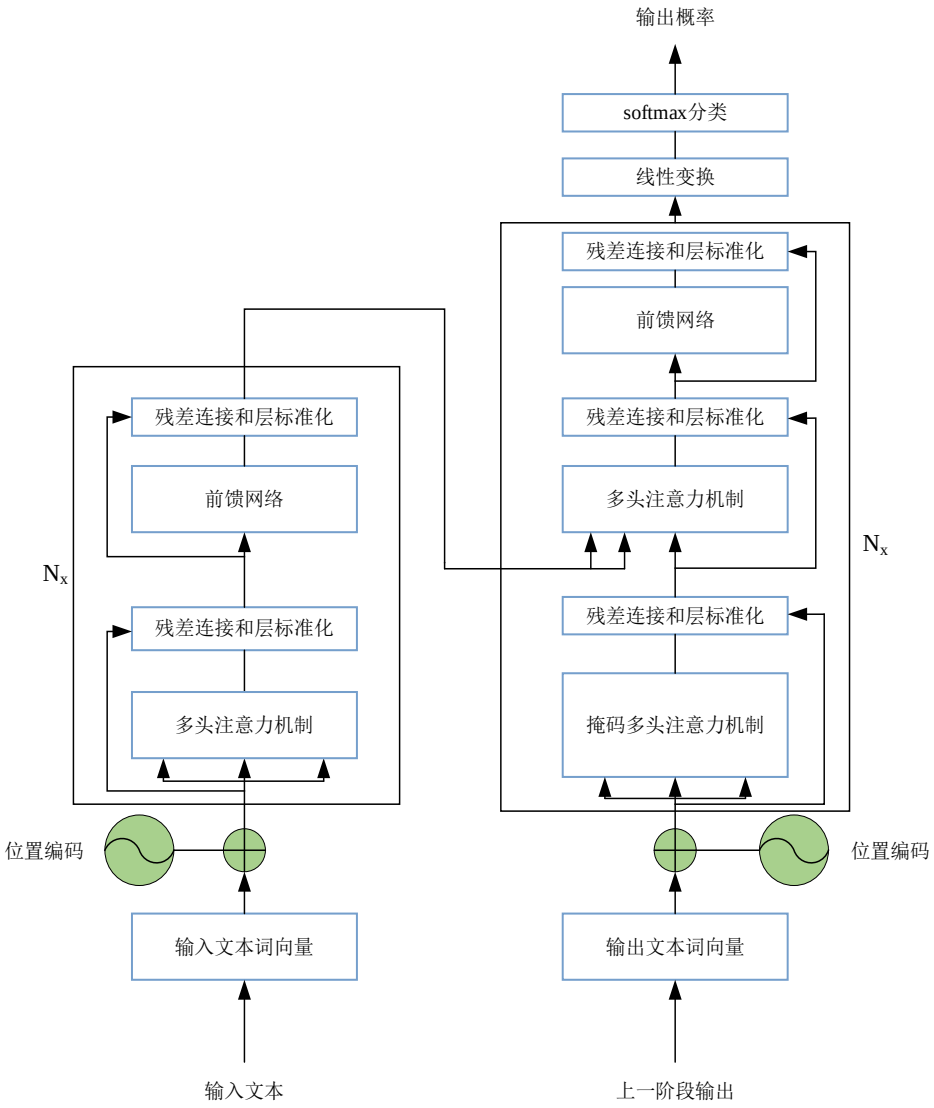

图 2.5 Transformer 模型结构

Transformer 的多头注意力机制将输入向量分成多个头来计算多个独立的注意力分布，以提高模型的表示能力。每个注意力头都有三个输入向量：Query 向量、Key 向量和 Value 向量。对于第 i 个头，其注意力计算公式为：

$$Head_i(Query, Key, Value) = Soft\max(\frac{Query_i \cdot Key_i^T}{\sqrt{d_k}}) \cdot Value_i \qquad (2.12)$$

$$Query_i = W_i^Q X$$

$$Key_i = W_i^K X$$

$$Value_i = W_i^V X$$

其中W_i^Q，W_i^K，W_i^V是将输入向量X线性转换成 Query 向量、Key 向量和 Value 向量的权重矩阵，d_k是注意头的维度。

最后将所有头的注意力拼接在一起，经过线性变换得到多头注意力值，多头注意力计算公式为：

$$MultiHead(Query,Key,Value)=Concat(head_1,head_2,...,head_n)W^o \quad (2.13)$$

其中W^o是输出向量的权重矩阵。

Transformer 编码器的主要功能是对输入文本进行指定的特征抽取，为解码器提供有效的语义信息支持。编码器由 6 个相同的层堆叠而成，每个层都包含一个多头注意力模块和一个前馈神经网络模块。为了防止梯度消失问题并增强模型性能和收敛速度，在每个模块后面都引入了残差连接和层标准化操作。

编码器经过多头注意力机制计算得到矩阵 X 后，将新得的矩阵经过残差连接和层标准化操作得到输出x，计算公式如下：

$$x=LayerNorm(X+sublayer(X)) \quad (2.14)$$

前馈神经网络由两个全连接层组成，通过激活函数连接。将得到的输出x作为前馈神经网络FFN的输入，前馈网络计算公式如下:

$$FFN(x)=\max(0,xW_1+b_1)W_2+b_2 \quad (2.15)$$

其中W_1，W_2是权重矩阵，b_1，b_2是偏置。

Transformer 解码器的作用在于从编码的表示中检索信息，同样有 6 个功能相似的层，除了比编码器多了一个掩码多头注意力机制外，其余的结构与编码器相同，这是因为解码器的输出带有时序效果，掩码多头注意力机制能够防止文本序列的某个位置提前得到后面位置的信息。

2.4.5 BERT 预训练模型

BERT 是在 2018 年由 Google 公司提出的一种基于 Transformer 架构的预训练语言模型[69]。BERT 的主要思想是先在大规模未标注的语料库上进行预训练，学习通用的语言表示，再使用少量的有标注数据进行微调训练。现已被广泛应用

于情感分类、实体识别、关系抽取等自然语言处理任务中。BERT 预训练模型结构如图 2.6 所示。

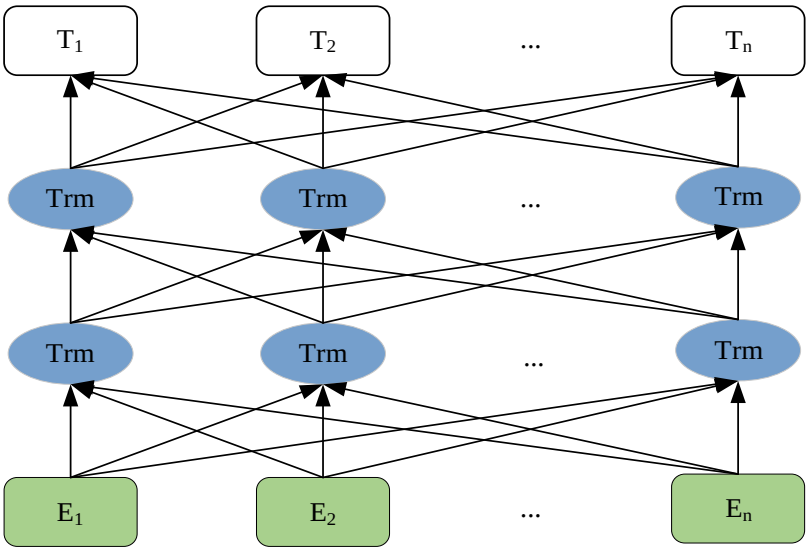

图 2.6 BERT 预训练模型结构图

BERT 的模型结构主要有三个模块：最底层的嵌入(Embedding)模块，中间层的 Transformer 模块和最上层的预微调模块。每个嵌入层由字符嵌入(Token Embedding)、分段嵌入(Segment Embedding)和位置嵌入(Position Embedding)组成。其中，字符嵌入将输入的句子转换成多个向量，BERT 模型会自动对输入的句子添加“[CLS]”与“[SEP]”来标记句子的开头与末尾，以此来划分不同的句子。分段嵌入用于区分句子的类型。位置嵌入将字符的位置信息表示成特征向量，解决字符相同但位置不同导致最终输出相同的问题。嵌入层结构图如图 2.7 所示。

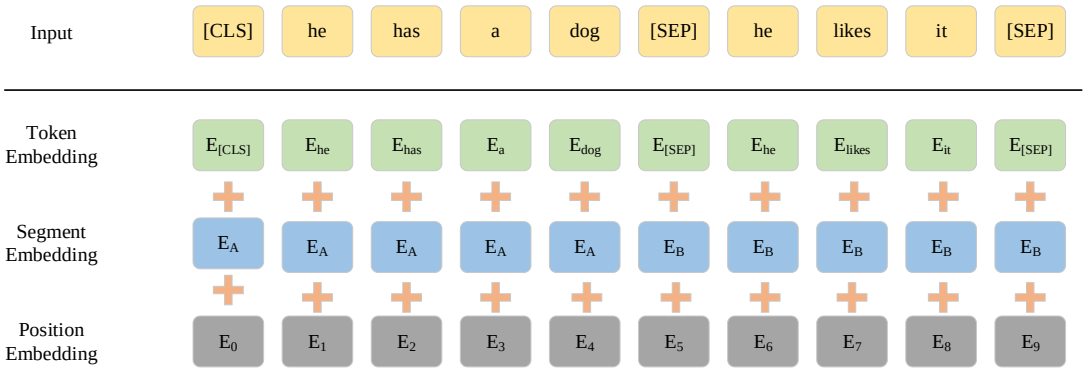

图 2.7 嵌入层模型图

BERT 模型在预训练阶段中主要有两个任务：掩码语言模型(Masked Language Model，MLM)和下句预测(Next Sentence Prediction，NSP)。掩码语言模型在单词序列输入之前随机掩盖 15%的字符，然后结合单词上下文预测这些被掩盖的词以获取深度语义表征能力。其中在这些被掩码的 15％字符中，80％的字符被[MASK]标记替换，10％的字符被随机替换成别的字符，剩下的字符保持不变。

下句预测任务旨在让模型理解文本中不同句子之间的关系，从而提供更好的句子级别的语义表示。首先在训练过程中挑选一对文本句子 A 和 B，其中，有 50％的概率挑选的句子 A 和 B 是两个连续的语句。然后模型同时观察 A 和 B 两个句子，以便正确地预测 B 是否是 A 的下一个句子。这两个任务同时进行训练后，可以将输出的预训练模型微调并迁移至特定的任务上，如文本分类、命名实体识别、问答等，以适应具体的任务要求。

2.5 本章小结

本章首先对知识图谱构建过程进行概述并对知识图谱存储 Neo4j 图数据库进行介绍，然后对构建技术中命名实体识别和实体关系抽取技术进行详细阐述，随后介绍了知识图谱构建过程中涉及到的相关神经网络模型，包括条件随机场、双向长短时记忆网络、注意力机制、Transformer 模型和 BERT 模型，为本文模型的构建奠定了基础。

3 融合注意力机制的命名实体识别模型

通过前文对知识图谱构建相关技术的介绍可知，命名实体识别是知识图谱构建的基础，也是其最重要的一环。实体是知识图谱构建的基本元素，其后续关系识别，知识存储的使用基础便是拥有一个意义明确的实体。本章主要研究内容为构建祁连山可持续发展数据集，并基于 BERT-BiLSTM-CRF 命名实体识别模型，提出融合注意力机制的 ALBERT-BiLSTM-Attention-CRF 轻量化模型,最后在构建的数据集上进行实验对比。

3.1 祁连山可持续发展数据集构建

3.1.1 数据来源与获取

目前在祁连山可持续发展领域中还没有公开的数据集用于实体识别和知识图谱构建，因此本文自行构建祁连山可持续发展数据集用于命名实体识别实验及后续的知识图谱构建工作。选择合适、优质的数据源是构建高质量知识图谱的前提，本文选取了中国知网数据库中公开发表的中文期刊论文作为主要数据源，这些论文文献高度凝练了祁连山可持续发展领域的知识要点，知识权威性较高，并且在中国知网数据库中具有较为规则的数据形式。

在确定了祁连山及祁连山可持续发展相关信息为主题的基础上，本文在中国知网数据库上以题目、关键词和主题词作为检索范围，进行检索并初步筛选符合要求的科技文献，将筛选出的文献相关信息导出为 excel 文件保存，由于摘要信息包含了几乎所有关键点，使用摘要基本不会遗漏重要信息，因此本文并未对文献全文进行导出。通过阅读题目和摘要删除掉不相关文献后，对剩余文献进行全篇阅读与摘要进行比对，防止遗漏重要信息，最终得到 13198 条原始数据，最后对保留的文献进行处理转换为 txt 文本以便于进行数据标注工作。获取的数据示例如表 3.1 所示。

表 3.1 数据示例

字段	数据示例
src	期刊
title	黑河中游绿洲农区地下水硝态氮污染调查研究
author	杨荣;苏永中
org	中国科学院寒区旱区环境与工程研究所
source	冰川冻土
keyword	地下水;硝酸盐;土壤;土地利用类型
summary	通过区域采样的方法,对黑河中游绿洲农田灌区 71 眼水井…
year	2008

3.1.2 实体定义

通过对相关生态环境领域知识图谱的研究与分析，并结合祁连山可持续发展文献文本特点和联合国可持续发展目标(SDGs)中有关生态领域的 SDGs2.3 农业生产力、SDGs2.4 可持续农业面积、SDG6.4.1 水生产力、SDG6.4.2 水压力水平、SDGs6.6 保护和恢复与水有关的生态系统、SDGs 11.6 负面环境影响-污染物、SDGs15.1.1 森林覆盖率、SDGs15.4.2 绿色覆盖指数等 8 个可持续发展目标，归纳整理了所构建数据集中 6 种实体类别，作为构建祁连山可持续发展数据集实体标注规范，具体如表 3.2 所示。

表 3.2 祁连山可持续发展实体类别

实体类别	相关词语
地理	山脉、河流、湖泊、城市
生物	生物、植物、动物、人类
生态系统	森林、草原、湿地、山地、冰川、土地
环境因素	农业、土壤、环境、水体、大气、沉积、气候、污泥、地表、水库、降水、径流、沙漠化、地表水、地下水、水资源

续表 3.2

污染物	金属、氮磷、风险、污水、废水、颗粒、臭氧、污染、化学、污染物、化合物、抗生素、重金属
评价标准	含量、指数、指标、压力、浓度、评价、水平、质量、活性、水质、表层、能力、粒径、功能、强度、通量、均值、梳理、性质、特性、产量、面积、覆盖率、生产力、生物量、排放量

3.1.3 实体数据标注

基于有监督的深度学习模型需要大量标注数据来训练模型，使其达到更优的性能。因此标注文本数据中的实体对于命名实体识别任务至关重要。因为数据量较大，为了提高标注效率，本文使用数据标注平台 Doccano 进行人工辅助标注工作。Doccano 是一款基于 web 端的开源标注工具，用于文本、命名实体识别、关抽取等自然语言处理任务的数据标注。它通过 web 界面进行使用和管理，使得标注人员可以轻松地对文本进行标注，并生成高质量的训练数据集。除了标注功能，Doccano 还提供了数据导入、导出、版本控制等功能，方便用户管理和共享标注数据，还可以将导出的标注数据经过 Doccano 官方工具 doccano-transformer 转换为通用数据集格式，以便后续的模型训练和评估。标注工作首先将处理好的祁连山可持续发展文本数据导入到 Doccano 平台中，导入数据后创建祁连山可持续发展数据集的 6 种实体类别，进行数据标注工作，标注数据示例如图 3.4 所示。

图 3.1 标注数据示例

3.1.4 数据集构建

通过上文对原始数据进行获取并完成预处理工作后，在标注阶段人工对原始数据进行标注和筛选，最终保留 4000 条人工标注数据用于本文命名实体识别模型实验，其中每条数据都包含有若干实体。祁连山可持续发展数据集的实体标注结果统计情况如表 3.3 所示。

表 3.3 实体标记结果统计

实体类别	数量
地理	3013
生物	1022
生态系统	4599
环境因素	5663
污染物	795
评价标准	1901

3.2 BERT-BiLSTM-CRF 模型

命名实体识别是本文构建祁连山可持续发展知识图谱的基础，是最重要的步骤，其在自然语言处理领域中的重要性也使得与其相关的研究方法陆续出现，其中 BERT-BiLSTM-CRF 模型是现阶段主流的实体识别模型，其在命名实体识别任务中有着出色的表现。该模型能够更好地处理复杂的文本序列，更好地处理实体中出现的一词多义现象，从而提高命名实体识别的准确度和泛化能力。BERT-BiLSTM-CRF 模型整体的框架结构如图 3.2 所示。

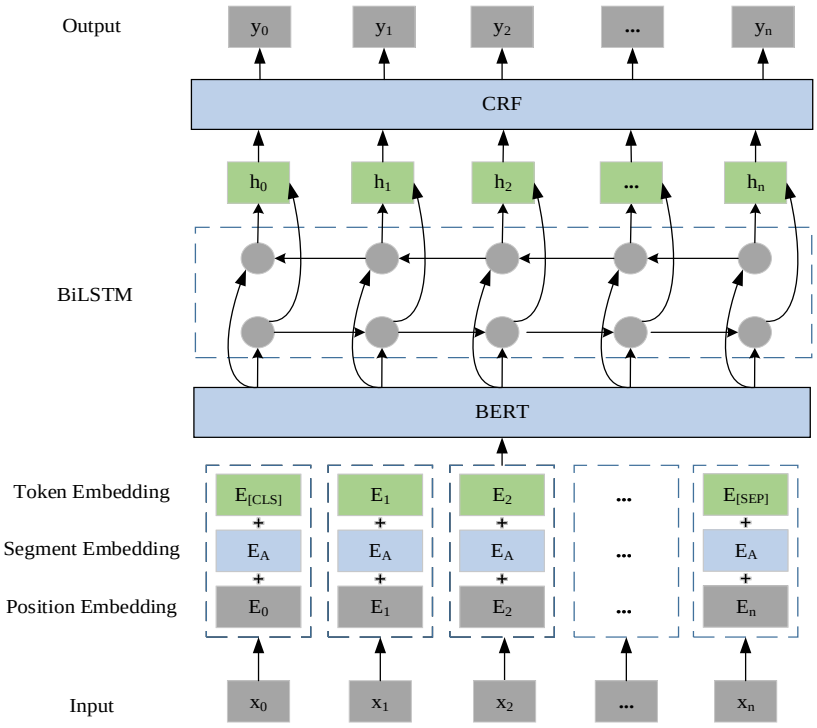

图 3.2 BERT-BiLSTM-CRF 模型结构图

从结构图中可知，BERT-BiLSTM-CRF 模型主要包含了三个模块，分别是：BERT 嵌入层，双向长短时记忆网络(BiLSTM)特征提取层以及 CRF 层。该模型首先将文本数据经过预处理后输入 BERT 层，由 BERT 作为嵌入层处理语言文本生成动态词向量，然后将 BiLSTM 作为特征提取层对词向量进行双向训练，进一步提取文本字符间的关系特征，获取实体识别双向依赖的特征编码，最后使用 CRF 层约束预测标签之间的依赖关系，输出全局最优的标签序列。

3.2.1 BERT 嵌入层

对于输入文本序列 $X=\{x_1,x_2,...,x_n\}$，模型将 X 输入 BERT 嵌入层中，在嵌入层会分别生成字符向量(Token Embedding)、句子向量(Segment Embedding)和位置向量(Position Embedding)三个词向量。具体过程为：

Token Embedding 层的作用是将文本序列中每一个字符转换为具有固定维度的向量表达形式。在进行字符向量化之前，需要对输入文本进行预处理，即将在文本的开头和结尾插入两个特殊的符号[CLS]和[SEP]，之后再输入到 Token Embedding 层，在该层中每个字符都将被转换为一个 768 维的向量表示。

Segment Embedding 层对输入的两个文本语义是否相似进行判断分类，从而将这对输入文本连接并输入到模型中。为实现这一连接，Segment Embedding 层使用了两种向量表示方式。其中第一个向量将第一个句子中的每个 Token 赋值为 0，第二个向量则将第二个句子中的每个 Token 赋值为 1。需要注意的是，如果输入只有一个句子，则其赋值将全被设置为 0。

Position Embedding 层的作用是对文本序列中的每个字符进行位置编码，通过让模型学习每个字符位置的向量表示来包含输入序列的顺序特征。共有两种编码方式，一种是绝对位置编码，另一种是相对位置编码，本文模型采用的是绝对位置编码，直接对不同的位置随机初始化位置编码，之后这个位置编码作为参数进行训练。

最后将上述三个词向量拼接起来就是嵌入层的最终输出，同时也是特征提取层的输入向量。

3.2.2 BiLSTM 特征提取层

模型中的特征提取层采用的是双向长短期记忆网络(BiLSTM)，BiLSTM 是一种用于捕捉文本双向依赖的模型，是一种特殊的循环神经网络。由于 LSTM 模型只能单向的从前往后学习语义的信息特征，无法处理后向语义信息，而在序列标注任务中，实体与上下文均存在强烈的依赖关系，因此 BiLSTM 模型被提出用于更好的捕捉双向的语义依赖。BiLSTM 模型由前向的 LSTM 和后向的 LSTM

组合而成，通过对文本的双向学习捕捉到双向依赖，充分考虑了输入语句的上下文信息。BiLSTM 模型结构如图 3.2 所示。

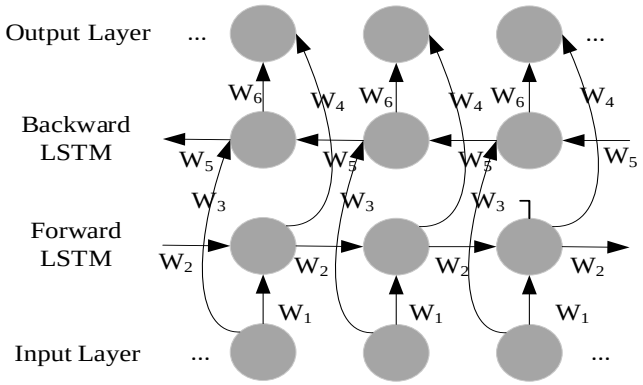

图 3.3 BiLSTM 模型结构图

在前向 LSTM 层中正向计算从 1 时刻到 t 时刻的向前隐层输出，在后向 LSTM 层中沿着 t 时刻到 1 时刻反向计算向后隐层输出，最后在每个时刻结合前向 LSTM 和后向 LSTM 的输出结果进行向量拼接得到最后输出。计算公式如下：

$$h_t = f(w_1 x_t + w_2 h_{t-1}) \tag{3.1}$$

$$h_t^{'} = f(w_3 x_t + w_5 h_{t-1}^{'}) \tag{3.2}$$

$$o_t = g(w_4 h_t + w_6 h_t^{'}) \tag{3.3}$$

在 BERT-BiLSTM-CRF 模型中，经过 BERT 嵌入层生成的词向量通过特征提取层 BiLSTM，将前向和后向的输出向量拼接在一起，得到一个包含更多上下文语境信息的输出向量，从而提高了模型的性能和泛化能力。

最后将 BiLSTM 层输出向量输入到条件随机场(CRF)中，输出符合标注转移约束条件的、最大可能的预测标注序列。在 BiLSTM 层后加入一个 CRF 层是因为在命名实体识别任务中，CRF 通过相邻标签的关系从而得到最优的预测序列，弥补了 BiLSTM 在处理短距离文本信息时无法很好预测其相邻标签的关系的问题。

3.3 ALBERT-BiLSTM-Attention-CRF 模型

虽然 BERT 模型具有强大的语言表示学习能力，在文本分类、实体识别和语义相似度计算等自然语言处理任务中取得了非常好的效果，但其本身效果依赖于大量的模型参数。随着模型规模的持续增大，数亿甚至数十亿的参数量随之出现，这导致了 BERT 模型的训练需要大量的计算资源、时间和大规模的语料库，训练成本非常高。

在 BERT-BiLSTM-CRF 模型中，除了 BERT 模型本身就需要大量的计算资源和时间进行训练外，BiLSTM 层和 CRF 层也增加了模型的复杂度和计算开销，这导致了模型有着更高的训练成本。同时 BiLSTM 层虽然可以更好的结合上下文，有效的利用输入的前向和后向特征信息，但对于过长的序列依然没法很好地传输序列起点的信息。因此本文针对上述问题，提出了融合注意力机制的轻量化模型 ALBERT-BiLSTM-Attention-CRF。在轻量化模型中 ALBERT 层大大减少了参数量，使得模型在训练时具有更高的效率和更快的速度，Attention 层提高了对实体分类的准确性。轻量化模型 ALBERT-BiLSTM-Attention-CRF 的结构如图 3.3 所示。

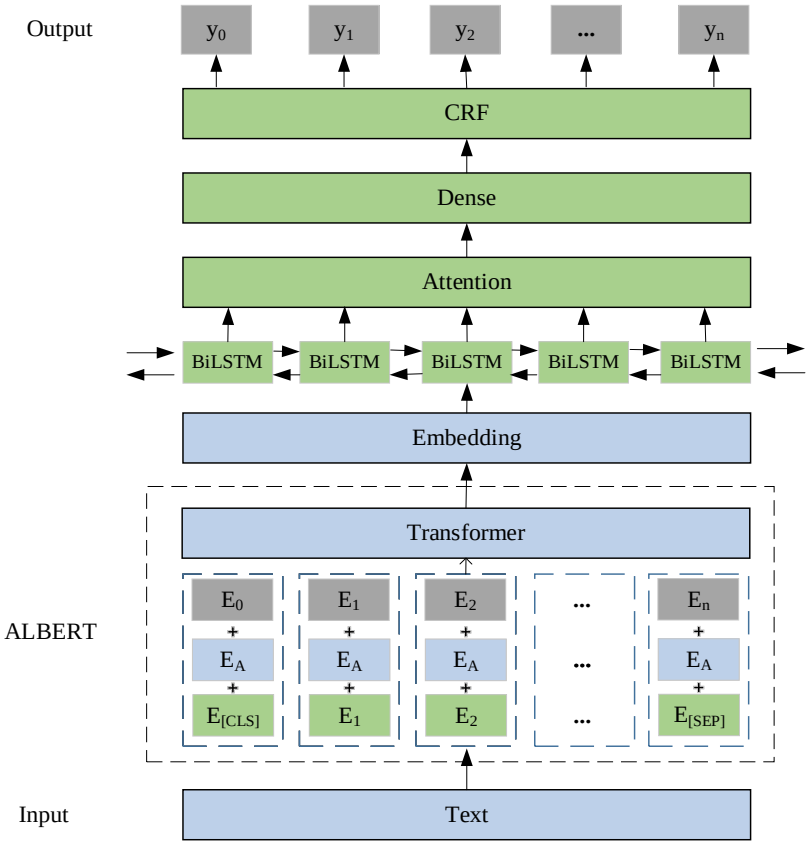

图 3.4 轻量化模型结构图

3.3.1 ALBERT 嵌入层

为了解决 BERT 本身存在的问题，谷歌公司于 2019 年提出了一种轻量级的 BERT 模型，并取名为 ALBERT(A Lite BERT)。ALBERT 模型在 BERT 的基础上大幅降低其参数规模，所占用的显存空间可以减少至 BERT 的十分之一乃至更少，在不损失训练准确度的前提下加快了训练的速度，更加方便的进行预训练语言模型的训练与部署。ALBERT 模型架构与 BERT 类似，都使用了基于 GELU 非线性激活函数的 Transformer，但与 BERT 不同的是，ALBERT 模型采用了两种参数精简的技术减少了参数量，这些技术可以消除扩展预训练模型时遇到的主要障碍。ALBERT 模型改进主要有以下几点：

（1）嵌入层参数因式分解

在 BERT 中，嵌入层的输出向量维度 E 和 Transformer Block 层输出的向量维度 H 默认相同，但在实际任务中，嵌入层只作为一个静态映射，词的向量化过程不涉及上下文信息，因此向量维度 E 不需要随着 H 的增大而增大，向量维度 E 的增大会造成参数冗余。针对这一现象，在 ALBERT 模型中对向量维度 E 进行了固定，并引入了一个变换矩阵，该矩阵将维度为 E 的嵌入层输出向量变换为维度为 H 的向量，以便输入至 Transformer Block 层。设模型词表大小为 V，这个变换可以视为如下的因式分解操作：

$$V \times H = V \times E + E \times H \tag{3.4}$$

通过上述的因式分解操作后，使得 ALBERT 模型嵌入层需要的参数量远远少于 BERT 模型。

（2）跨层参数共享

传统的 Transformer 中的每一层参数都是独立的，包括各层中的注意力机制和全连接，这样在层数增加时，参数量也会大幅增加，因此需要减小参数规模，对参数进行共享。在 Bert 中对 Transformer 的共享参数方案有只共享全连接层或只共享注意力层两种，而 ALBERT 中结合了这两种方法，在全连接层和注意力层都进行参数共享，使得隐藏层的参数量变为原来的 1/12 或 1/24，解决了参数规模不断随网络深度增加而增加的问题，在稳定了网络参数的同时进一步减小参数规模。

（3）SOP 方法

在 ALBERT 模型中引入了一种改进的下一句预测方法，称为自监督的句子顺序预测损失(Sentence-Order Prediction, SOP)。BERT 中的 NSP 任务类似于语义匹配，预测两个语句在文本中是否连续出现，NSP 任务可以提高下游任务的性能，但有的研究表明 NSP 任务也是可以去掉的，去掉之后反而也可提高性能，因此 NSP 任务对模型的影响并不可靠。同时 NSP 任务与 MLM 任务相比过于简单，学习到的内容相对有限，NSP 中将句子的主题预测和连贯性预测同时进行，然而与一致性预测相比，主题预测更容易学习，也与使用 MLM 任务学习到的内容重叠。因此，Albert 中提出了使用句子顺序预测损失 SOP 来代替 NSP 任务，这种方法避免了主题预测，关注于句子之间的连贯性而不是句子之间的匹配性，SOP 的正样本同样从原始语料中获得，但负样本为原始语料中的两个句子交换顺序，迫使模型更加关注于句子之间的连贯性信息，使得 SOP 的任务更难，学习到的信息更丰富。

（4）去除 Drop out

在 Albert 模型长期的训练中，过拟合现象都未曾出现，因此 Albert 模型中去除了 Drop out，去除后显著优化了在训练当中的内存占用，加快了训练速度。这是因为 Drop out 会产生许多临时变量，而这些变量对于 ALBERT 来说是没有用处的。

在本文模型 ALBERT 嵌入层中，ALBERT 通过参数因式分解和跨层参数共享，大大减少了模型的参数量。对于输入文本序列 $X=\{x_1,x_2,...,x_n\}$，在嵌入层经过 ALBERT 中 MLM 机制和 SPO 方法处理后，生成字符向量、句子向量和位置向量三个词向量。最后将三个词向量拼接起来作为特征提取层的输入。

3.3.2 特征提取层

在本文模型架构中，特征提取层由 BiLSTM 层、注意力层和全连接层组合而成。在特征提取层，BiLSTM 能够有效地捕捉输入序列中的长距离依赖关系，它被用来处理 ALBERT 输出的嵌入表示，进一步学习上下文信息，而注意力机制允许模型关注输入序列的不同位置信息，从而有助于模型更好地捕捉全局依赖关

系，因此在特征提取层中注意力层被应用于 BiLSTM 的输出，以强调对当前预测任务更为重要的信息。全连接层用于将注意力层的输出转换为最终标签空间的维度输出到 CRF 层中。对第 i 个词，首先将嵌入层输出向量作为输入经过 BiLSTM 模型可以得到的两个隐层向量，假设这两个向量分别为 h_1 和 h_2，h_1 是正向处理系列，h_2 为反向处理系列，处理完成后将 h_1 和 h_2 拼接在一起，得到最终的 BiLSTM 输出结果 $h_i=[h_1,h_2]$ 向量。然后将输出向量 h_i 作为注意力层的输入，通过全连接层将 h_i 转换为 k_i，k_i 计算公式为：

$$k_i=\tanh(Wh_i+b) \tag{3.5}$$

得到 k_i 后，计算 k_i 与上下文向量 q_i 的相似度，q_i 是由随机初始化并通过训练获得，然后通过 *soft* max 函数进行归一化处理，计算出权重系数，计算公式为：

$$a_i=\frac{\exp(q_i k_i^T)}{\sum_i (q_i k_i^T)} \tag{3.6}$$

a_i 可以看做是每个单词的句子的重要程度，最后使用 a_i 作为全局对 h_i 加权求和得到注意力层的输出，计算公式为：

$$s_i=\sum_{i=1}^{n} a_i h_i \tag{3.7}$$

通过注意力机制，模型能够对输入序列中的不同位置进行加权，提高对关键信息的关注度，解决了 BiLSTM 对于过长的序列没法很好地传输序列信息的问题。

最后本文模型结构中的 CRF 层能够有效约束预测标签之间的依存关系，通过 CRF 计算得出的条件概率约束能够充分保证预测的合法性，弥补了特征提取层中 BiLSTM 模型无法处理相邻标签之间依赖关系的缺点，最终输出最优的预测标注序列。在本文模型进行实验时，模型将使用 CRF 定义的损失函数和准确率指标进行训练和评估。

3.4 实验与分析

3.4.1 实验数据

在进行实验时，将已经进行预处理和标注的数据集按照 4:1 的比例划分为训练集和测试集，3200 条文本用于训练，800 条文本用于测试。此时划分好的数据集还需要经过标注序列转换才能用于命名实体识别模型实验，本文选用目前常用的 BIO 序列标注方法对训练集和测试集进行转换。BIO 方法将文本中每个字元素分别标注为“B-XX”、“I-XX”或者“O”。其中“B-XX”表示为该字元素是 XX 类型实体的开头，“I-XX”表示该字元素是 XX 类型实体的中间或结尾位置，“O”则表示文本中非实体的部分。最后将经过 BIO 序列标注方法转换的数据集作为实验数据，序列标注示例如表 3.4 所示。

表 3. 4 实验数据标注示例

文本	标注	文本	标注	文本	标注
在	O	、	O	站	I-地理
黑	B-地理	祁	B-地理	逐	O
河	I-地理	连	I-地理	日	O
上	I-地理	及	O	采	O
游	I-地理	莺	B-地理	集	O
扎	B-地理	落	I-地理	河	B-环境因素
马	I-地理	峡	I-地理	水	I-环境因素
什	I-地理	水	I-地理	样	I-环境因素
克	I-地理	文	I-地理	，	O

3.4.2 实验评价指标

深度学习模型在不同任务中的表现需要用评价指标进行实验，在实体识别任务中，评价指标能够更好地反映和评价本文模型对于实体识别的影响。本文在实验中通过精准率、召回率、F1 值进行评估和检测模型的性能。

精准率(Precision, P)表示预测结果与标记结果均为真的样本数量占全部预测

为真样本数量的比例，用来衡量模型预测的准确性。计算公式为：

$$p = \frac{TP}{TP + FP} \tag{3.8}$$

其中 TP 表示预测结果与真实标记结果均为真的样本数量；FP 表示预测结果为真但真实标记结果为假的样本数量。

召回率(Recall, R)表示预测结果与标记结果均为真的样本数量占全部标记结果为真样本数量的比例。计算公式为：

$$R = \frac{TP}{TP + FN} \tag{3.9}$$

其中 TP 的含义与精准率公式定义相同，FN 表示真实标记结果为真但预测结果为假的样本数量。

F1 指标是精准率和召回率的调和均值，用来整体兼顾准确率和召回率效果的指标，计算公式为：

$$F1 = \frac{2 * P * R}{P + R} \tag{3.10}$$

3.4.3 实验环境及参数设置

本实验训练通过 AutoDL 算力云平台，该平台是一款云 GPU 深度学习环境出租平台。所用的服务器配置：GPU 为 RTX 3090、24GB 显存；CPU 为 Intel(R) Xeon(R) Gold 6330，具有 Cuda11.2 的加速框架，80G 内存。深度学习配置的环境为 TensorFlow 2.9.0、Python 3.8，编码格式统一为 utf-8，详细参数设置如表 3.5 所示。

表 3.5 实验参数设置

参数名称	参数值
Batch size	32
Epoch	100
Learning rate	2e-5
Optimizer	Adam
Dropout	0.4

续表 3.5

Embedding dim	128
LSTM hidden size	64

3.4.4 实验结果与分析

为了验证本模型的效果，选取了几种主流的命名实体识别模型进行对照实验，分别为 BERT-CRF 模型、采用 Word2Vec 词向量的 Word2Vec-BiLSTM-CRF 模型、采用 BERT 预训练模型的 BERT-BiLSTM-CRF 模型以及本文融合了注意力机制的 ALBERT-BiLSTM-Attention-CRF 轻量化模型，训练数据均为本文经过 BIO 标注方案转换的祁连山可持续发展数据集。对比结果如表 3.6 所示。

表 3.6 对比实验结果

模型	P/%	R/%	F1/%
BERT-CRF	80.40	78.02	79.25
Word2Vec-BiLSTM-CRF	84.65	79.62	82.10
BERT-BiLSTM-CRF	82.32	82.04	82.24
ALBERT-BiLSTM-Attention-CRF	86.04	82.17	84.11

从表中数据可以看出，Word2Vec-BiLSTM-CRF 模型、BERT-BiLSTM-CRF 模型和 ALBERT-BiLSTM-Attention-CRF 模型的准确率、召回率、F1 值皆明显高于 BERT-CRF 模型。通过对 BERT-CRF 模型与 BERT-BiLSTM-CRF 模型进行对比，发现在增加 BiLSTM 模块后，精准率、召回率和 F1 值均有提升，这是因为 BERT 模型中使用了 Transformer 架构，而 Transformer 的自注意力机制弱化了位置信息，BERT 模型中的位置信息仅依靠 Position Embedding，但在序列标注、实体识别任务中位置信息很重要，BiLSTM 的作用是学习到输入序列的上下文依赖，因此 BiLSTM 模块的增加能有效提升特征提取的效果，算法能够发现更多的祁连山可持续发展相关的实体，并且对这些实体的判断准确率也有提升。

从 Word2Vec-BiLSTM-CRF 模型与 BERT-BiLSTM-CRF 模型可以看出，词

向量编码模块换为 BERT 后，召回率提升比较明显，基于 BERT 进行词向量编码的命名实体模型能发现到更多的命名实体，但是识别准确率由于识别到更多的实体而有所降低，但总体来说基于 BERT 词向量编码的命名实体模型相比于基于 Word2vec 词向量编码的模型来说性能要好。这说明 BERT 预训练模型能够更好的表达词语的特征信息。

在 BERT-BiLSTM-CRF 模型和 ALBERT-BiLSTM-Attention-CRF 模型中，会发现使用轻量化 ALBERT 模型的本文模型精准率、召回率和 F1 值均有提升，这说明在本文的命名实体语料规模相对较小的情况下，ALBERT 以其参数少的优势，能够更快的达到模型泛化的效果。

为了进一步研究模型改进的有效性，本文将在数据集上进行消融实验，主要验证本文模型特征提取层对模型性能的影响，对比模型为 ALBERT-BiLSTM 模型、去掉注意力机制的 ALBERT-BiLSTM-CRF 以及本文 ALBERT-BiLSTM-Attention-CRF 模型，实验结果如表 3.7 所示。

表 3.7 消融实验结果

模型	P/%	R/%	F1/%
ALBERT-BiLSTM	80.64	78.53	79.68
ALBERT-BiLSTM-CRF	84.59	82.10	83.36
ALBERT-BiLSTM-Attention-CRF	86.04	82.17	84.11

从表中数据可以看出，本文提出的模型精准率、召回率、F1 值均高于另外两个对比模型，这说明在实体识别任务中，在特征提取层 BiLSTM 的基础上融合注意力机制有助于提高对输入序列不同位置的信息关注度，能够有效提升模型性能。

3.5 本章小结

本章提出了一种融合注意力机制的 ALBERT-BiLSTM-Attention-CRF 模型，用来提高祁连山可持续发展实体识别的准确率。模型在 BERT-BiLSTM-CRF 的

基础上将嵌入层由 BERT 改为轻量化模型 ALBERT，在特征提取层 BiLSTM 的后面加上一个注意力层，减少了模型的参数量并提高了模型对输入文本序列位置信息的关注度。通过归纳确立祁连山可持续发展实体类别后，在本文自行构建的数据集上进行对比实验和消融实验，实验结果证明本文模型相比其他对比模型在准确率、召回率和 F1 值上都得到了提升，构建的特征提取层对模型性能提高有较好的作用。

4 知识图谱构建与存储

知识图谱可被描述为一个由节点和边构成的图结构。其中节点代表着实体，实体间的相互关系通过边来表示。通过<节点，边，节点>的关联，知识图谱能够以结构化和语义化的方式来表示现实世界中的实体及其关系，进而实现知识的存储、查询、推理和应用。在知识图谱的构建工作中，命名实体识别和实体关系抽取是其最重要的任务，在前文中完成了命名实体识别任务，在本章完成实体关系定义工作，并基于构建框架，通过 Neo4j 图数据库完成知识图谱的构建和存储可视化工作，最后通过 CiteSpace 软件对祁连山可持续发展文献数据的研究热点进行了分析。

4.1 祁连山可持续发展知识图谱构建框架

本文研究构建的祁连山可持续发展知识图谱，主要由前期的数据准备工作和技术构建两部分构成，前期的数据准备工作包括制作数据集以及实体和关系定义，通过规范目标领域内的实体和不同实体之间的关系，明确祁连山可持续发展数据特征、可持续发展概念以及联合国可持续发展目标(SDGs)，借此梳理得到祁连山可持续发展的实体与关系；然后技术构建部分针对获取的祁连山可持续发展文本数据应用信息抽取技术提取实体和关系信息组建<实体，关系，实体>三元组，并将三元组信息存入图数据库 Neo4j 中，以三元组的形式表征知识结构，完成知识图谱的构建。构建过程具体可划分为数据获取、实体关系定义、信息抽取、知识存储 4 个部分。

（1）数据获取。将中国知网数据库中公开发表的中文期刊论文作为主要数据源，对祁连山可持续发展的相关数据信息进行获取与预处理。在本文第三章完成了祁连山可持续发展数据集的构建工作。

（2）实体关系定义。依据祁连山可持续发展数据特征和可持续发展概念，明确定义实体、关系以及实体关系对应信息，构建一个能够反映祁连山及其可持续发展信息现状的实体和关系。

（3）信息抽取。第一步命名实体识别，基于定义的实体类别信息，通过人

工借助标注工具的方式对经过预处理的数据集进行标注，将标注好的数据作为输入，使用本文提出的命名实体识别模型抽取出该文本语料中自定义的命名实体，将其标注为知识图谱中的实体节点。第二步关系抽取，根据抽取的实体词语特征，根据定义的关系和实体关系对应信息，通过规则模板和人工标注结合的方式完成实体关系抽取，并完成三元组的构建。

（4）知识存储。在经历了以上的步骤之后，得到祁连山可持续发展的结构化知识，使用 Neo4j 图数据库将获得的三元组知识进行存储和更新，从而实现图谱存储的持久性，最后完成祁连山可持续发展知识图谱的构建。本文构建祁连山可持续发展知识图谱流程如图 4.1 所示。

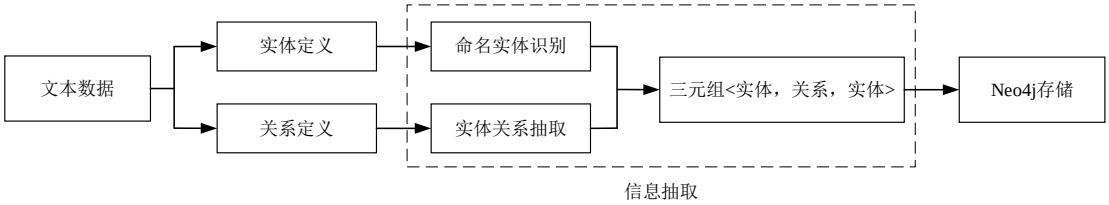

图 4. 1 知识图谱构建框架

4.2 祁连山可持续发展知识图谱关系定义

定义知识图谱中的实体与关系的类型是构建知识图谱的基本框架，决定了知识图谱的发展方向与质量，是知识图谱构建过程中十分重要的环节。本文在第三章根据祁连山可持续发展数据特征和联合国可持续发展目标(SDGs)中有关生态领域的 8 个可持续发展目标定义了祁连山可持续发展知识图谱的 6 种实体类别。基于 6 种实体关系，本文关系定义研究根据张华[40]的生态环境领域知识图谱模式层构建方法以及王天一的生态环境领域知识图谱构建方法[70]，并在此基础上结合 8 个可持续发展目标进行改进，最后定义了祁连山可持续发展知识图谱的 6 种关系信息以及 12 种实体-关系对应信息，如表 4.1 和表 4.2 所示。

表 4.1 祁连山可持续发展关系类型

关系	相关关系词语
包含	包含、属于、组成
具有	具有、拥有、含有
存在	存在、栖息、分布、位于
生产	产生、造成、导致、释放、来源于、是因为
升降增减	升高、降低、增加、减少、促进、提高、改善、抑制、影响
分析研究	分析、研究、处理、采集、采样、吸附、去除、提取、萃取、方法、调查、方式、测定

表 4.2 实体与关系对应信息

实体类别	关系类别	实体类别
地理	包含	地理
地理	包含	生物
地理	包含	生态系统
生态系统	具有	环境因素
环境因素	具有	评价标准
污染物	具有	评价标准
生物	存在	生态系统
生物	生产	污染物
环境因素	升降增减	污染物
污染物	升降增减	生物
地理	分析研究	环境因素
生态系统	分析研究	污染物

4.3 知识图谱存储及可视化

4.3.1 基于 Neo4j 的知识图谱存储

鉴于 Neo4j 图数据库的界面友好、拓展性强等特点，本文采用 Neo4j 图数据库作为知识图谱存储及可视化工具。Neo4j 图数据库数据导入一般有两种方式，一是通过数据导入的方式，导入 CSV 文件，另一种方式是通过 Cypher 语言创建节点、关系和属性。本文通过使用 py2neo 模块包实现与 Neo4j 图数据库的连接，大批量构建节点、建立节点之间的关系和创建节点属性，从而构建大规模的知识图谱。其中 py2neo 是用于连接 Neo4j 的一个模块包，通过引用这个模块可以编写 Python 代码将三元组信息导入到数据库中。

在存储知识图谱的过程中，Neo4j 图数据库的地址为 http://localhost://7474，在用户第一次登陆时，需要重置密码后才可使用 NEO4j 数据库；启动方式为在系统配置中添加 Neo4j 图数据库变量后，每次启动都只需在命令提示符下执行 neo4j.bat console 命令，就可以进入数据库浏览器页面；启动后运行 Python 代码，将文件中的实体节点和关系导入到数据库中即完成知识图谱存储。本文构建的祁连山可持续发展知识图谱部分图谱如图 4.2 所示。

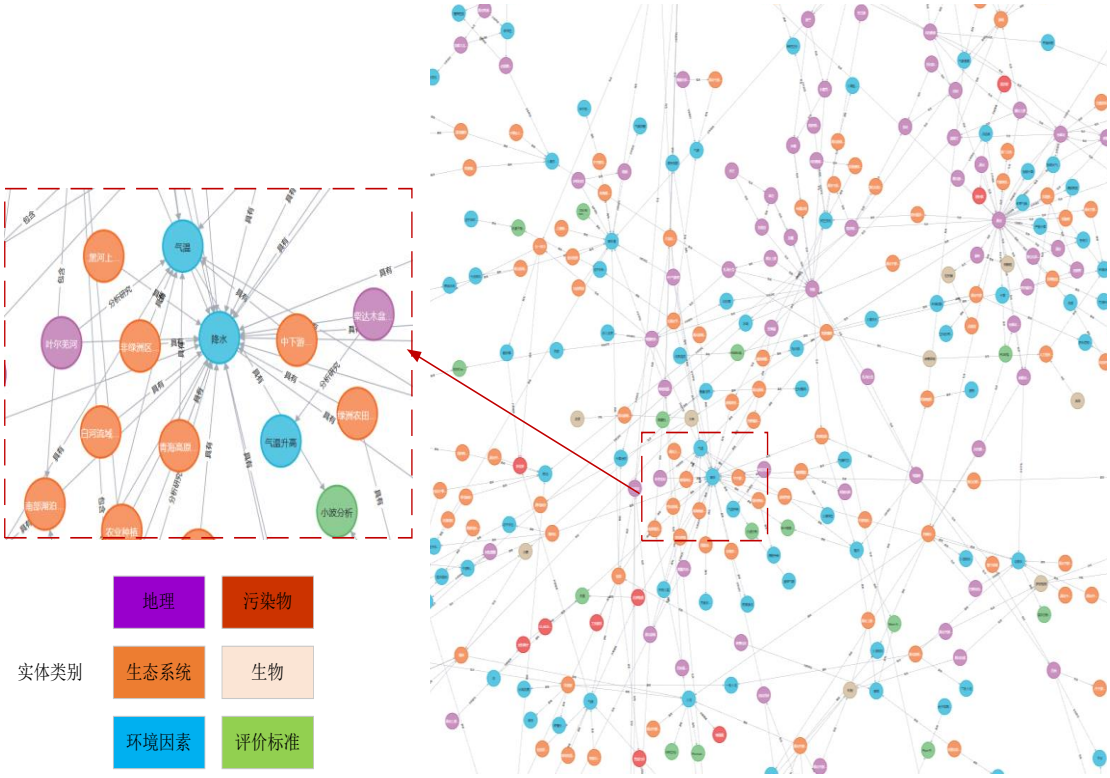

图 4.2 祁连山可持续发展知识图谱部分图

4.3.2 祁连山可持续发展信息可视化查询

祁连山可持续发展知识图谱为祁连山的可持续发展相关研究提供了一个可以进行可视化查询的知识库，通过知识图谱能够查询到祁连山地区及其可持续发展相关信息。在完成知识图谱存储后，在 Neo4j 中可以通过 Cypher 语言对知识图谱进行相关的可视化查询和修改。Cypher 语言也称为 CQL，它借鉴了 SQL 语言的结构，是一个声明式图查询语言，可以对图数据进行高效的查询和更新，其常用命令如表 4.3 所示。

表 4.3 Cypher 语言用法

命令	说明	Cypher 语言
CREATE	创建实体、关系	CREATE (<node-name>:<label-name>)
MATCH	检索实体、关系	MATCH (<node-name>:<label-name>)
RETURN	返回查询结果	RETURN <relationship-label-name>
WHERE	提供条件过滤检索数据	WHERE <condition>
DELETE	删除实体、关系	DELETE <node-name-list>
MERGE	CREATE 与 MATCH 命令组合	MERGE (<node-name>:<label-name>{<Property1-name>:<Pro<rty1-Value>})

通过 Cypher 语言可以完成对祁连山可持续发展信息的可视化查询，以 MATCH 命令为例，在 Neo4j 搜索框中查询祁连山可持续发展研究中生态系统中具有的环境因素相关信息，输入指令 MATCH (p:生态系统) -[:具有]-> (n:环境因素) RETURN p, n LIMIT 25，执行命令后返回所有生态系统中具有的环境因素，结果如图 4.3 所示。

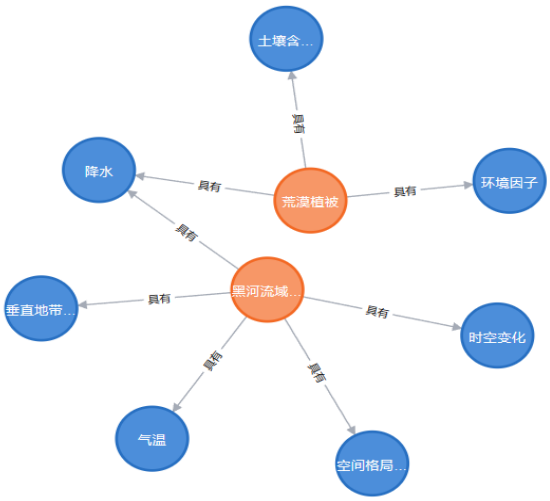

图 4.3 查询结果

4.4 基于 CiteSpace 的祁连山可持续发展文献分析

知识图谱分析软件 CiteSpace 是一款功能强大的可视化文献分析软件。该信息可视化软件利用共引分析和寻径网络等方法对文献数据进行可视化处理，能够将发文作者、发文机构、文献关键词等之间的关系通过可视化图谱的方式呈现出来，可以清晰地展示某一学科领域的演化过程。

某一学科的研究热点始终是该学科领域的热门话题，受到研究者和研究机构的密切关注，对推动学科创新与发展具有关键作用。利用 CiteSpace 软件能帮助研究者了解某一学科领域过去的研究轨迹和研究现状，掌握研究热点，把握该领域最集中的研究问题和最新的研究动态，并预测该学科未来的发展方向。同时在一篇文献中，关键词是反映作者的核心思想或核心观点的词汇，是对研究内容的高度凝练。因此本小节利用 CiteSpace 软件对本文构建的祁连山可持续发展文献数据进行关键词分析，掌握祁连山可持续发展领域的研究热点和研究主题，为进一步完善祁连山可持续发展知识图谱提供重点研究方向。祁连山可持续发展研究领域文献关键词聚类结果如图 4.4 所示。

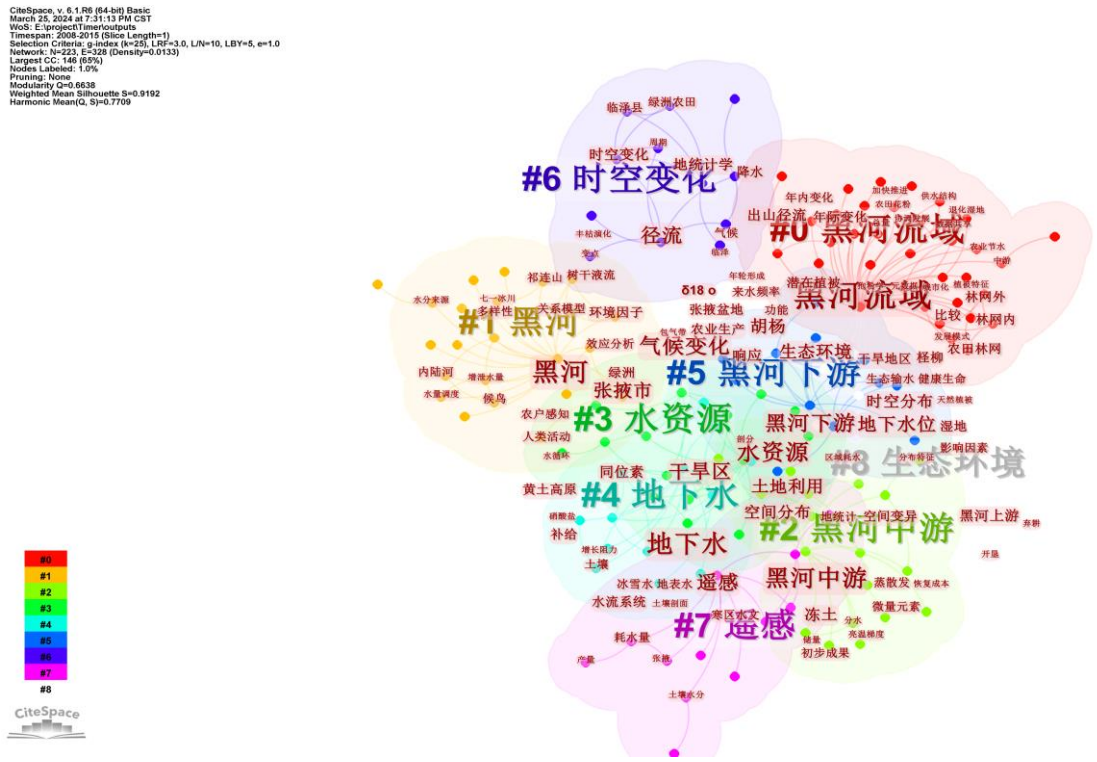

图 4.4 文献关键词聚类分析结果

从图 4.4 中可以看出，祁连山可持续发展研究领域文献关键词主要有 8 个聚类，对这些聚类进行汇总分析，可以发现关于祁连山的可持续发展研究主要集中在黑河流域，主要通过祁连山地区的水资源与地下水信息、时空变化信息、遥感信息和生态环境领域信息进行可持续发展研究。文献关键词聚类分析的结果为后续本文知识图谱的完善工作提供了重点修改、补充方向以及补充构建工作中需要关注的重点研究领域。

4.5 祁连山综合集成研讨厅与知识图谱应用

祁连山“山-水-林-田-湖-草”复杂共同体是一个复杂巨系统，其各要素之间的相互作用影响着祁连山及其各区域发展可持续性。祁连山综合集成研讨厅围绕祁连山“山水林田湖草”系统，利用流域综合集成模型、机器学习和元分析方法，分别融合硬数据（观测和统计数据）和软数据（社交网络大数据和科技文献大数据），通过 D-S 证据理论将软数据和硬数据融合，实现定性和定量数据综合集成。结合 Internet、GIS、DSS，集成流域综合集成模型、社交网络大数据融合模

型、元分析模型和知识图谱到综合集成研讨厅框架，研建面向不同科学议题的祁连山可持续综合集成研讨厅系统，实现科学模型在决策管理中的应用，为祁连山“山水林田湖草”资源优化配置提供科学依据。

祁连山综合集成研讨厅总体设计共包含六个部分：议题分解与识别框架构建、流域集成模型数据融合、社交网络大数据融合、科技文献大数据融合、知识图谱和证据转换与软硬数据融合。祁连山综合集成研讨厅系统总体架构图如图 4.5 所示。其中本文构建的祁连山可持续发展知识图谱为研讨厅总体设计的一部分。知识图谱将与祁连山可持续发展相关的知识有效地组织起来形成计算机可理解和可计算的知识体系，为祁连山综合集成研讨厅数据资源层提供知识图谱数据，为研讨厅快速知识获取、知识推理和知识融合提供基础知识智库。

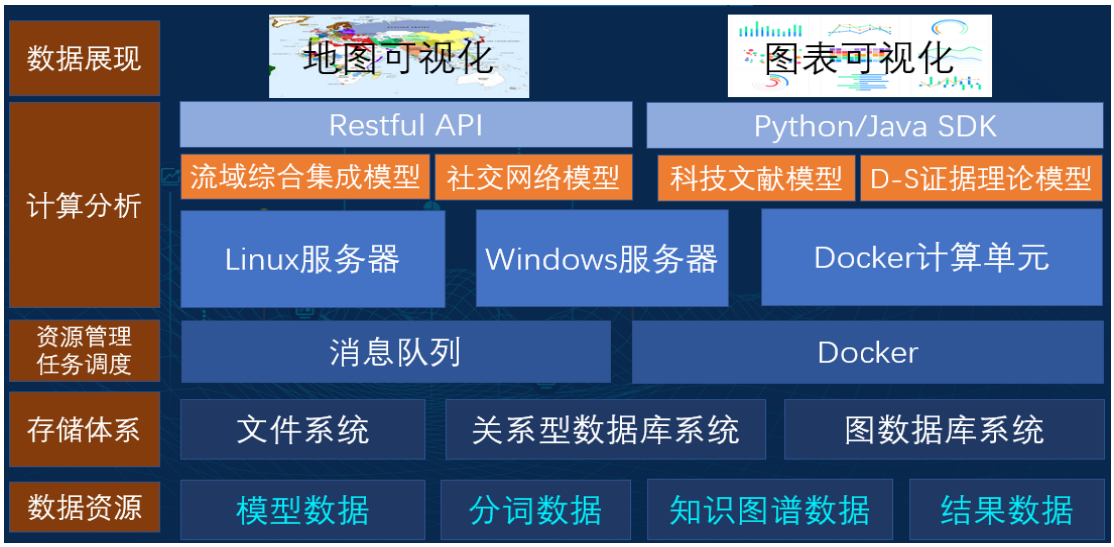

图 4.5 祁连山综合集成研讨厅系统架构图

4.6 本章小结

本章首先对祁连山可持续发展知识图谱的构建框架进行了介绍，框架包括数据获取、实体关系定义、信息抽取和知识存储 4 个部分。其次对关系定义进行阐述，共定义了知识图谱中 6 种关系和 12 种实体关系对应信息。将抽取好的祁连山可持续发展的实体以及关系导入 Neo4j 图数据库中完成知识图谱的构建和可

视化查询工作。然后，基于 CiteSpace 软件对本文构建的祁连山可持续发展文献数据根据关键词信息进行研究热点分析，分析结果为后续完善祁连山可持续发展知识图谱提供了重点研究方向。最后对祁连山综合集成研讨厅进行了简要的介绍以及本文构建的知识图谱在其中的应用。

5 总结与展望

5.1 工作总结

随着可持续发展理念越来越受到重视，有关祁连山可持续发展的研究也在随之增加，同时产生了许许多多的祁连山相关数据信息，不过这些数据没有规范成体系的建立为知识库，对于研究人员进行祁连山可持续发展相关信息查询和学术研究非常不便，因此借助知识图谱技术，构建一个直观有效的祁连山可持续发展知识图谱具有较高现实意义。本文基于深度学习模型，针对祁连山实体识别任务以及知识图谱构建展开了一系列研究，具体工作总结如下：

（1）建立祁连山可持续发展数据集。在祁连山的可持续发展研究领域目前还没有较好的公开数据集用于实体识别和构建知识图谱，并且根据实验所需的实体类别和关系类别的要求不同，公开数据集无法达到通用，因此本文选择自行构建数据集，数据集可分为原始数据集与标注数据集，原始数据集可用于祁连山可持续发展研究的文献计量分析，标注数据集为后续实验提供了基础。

（2）提出 ALBERT-BiLSTM-Attention-CRF 命名实体识别模型。为提高实体识别准确率，本文采用轻量化 BERT 模型，缩减模型参数量，并融合注意力机制提高模型对输入文本序列不同位置信息的关注度，并最终在本文的命名实体标注数据中，以其参数少的优势，取得了更好的性能。

（3）构建祁连山可持续发展知识图谱。目前关于祁连山可持续发展相关实体的识别和构建知识图谱、知识库的应用都处于空白，因为本文利用确定好的实体类型和关系类型，使用深度学习模型和人工标注的方法提取三元组，最后采用 Neo4j 图数据库对知识图谱进行创建与存储。

5.2 工作展望

本文针对祁连山可持续发展数据进行实体识别实验，并在此基础上经过规则模板和人工标注结合的方式对数据进行了实体关系抽取，并完成三元组的构建，最后构建了祁连山可持续发展知识图谱，取得了一定成效，但是仍存在一些局限

亟待解决，这也是未来工作的相关研究方向：

（1）扩充祁连山可持续发展数据集。目前数据集的文本数据量偏少，尤其是可用于实体识别实验的标注数据量较少。扩充数据集的文本数量可以让深度学习模型有更好的学习效果。同时，祁连山地区在社会、经济、生态三方面具有复杂的实体和实体关系，需要进一步的细化、增加调整实体类别和关系类别以更好的适应实际研究需求。

（2）完善祁连山可持续发展知识图谱。现有的知识图谱节点与关系并未完全包含祁连山可持续发展文献数据的研究热点，除了用于知识图谱构建的数据只占原始数据的一部分原因外，还因为面对祁连山复杂巨系统在时空上的关联关系，现有的模型进行关系抽取性能不佳，因此本文后期主要的工作将依据文献研究热点，完善实体关系和关系抽取研究，最后进一步完善祁连山可持续发展知识图谱。

参考文献

[1] 王晓琪,赵雪雁.人类活动对国家公园生态系统服务的影响——以祁连山国家公园为例[J].自然资源学报,2023,38(04):966-982.

[2] Steffen W, Broadgate W, Deutsch L, et al. The trajectory of the Anthropocene: The Great Acceleration[J]. Anthropocene Review, 2015, 2(1):81-98.

[3] 汪孝贤,张秀霞,李旺平等.基于遥感生态指数(RSEI)改进模型的祁连山国家级自然保护区生态环境质量评价[J].生态与农村环境学报,2023,39(07):853-863.

[4] Zhigang Z ,Xiong L ,Maojian C, et al.A Survey of Knowledge Graph Construction Using Machine Learning[J]. School of Computer and Communication Engineering, University of Science and Technology Beijing ,Beijing, 100083 ,China ,2023,139(1):225-257.

[5] Peng W.A Survey of Research on Deep Learning Entity Relationship Extraction[J].Natural Language Processing and Speech Recognition,2019,1(1)

[6] 郭华东,梁栋,陈方等.地球大数据促进联合国可持续发展目标实现[J].中国科学院院刊,2021,36(08):874-884.

[7] Allec A L ,Alejandro J C .Assessments under the United Nations Sustainable Development Goals: A Bibliometric Analysis[J].Environmental and Climate Technologies,2022,26(1):166-181.

[8] 王有恒,李丹华,卢国阳等.祁连山气候变化特征及其对水资源的影响[J].应用生态学报,2022,33(10):2805-2812.

[9] 王涛,高峰,王宝等.祁连山生态保护与修复的现状问题与建议[J].冰川冻土,2017,39(02):229-234.

[10] 蒋强,魏林波,李艳等.基于高精度观测资料的 2009-2019 年祁连山地区降水特征分析[J].兰州大学学报(自然科学版),2022,58(01):89-98.

[11] Yingchun Ge, Xin Li, Chunlin Huang, et al. A Decision Support System for irrigation water allocation along the middle reaches of the Heihe River Basin, Northwest China[J]. Environmental Modelling and Software,2013,47.

[12] Xin Li, Guodong Cheng, Shaomin Liu, et al. HEIHE WATERSHED ALLIED

TELEMETRY EXPERIMENTAL RESEARCH (HiWATER): Scientific Objectives and Experimental Design[J]. Bulletin of the American Meteorological Society,2013,94(8).

[13] 盖迎春,李新.黑河流域中游水资源管理决策支持系统设计与实现[J].冰川冻土,2011,33(01):190-196.

[14] 盖迎春,李新.水资源管理决策支持系统研究进展与展望[J].冰川冻土,2012,34(05):1248-1256.

[15] 张江蕾,陈少辉.祁连山自然保护区植被覆盖时空变化及地形分异研究[J].西部林业科学,2023,52(01):106-112+121.

[16] 杨欣,薛华柱,董国涛等.1982—2022 年祁连山植被变化及其驱动因子[J/OL].生态学杂志:1-16[2024-03-27].http://kns.cnki.net/kcms/detail/21.1148.Q.20240305.0924.002.html.

[17] 张强,冯悦,魏伟等.基于 GIS 的祁连山生态敏感性评价[J].安全与环境学报,2019,19(03):1056-1064.

[18] 汪慧玲,李励恒.祁连山国家级自然保护区周边经济发展模式探讨[J].林业经济问题,2008,28(06):521-525.

[19] 邸华,贺晓香.加快祁连山林区经济发展对策[J].中国林业,2012(14):59.

[20] 蒋志成,汪有奎,陈志宏等.生态旅游对祁连山国家级自然保护区居民的影响[J].旅游纵览(下半月),2013(11):200-201.

[21] 张文昌.祁连山保护区生态保护与经济发展途径[J].中国林业经济,2018(04):47-48.

[22] 李华芸.天祝县生态建设与绿色发展的实践与思考[J].甘肃农业,2019(11):50-53.

[23] 祁帜.祁连山地区生态保护与转型发展路径研究——以甘肃省肃南县为例[J].财会研究,2020(02):76-80.

[24] 王天雁.草地承包经营权之殇:祁连山牧区生态保护与牧民生计保障调查[J].中国不动产法研究,2020(01):155-170.

[25] 张吉祥,张祥森,武长旭等.知识图谱构建技术综述[J].计算机工

程,2022,48(03):23-37.

[26] Fensel D, Şimşek U, Angele K, et al. Introduction: what is a knowledge graph?[M]//Knowledge Graphs.Springer,Cham,2020:1-10.

[27] Zou X .A Survey on Application of Knowledge Graph[J].Journal of Physics Conference Series,2020,1487(1):012016.

[28] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]. Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008: 1247-1250.

[29] Suchanek F M, Kasneci G, Weikum G.YAGO: A Large Ontology from Wikipedia and WordNet[J]. Journal of Web Semantics, 2008,6(3):203-217.

[30] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia-A crystallization point for the Web of Data[J].Web Semantics Science Services & Agents on the World Wide Web,2009,7(3):154-165.

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

[32] Zheng X, Wang B, Zhao Y, et al. A knowledge graph method for hazardous chemical management: Ontology design and entity identification[J]. Neurocomputing, 2021, 430: 104 111.

[33] Rotmensch M, Halpern Y, Tlimat A, et al. Learning a health knowledge graph from electronic medical records[J]. Scientific reports, 2017, 7(1): 1-11.

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

[35] Liu J, Schmid F, Li K, et al. A knowledge graph-based approach for exploring railway operational accidents[J]. Reliability Engineering & System Safety, 2021, 207: 107352.

[36] 陈彦光,刘海顺,李春楠等.基于刑事案例的知识图谱构建技术[J].郑州大学学报(理学版), 2019, 51(3): 85-90.

[37] 常晋义.生态环境监测与管理系统知识库的设计[J]. 计算机应用与软件, 2003, 20(10):3.

[38] 徐超.生态红线知识库与服务管理工具的设计与实现[D]. 武汉大学, 2018.

[39] 陈兰鑫.洞庭湖生态环境监测系统知识图谱的构建[D]. 湖南农业大学, 2019.

[40] 张华.面向文献文本的生态环境领域知识图谱构建研究[D].武汉大学,2022.

[41] 李长哲.面向唐卡文化的知识图谱构建与研究[D].青海大学,2022.

[42] Klyne G. Resource description framework (RDF): Concepts and abstract syntax[J]. http://www.w3.org/TR/rdf-concepts/, 2004.

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

[44] 刘凯.虚拟健康社区中的命名实体识别方法研究[D].北京理工大学,2016.

[45] 尹越.财经问答系统数据服务子系统的设计实现[D].哈尔滨工业大学,2012.

[46] 赵辉,庞海婷,冯珊珊等.中文命名实体识别技术综述[J].长春工业大学学报,2021,42(05):444-450.

[47] Vishwanathan S, Murty M N.SSVM:a simple SVM algorithm[C].Proceedings of the 2002 International Joint Conference on Neural Networks.2002:2393-2398.

[48] RATNAPARKHI A.A Maximum Entropy Model for Part of speech Tagging[C].Conference on Empirical Methods in Natural Language Processing,1996:133-142.

[49] Vogel,Stephan,Hermann Ney,and Christoph Tillmann.HMM-based word alignment in statistical translation[C].The 16th International Conference on Computational Linguistics.1996.

[50] Han A L F, Wong D F, Chao L S. Chinese named entity recognition with conditional random fields in the light of Chinese characteristics[C]//Language Processing and Intelligent Information Systems: 20th International Conference, IIS 2013, Warsaw, Poland, June 17-18, 2013. Proceedings. Springer Berlin Heidelberg, 2013: 57-68.

[51] 马瑞祥.面向中文医学领域的知识图谱构建关键技术研究[D].浙江大学,2023.

[52] Chiu J P, Nichols E. Named entity recognition with bidirectional lstm-cnns[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[53] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(01): 2493-2537.

[54] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(08): 1735– 1780

[55] Dong C, Zhang J, Zong C, et al. Character-based lstm-crf with radical-level features for chinese named entity recognition[M]. Natural Language Understanding and Intelligent Applications. Berlin, Germany: Springer, 2016: 239-250.

[56] 齐宁.基于深度学习的中文实体关系抽取研究[J].长江信息通信,2024,37(01):64-66.

[57] 刘辉,江千军,桂前进等,实体关系抽取技术研究进展综述[J].计算机应用研究,2020, 37(S2):1-5.

[58] 秦伟德.基于深度学习的政务网站人事信息知识图谱构建研究[D].兰州财经大学,2023.

[59] 马江微,吕学强,游新冬,肖刚,韩君妹.融合 BERT 与关系位置特征的军事领域关系抽取方法[J].数据分析与知识发现,2021,5(08):1-12.

[60] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[C]//Proceedings of the 25th International Conference on Computational Linguistics. Stroudsburg: ACL,2014: 2335-2344.

[61] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012:1201-1211.

[62] 鄂海红,张文静,肖思琪等.深度学习实体关系抽取研究综述[J].软件学报,2019,30(6):1793-1818.

[63] 陈佳沣. 基于强化学习的实体关系联合抽取模型研究[D].武汉大学,2019.

[64] Miwa M, Bansal M. End-to-end relation extraction using lstms on sequences and tree structures[J]. arXiv preprint arXiv:1601.00770, 2016.

[65] John Lafferty, Andrew McCallum, Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001), 2001.

[66] Jin W ,Yu Z ,Fu Q D , et al.A bidirectional long short-term memory network for electron density diagnostic with double probe[J].Measurement Science and Technology,2023,34(12):

[67] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Annual Conference on Neural Information Processing Systems, 2017:5998-6008.

[68] Ming J ,Junlei W ,Xiangrong S , et al.Transformer Based Memory Network for Sentiment Analysis of Web Comments[J].IEEE Access,2019,7179942-179953.

[69] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// The Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2019:4171–4186.

[70] 王天一,孟小亮,张华.一种面向生态环境领域的知识图谱构建方法[J].地理空间信息,2023,21(01):14-19.

攻读硕士学位期间发表的论文及科研情况

发表论文：

[1]韩运龙,尚庆生,赵薇等.结合密度峰值和集成过滤器的自训练算法[J/OL].宜宾学院学报,1-12[2024-03-27].http://kns.cnki.net/kcms/detail/51.1630.Z.20231205.1101.004.html.

致 谢

三年的时光仿佛弹指一挥间，我的校园生活就要划上句号。回首两年来的学习生涯，有初入校园的斗志昂扬，亦有面对困难的彷徨失落，有同学相伴的甜蜜情谊，亦有不随人意的失之交臂。种种回忆，历历在目。

即将走出校园，心中充满了无尽的感激，首先要感谢我的导师尚庆生老师对我的体谅和包容，在学习、生活等各个方面对我的悉心指导和殷切关怀。初次见面老师的随和就给我留下了深刻的印象，在之后的接触中老师对待学生和蔼可亲的态度，带给了我们无限的温暖。感谢管理科学与工程专业的所有老师们，是你们认真负责的态度，无私的授课让我在这三年的学习中受益匪浅，衷心祝愿各位老师工作顺利，生活美满。

其次感谢实验室的伙伴们，让实验室充满了欢声笑语，让学习变成一件愉快的事，感谢两位师姐在我入学什么都不懂时的无私帮助，感谢我的同门郭泓同学在学习和生活中的帮助，在找工作时无私的分享。

此外，很幸运在读研期间结交了许多志同道合的好友，特别是我的饭搭子们，在一起吃了无数顿饭的吴义稳，陈贵富，孙梦泽和赵金雨同学，除了吃饭外在三年时光中还有着数不清次数的帮助，感谢有你们让我不再孤单，很高兴在兰财认识你们，祝愿大家毕业快乐，友谊长存。

最后感谢我的家人，他们永远是我成长道路上最无私的奉献者，二十载求学生涯，他们一直给予我最温馨的鼓励和关怀。谨以此致谢最后，我要向百忙之中抽出时间对本文评审的各位老师表示衷心的感谢。