分类号 TP391.1　　密级
U D C　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目　基于深度学习的政务网站人事信息知识图谱构建研究

研 究 生 姓 名：秦伟德

指导教师姓名、职称：杨海军　教 授

学 科、专 业 名 称：管理科学与工程

研 究 方 向：信息管理与信息系统

提 交 日 期：2023 年 6 月 6 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：秦伟德 签字日期：2023.5.20

导师签名： 签字日期：2023.5.20

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：秦伟德 签字日期：2023.5.20

导师签名： 签字日期：2023.5.20

Research on Knowledge Graph Construction of Personnel Information of Government Affairs Website Based on Deep Learning

Candidate : Qin WeiDe

Supervisor : Yang HaiJun

摘 要

随着政务信息化发展水平的不断提高，政务网站也产生了越来越多的政务数据，在这些数据中，人事信息数据是非常重要并且具有较高研究价值的一种数据。一方面，人是各单位各部门的主体，与人相关的信息是各个决策的基础，具有重要作用；另一方面当前的人事信息主要存在于独立的人事任免、政事新闻等文本里，这些信息各自分散缺少关联，浪费了信息本身存在的价值。当下自然语言处理相关技术正在飞速发展，为处理政务人事信息数据提供了技术支撑，知识图谱的应用也为相关学术研究和领域发展提供了直观有效的工具。深度学习能够从非结构化的文本数据中提取结构化的三元组信息，在此基础上可以构建政务人事信息知识图谱，但是目前仍然存在着数据集的匮乏、现有算法模型精度不够高等问题。针对以上问题，本文的主要研究内容如下：

（1）建立政务网站人事信息数据集。使用 Python 语言编写爬虫程序从政务网站中获取原始文本数据并进行数据清洗，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数据集的创建。数据集共包含 10 种关系类别和对应的实体类别，为后续的实验研究提供数据支撑，除此之外也能推动政务人事信息实体关系抽取领域的发展。

（2）提出了融合依存句法分析的 GCN-CasRel 实体关系联合抽取模型。模型基于 CasRel 模型端到端的级联二进制标记框架，在解决政务人事信息文本中出现的三元组重叠问题的同时，采用图卷积神经网络对依存句法关系进行建模，使模型更好的捕获句法结构信息，同时引入注意力机制过滤依存句法树的噪声，提高实体关系抽取性能。经过与其他联合抽取模型实验对比，本模型的查准率和 F1 值都有较高的提升，证明了模型的有效性。

（3）构建政务网站人事信息知识图谱。利用已提出的实体关系联合抽取模型开发了一款基于微信平台的政务人事信息实体关系抽取小程序。微信小程序分为前端页面、后端服务器以及数据端等三部分，使用者可以通过上传文本或者网址来直接获得三元组抽取结果，并在此基础上采用 Neo4j 图数据库对小程序抽取的人事信息三元组进行可视化，完成政务网站人事信息知识图谱的创建。

本文通过爬取政务网站文本数据，经过预处理建立政务网站人事信息数据集，在端到端的级联二进制标记框架基础上引入图卷积神经网络建模依存句法关系，

并利用注意力机制过滤依存句法树的噪声，经实验证明模型具有良好的效果。基于微信平台开发了政务人事信息实体关系抽取小程序，利用小程序对甘肃省部分政务文本进行实体关系抽取，获得人事三元组信息后利用 Neo4j 图数据库完成知识图谱可视化。

关键词: 人事信息 联合抽取 知识图谱 小程序

Abstract

With the continuous improvement of the development level of government information, the government website has also produced more and more government data, among which the personnel information data is a very important and high research value data. On the one hand, people are the subject of each unit and department, and the information related to people is the basis of each decision, which has an important role; on the other hand, the current personnel information mainly exists in the independent text of personnel appointments and dismissals, political news, etc., which are scattered and unrelated, which wastes the value of the information itself. The technology related to natural language processing is developing rapidly nowadays, which provides technical support for processing personnel information data of government affairs, and the application of knowledge graph also provides an intuitive and effective tool for related academic research and field development. Deep learning can extract structured triadic information from unstructured text data, on which the knowledge graph of government personnel information can be built, but there are still problems such as the lack of data sets and the lack of accuracy of existing algorithm models. To address the above issues, the main research of this paper is as follows:

(1) Create a dataset of personnel information from the government website. We use Python language to write a crawler program to get the

original text data from the government website and clean the data, determine the entity type and relationship type and then annotate the text data, and finally convert the text into a specific format to complete the creation of the dataset. The dataset contains a total of 10 relationship categories and corresponding entity categories, which can provide data support for subsequent experimental research and promote the development of entity relationship extraction in the field of government personnel information.

(2) A joint GCN-CasRel entity relationship extraction model incorporating dependency syntax analysis is proposed. The model is based on the end-to-end cascaded binary tagging framework of CasRel model. While solving the triad overlap problem in the text of government personnel information, the model uses graph convolutional neural network to model the dependent syntactic relations so that the model can better capture the syntactic structure information, and introduces the attention mechanism to filter the noise of the dependent syntactic tree to improve the entity relation extraction performance. After the experimental comparison with other joint extraction models, the accuracy rate and F1 value of this model have been improved, which proves the effectiveness of the model.

(3) Constructe a knowledge graph of personnel information of governmental websites. An entity relationship extraction applet based on the WeChat platform is developed using the proposed joint entity

relationship extraction model. The applet is divided into three parts: front-end page, back-end server and data side. Users can directly obtain the triad extraction results by uploading text or URL, and on this basis, the Neo4j graph database is used to visualize the personnel information triad extracted by the applet and complete the creation of the personnel information knowledge map of the government affairs website.

In this paper, we crawl the text data of government websites, establish the personnel information dataset of government websites after pre-processing, introduce graph convolutional neural network to model the dependent syntactic relations based on the end-to-end cascaded binary tagging framework, and filter the noise of the dependent syntactic tree by using the attention mechanism, and prove the model has good effect by experiment. Based on the WeChat platform, we developed an applet for extracting entity relations of government personnel information, and used the applet to extract entity relations of some government texts in Gansu Province, and completed the knowledge graph visualization using Neo4j graph database after obtaining the personnel triad information.

Keywords: Personnel Information; Joint Extraction; Knowledge Graph; Mini Programs

目 录

1 绪论

1.1 研究背景及意义

自《中华人民共和国政府信息公开条例》等文件公布以后，各级政府部门大力推进政府信息公开，推行一系列信息惠民措施。而随着大数据、云技术的发展，“互联网＋政务服务”应运而生，政务网站的建设成为“互联网＋政务服务”这一概念的重要体现，国务院出台的一系列政策文件加以肯定，并强调加强政务网站的建设和管理、整合相关网站之间的协调联动，打造更加全面的便民平台。

政务网站蕴含着大量的政务数据，这些数据内容丰富，形式多样，涉及政策法规、行政规划、机构介绍、人事信息、统计数据等多个领域。据统计，2021 年广东省各级政府公开各类行政信息超过 2000 万条，浙江省主动公开各类政府信息超过 800 万条，甘肃省各地各部门也通过政府网站、政务新媒体等平台主动公开各类政府信息达 250 万条。作为所有政务网站和政务数据的基础信息，人事信息有着较大的数据规模和较高的利用价值，随着我国政府机关和事业单位人事制度不断的发展改革中，人事规模和部门种类也随之扩大，越来越多的人员加入也带来更多的信息资源，如果被充分利用能够对开展学术研究、商业经营带来新的机遇和发展。目前人事信息的利用仍然停留在相关人员的人事任免这一层面，海量的人事数据因为数据庞大、关联复杂等因素没有得到有效的整理归纳，因此当下的人事信息有着进一步可挖掘的空间。人事信息一般隐藏在非结构化文本数据之中，如人事任免通知、人物事迹宣传等文本文件，利用人工识别进行有效人物信息的提取面临着工作量巨大、效率较低的困难，因此利用自动化的文本识别算法能够快速、全面、高效地识别提取相关的结构化信息数据，再将信息数据进行存储能够解决以上难题。而从非结构化数据提取出结构化数据，挖掘出信息数据之间的隐藏关联，进而构建知识图谱，能够为个人用户、企业机构在智能问答、推荐系统等方面提供进一步的帮助，对相关研究和产业发展都有着重要的意义。

知识图谱是一种特殊的语义网络，用来描述真实存在的实体以及实体之间的相关联系[1]，通过 Neo4j 等图数据库，知识图谱可以进行可视化，能够直观的表达实体、知识、概念之间的关联关系。知识图谱构建的核心任务在于信息抽取技

术，从非结构化数据中提取结构化数据是信息抽取的主要目的，通过统一形式集成，以便在海量文本数据中找出所需数据[2]。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的边界和类别，在人事信息知识图谱构建中，主要是识别人物的姓名、民族、籍贯、学历、职务、政治面貌等信息；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，在人事信息任务中，主要的关系有民族关系、出生地、教育经历、担任职务、出生日期、毕业院校等。通过信息抽取，从非结构化文本数据中提取结构化数据，能够获取文本的语义信息，得到三元组<Subject，Predicate，Object>，三元组的两端分别代表两个实体，中间由实体关系进行连接。例如，<成龙，出生地，香港>就是表达成龙的出生地点，构建成龙与香港之间的关系。通过从政务网站的文本信息中提取大量的三元组并进行相互关联，在此基础上可以构建人事信息知识图谱。政务人事信息知识图谱的建立既能够充分挖掘政务文本信息，降低个人用户、企业机构等获取政务人事数据的门槛，促进政务信息化的发展，又能为智能问答、信息搜索等应用领域发展提供数据基础。

1.2 实体关系抽取研究现状

在信息抽取领域，实体关系抽取是核心任务和重要环节[3]，能够从非结构化或半结构化的文本中自动识别实体、实体类型以及实体之间特定的关系类型，为知识图谱构建提供关键支持[4]。实体关系抽取这一概念于 1998 年 MUC-7 会议中首次提出[5]，经过 20 年的发展，先后形成了基于规则和词典驱动的抽取方法、基于传统机器学习的抽取方法、基于深度学习的抽取方法和基于开放领域的抽取方法。

1.2.1 基于规则和词典驱动的方法

基于规则的实体关系抽取方法需要依靠人工进行语义规则制定，通过专用语言描述实体与实体之间的关系，并且将语料进行预处理，处理完毕后将语料与规则进行匹配，最终完成关系的分类抽取。McDonald[6]、Aone[7]和 Fukumoto[8]等都在英文领域构建了相应的语义规则进行了实体关系抽取，均取得了一定的成效。

在中文领域，邓擘等人[9]发现利用基于模板进行关系匹配的方法效果不明显，进而在模式匹配的基础上引入了词汇语义信息，大大提高了关系抽取的准确率。基于规则的方法在规则性较强的领域时能够发挥较好的作用，但是也有着明显的缺点，比如需要规则的制定者对面向的领域有着较深的理解、规则时间制定周期长、面对不同的领域可移植性差等。相比基于规则的方法，基于词典驱动的实体关系抽取方法更加灵活[10]。基于词典驱动的方法需要构造一个词典，词典中囊括表示各种实体关系类型的动词，面对新的实体关系类型，词典需要扩充，这种情况下，只要往词典里加入相应动词即可。Aone 等人[11]提出了一种快速、灵活的实体关系抽取方法 REES(Large-Scale Relation and Event Extraction System)，主要目的是面对大规模事件进行关系抽取，这种方法在 39 种关系类型构成的数据集中进行测试，F1 值达到了 75.35%，对比基于规则的方法提高了 6.35%。基于词典驱动的抽取方法简洁高效，使用者不需要对相关领域的语言知识有较深的了解，但是因为这种方法围绕的中心是动词，对于其他非动词的关系类型抽取难以实现，因此这种方法很快被新的实体关系抽取方法所取代。

1.2.2 基于传统机器学习的方法

基于机器学习的实体关系抽取方法通过训练样本学习出一个实体关系抽取模型，再将模型用于测试样本的预测。根据对语料库的依赖程度，基于机器学习的方法可以划分为有监督学习的方法、半监督学习的方法、无监督学习的方法。有监督学习的方法本质上是把关系抽取问题看作多个关系分类问题进行处理，使用的语料库是全部经过标注的，有监督的机器学习关系抽取方法包括基于特征向量的算法[12]和基于核函数的算法[13]两种。基于特征向量的方法是根据语料库文本的上下文信息选取合适的句法语法特征进行特征向量的构建，然后进行实体关系识别模型的训练来实现文本中实体关系的抽取。Kambhatla[14]在最大熵模型的基础上结合多种文本特征进行关系分类，利用少量词汇在 ACE RDC2003 数据集中测试取得了优秀的成绩。Sun 等人[15]利用两个实体之间的长期相关性特征、实体顺序特征、标点符号特征并融合上下文特征，使用朴素贝叶斯模型和投票感知模型两种算法进行实体关系的分类。Jiang 等人[16]研究对比了多个特征对关系抽取结果的影响，提出了过多的特征会对关系抽取结果产生副作用，只需要基础特

征就能够改善关系抽取模型。在中文领域，车万翔等人[17]采用 Winnow 和支持向量机(Support Vector Machine，SVM)两种机器学习算法进行中文实体关系的识别，结果表明当实体的左右两个词作为特征时，抽取效果最好。甘丽新等人[18]在基本特征的基础上融入句法语义特征，以《人民日报》版面内容作为语料库，使用 SVM 分类器进行关系抽取，实验取得了较好的结果。基于核函数的方法使用隐形特征映射代替显性特征映射，利用本文之间远距离特征和结构化特征弥补上下文信息利用不充分的问题，提高语义识别能力。Zelenko 等人[19]率先将核函数应用于浅层解析树结构中，通过使用 SVM 分类器和投票感知模型在关系抽取分类任务上取得了较好的效果。Zhang 等人[20]首次融合多个单一核函数组成复合核函数并用于实体关系抽取任务。实验结果表明，复合核函数的多项实验指标均好于单一核函数。在中文关系抽取领域，刘克彬[21]利用基于核函数结合 K 近邻(K-Nearest Neighbor，KNN)算法实现中文实体关系抽取。虞欢欢等人[22]也基于卷积树核函数方法结合语义信息在中文数据集关系抽取上取得了较好的效果。有监督的机器学习方法在各种数据集的关系抽取任务上都取得了一定效果，但是这种方法依赖于完全标注好的语料库，成本较高工作复杂。

半监督学习的抽取方法利用少数标注好的数据作为种子样本，不断地迭代学习最终训练出分类模型[23]，这显著减少了人工标注数据的工作量。Brin[24]首次将基于 Bootstrapping 的方法应用于实体关系抽取任务中，并最终构建了 DIPRE 系统，实验以少量的书名及作者名作为种子实体关系对，通过不断的迭代实现自动从语料库中获取抽取模板和关系实例。Agichtein 等人[24]在 DIPRE 基础上改进了表示实体关系的向量，设计了 Snowball 系统。除了 Bootstrapping 的方法，常用的半监督学习的抽取方法还包括协同训练[26]和标注传播[27]。半监督学习的方法避免了大量人工标注语料，降低了构建数据集的难度，同时也存在着极度依赖初始种子质量、迭代中会引入噪声的等问题。

无监督学习的抽取方法完全不依赖人工标注数据，在大规模语料库中，通过聚类的方法自底向上抽取实体关系。Hasegawa 等人[28]于 2004 年首次提出利用无监督的机器学习方法进行关系抽取，通过识别具有相似性的实体对进行聚类。Rozenfeld 等人[29]进一步完善了 Hasegawa 的模型，并利用基于上下文的模式大幅地提升了关系抽取的性能。Shinyama 等人[30]提出了多层级聚类的抽取方法，

在 12 家美国主流报刊的文章进行测试，获得了大量的模式和类型。在中文关系抽取领域，秦兵等人[31]提出了基于无监督机器学习的中文关系抽取模型，通过句式规则对三元组和指示词进行筛选，实验获得了较高的准确率。无监督学习的方法不依赖语料标注，可以很好的适应大规模文本和无规则内容，同时拥有更强的可移植性。但是相比有监督和半监督，无监督的方法准确率偏低。

1.2.3 基于深度学习的方法

在基于机器学习的关系抽取方法中，人工构建语义特征向量是必不可少的，这就要求研究者需要一定的专业领域知识并投入人力提取特征。相比之下，基于深度学习的关系抽取方法只需要训练大量的文本数据就能自行进行关系抽取。深度学习的概念由 Hinton[32]等人正式提出，随着不断的发展，研究者们逐渐将深度学习应用于文本实体关系抽取领域。

目前主流的基于深度学习的实体关系抽取方法分为两大类型：流水线方法(pipeline)和联合抽取方法(joint)。

流水线方法是将实体关系抽取任务分解成两个子任务：命名实体识别和关系抽取。两个子任务按顺序分别独立执行，在命名实体识别的基础上进行关系抽取，最后输出三元组结果。两个任务可以使用不同的深度学习模型，也可以在不同的训练集上训练模型。Zeng 等人[33]首次将卷积神经网络(Convolutional Neural Network，CNN)应用于文本关系抽取领域，利用 CNN 模型提取词级特征和句子级特征，大幅提升了关系分类的准确性。Xu 等人[34]在 CNN 基础上结合了最短依存路径并在公开数据集上进行测试，实验结果表明，最短依存路径的加入给模型带来了有效的提高。除了卷积神经网络，循环神经网络(Recurrent Neural Network，RNN)也被研究者们应用到关系抽取领域。Socher 等人[35]将 RNN 引入到文本关系抽取任务中，利用循环神经网络模型获取句子的向量表示，提高了实体关系抽取的性能。由于 RNN 因为模型层数较多容易出现梯度爆炸和梯度消失的问题，长短期记忆模型(Long Short Term Memory，LSTM)被研究者们引入到关系抽取任务中来。Xu 等人[36]提出了 SDP-LSTM 模型，该模型在 LSTM 模型的基础上，融合了依存句法分析树的最短路径、词向量、WordNet 和句法等特征，通过迭代学习进行关系分类。依存句法树能够让模型结合词语间的依存信息来挖掘

更深层次的语义关系，有效的改善关系抽取性能。图卷积神经网络(Graph Convolutional Network，GCN)为依存树结构信息的引入提供了新思路，GCN 最初应用于图数据结构处理，后来被逐渐应用于自然语言处理领域。Zhang 等人[37]在图卷积神经网络模型上进行改进，将修剪后的依存句法树引入到模型之中，这种修剪策略能够快速识别两个实体间的最短路径，过滤掉无关信息。流水线方法使得关系抽取模型的性能得到了提升，但是同时存在着误差传播、子任务缺少交互和产生冗余信息的问题[38]。为了解决这些问题，研究者们将两个子任务融合成一个任务，进行联合抽取。

联合抽取模型运用同一个模型进行实体识别和关系抽取两个任务，二者彼此交互，能够充分利用子任务的信息实现直接从文本中抽取三元组的目标。根据解码方式的不同，联合抽取方法可分为基于共享参数的方法和基于联合解码的方法。在基于共享参数的方法中，两个任务的共享编码层的参数，训练后得到全局最优参数。Miwa 等人[39]首次利用双向循环神经网络提取词序列和依存树结构信息，通过共享参数训练模型。Zheng 等人[40]提出一种新的混合模型，利用 LSTM 对实体进行提取和 CNN 对关系进行分类，改进了关系抽取的性能，但是这些模型都无法解决实体关系重叠问题。Wei 等人[41]在 2019 年提出了一种级联二进制标记框架 CasRel，该模型将关系看作成主体到客体的映射，有效的解决了实体关系重叠问题。基于联合解码的方法在序列编码层叠加统一的解码器，进一步加强了子任务的交互性。Zheng 等人[42]首次将实体识别和关系抽取结合成一个序列标注问题，通过唯一的解码器直接解码句子序列获得三元组。Zeng 等人[43]基于 Sequence-to-Sequence 提出了 CopyRE 模型，关系抽取的过程类似于机器翻译，解码器依次产生关系、主体和客体。基于深度学习的关系抽取方法各有优劣，但是都能够解决基于其他方法出现的难题，不需要人工选择特征的同时可移植性较强。

1.2.4 基于开放领域的方法

基于开放领域的实体关系抽取方法特点是对于文本类型、语料标注、关系类型都没有要求，主要通过前后的短语进行语义构建，接着借用大型非限定性语料库进行实体关系的抽取[44]。随着互联网的发展，许多开放性语料如网页数据等，由于自身数据量庞大、实体类型多、实体关系复杂，无法进行归纳概括，研究者

们尝试着借助机器学习的方法，对开放领域的信息进行实体关系抽取。2007 年 Banko 等人[45]发布了面向开放领域的信息抽取系统 TextRunner，该系统以自监督的方式训练朴素贝叶斯模型来进行关系分类，接着该系统在开放网页进行测试取得了优秀的效果。针对 TextRunner 无法提供有效信息的问题，Fader 等人[46]设计了 ReVerb 系统，该系统对于较长的文本采用先关系后实体的识别模式，大幅度的提高了关系抽取系统的性能，有力的促进了该领域的发展。面对开放的中文文本领域，鉴于中英文在语法使用上的差异，Qiu 等人[47]发布了面向中文开放领域的文本关系抽取模型，该模型利用依存解析树识别三元组，通过双向传播算法迭代抽取三元组，在 Wikipedia 中文数据中取得了 76.8％的实验准确率。秦兵等人[31]研究实体关系、实体间距离和关系词所在位置的关系并发布了 UnCORE 系统，该系统在网页上进行测试，测试结果表明能够有效的提取大量的关系三元组。郭喜跃等人[48]采用半监督的方式在百度百科等中文开放领域进行实体关系抽取，获得了质量较高的实体间二元关系。姚贤明等人[49]针对中文领域多元关系抽取的问题，以依存句法分析结果的根节点为入口，不断地迭代所获取的语句成分并完善定语成分，最终获得多个实体的语义关系。相比其他方法，基于开放领域的方法不需要太多人工关系，对数据也不做要求，但是目前该方法缺少客观的评价标准，同时一般模型测试的数据环境与实际的互联网环境有着较大的区别，后续还有较大的进一步发展改进空间。

1.3 知识图谱的研究现状

知识图谱由 Google 提出，最初被应用于搜索引擎[50]，随着对应用场景的不断拓展，知识图谱技术也被用于其他领域。2016 年李文鹏等人[51]为了解决软件复用中存在的一些问题,基于开源软件项目，构建了软件知识图谱，实现了软件知识检索的便利化。2017 年陈亚东等人[52]从多方面对苹果产业的知识图谱架构进行设计，推动知识图谱技术在水果产业的应用和发展。2018 年由丽萍等人[53]基于在线评论的语义信息构建情感知识图谱，实现智能化情感语义检索。2019 年奥德玛等人[54]针对复杂医学知识的精准描述问题，结合自然语言处理和文本挖掘技术，发布了中文医学知识图谱 CMeKG 1.0。2020 年陈璟浩等人[55]构建了“一带一路”投资知识图谱，并在知识图谱的基础上开发了问答系统。

在政务领域，华斌等人[56]创建了电子政务领域知识谱图，以此为基础为电子政务项目建设评价提供一种有效的辅助决策方法。高晨翔等人[57]对区域微博进行建模，对微博内容进行三元组抽取，构建了基于主题划分的区域政务微博知识图谱。朱宗尧[58]分析了上海“一网通办”服务场景的知识关联问题，提出利用知识图谱能够有效解决当前困境并构建了政务知识图谱的相关结构。黄贵辉等人[59]基于 CiteSpace 软件，分析几十年来的高质量论文构建知识图谱，并以知识图谱为基础研究国内行政改革的热点和趋势。

在人事领域，于娟等人[60]根据已有人物关系和隐藏人物关系，利用图数据库构建人物关系知识图谱。黄娟等人[61]以民国历史人物为研究对象，针对这些人的社交关系、生平事迹等信息构建知识图谱，将历史人物的多个方面可视化展示。孙洪伟等人[62]为了处理复杂人物关系，提出以本体为模型建立家谱知识图谱。

目前对于政务领域的研究重点在研究政策文本、政务服务和政务软件方面，缺少对政务人事信息的研究。而对于人事领域的知识图谱研究则聚焦于人物与人物之间的关系，较少涉及以个人自身信息的图谱建立。

1.4 主要研究内容与研究架构

1.4.1 主要研究内容

基于以上现状研究表明，一方面，深度学习在信息抽取领域有着很好的应用场景，但是目前缺乏针对政务领域人事信息的深度学习训练数据集；另一方面，知识图谱普遍应用于政务领域和人事信息领域，但二者结合的领域研究相对较少。

因此本文针对以上问题做的研究工作如下：

（1）数据集采集和构建。当前在政务人事信息领域，用于训练实体关系三元组抽取的公开数据集较为匮乏，并且根据任务需求的不同，数据集的样式、内容、标注情况等都有特定的要求。本文针对研究内容对政务网站人事相关的文本进行收集获取，通过预处理获得实验需要的数据集，填补政务人事信息抽取领域的空白，提高后续实验的效率。

（2）信息抽取模型构建。深度学习在信息抽取任务有着广泛的应用，当前主流的方法有流水线法和联合抽取两种，联合抽取方法能较好的解决信息抽取中

传播误差的问题。针对三元组重叠问题，本文基于 CasRel 模型提出 GCN-CasRel 实体关系联合抽取模型，使用图卷积神经网络对依存句法树进行建模，使模型更好的捕获句法结构信息，同时引入注意力机制过滤依存句法树的噪声，构建一个性能更佳的实体关系联合抽取模型，满足本文实验中信息抽取的需求。

（3）信息抽取移动应用的开发。为了降低使用者对文本进行三元组抽取进而构建知识图谱的难度，本文基于微信平台开发政务人事信息抽取微信小程序，让使用者能够简单快捷的识别文本中的人物关联信息，并为知识图谱的构建和扩展提供了便利可行的工具。

（4）政务人事信息知识图谱构建。本文在政务人事信息抽取微信小程序构建完毕后，对相关的数据进行三元组的抽取，基于 Neo4j 图数据库实现知识图谱的可视化，并演示查询功能，更加清晰地展示人物自身的关联信息，有效提高人事信息分析的效率，满足智能问答、信息搜索等应用领域的需求。

1.4.2 研究组织架构

本论文拟分为五个章节，研究内容和论文架构如图 1.1 所示：

第一章绪论。绪论部分对本文的研究背景及意义、信息抽取的研究现状以及知识图谱的研究现状做出阐述，重点介绍了信息抽取的主流方法以及各种方法的优缺点，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。本章节详细的介绍了文本进行信息抽取时涉及的相关理论和技术，包括知识图谱的构建和存储、爬虫技术、BERT 预训练模型、依存句法分析、图卷积神经网络、数据标注策略、移动应用开发框架等。

第三章政务网站人事信息数据集构建。本章节详细的介绍了政务网站人事信息数据集的构建过程，包括数据来源和获取、数据预处理、数据标注和数据集构建等内容。

第四章基于深度学习的实体关系联合抽取模型。本章节对 CasRel 模型进行介绍，提出 GCN-CasRel 实体关系联合抽取模型，并在第三章构建的数据上对模型进行对比试验和消融实验，实验证明 GCN-CasRel 模型在三个实验指标上都得到了提升。

第五章政务网站人事信息知识图谱构建。本章节主要介绍政务网站人事信息

知识图谱构建过程，为了提高知识图谱构建效率，设计开发了人事信息实体关系联合抽取微信小程序，并在小程序抽取的三元组信息基础上构建了政务网站人事信息知识图谱。

第六章总结与展望。本章节总结了全文的研究内容、分析本文研究中模型和移动应用的问题和不足以及针对这些问题的解决方法，并指出进一步的研究方向。

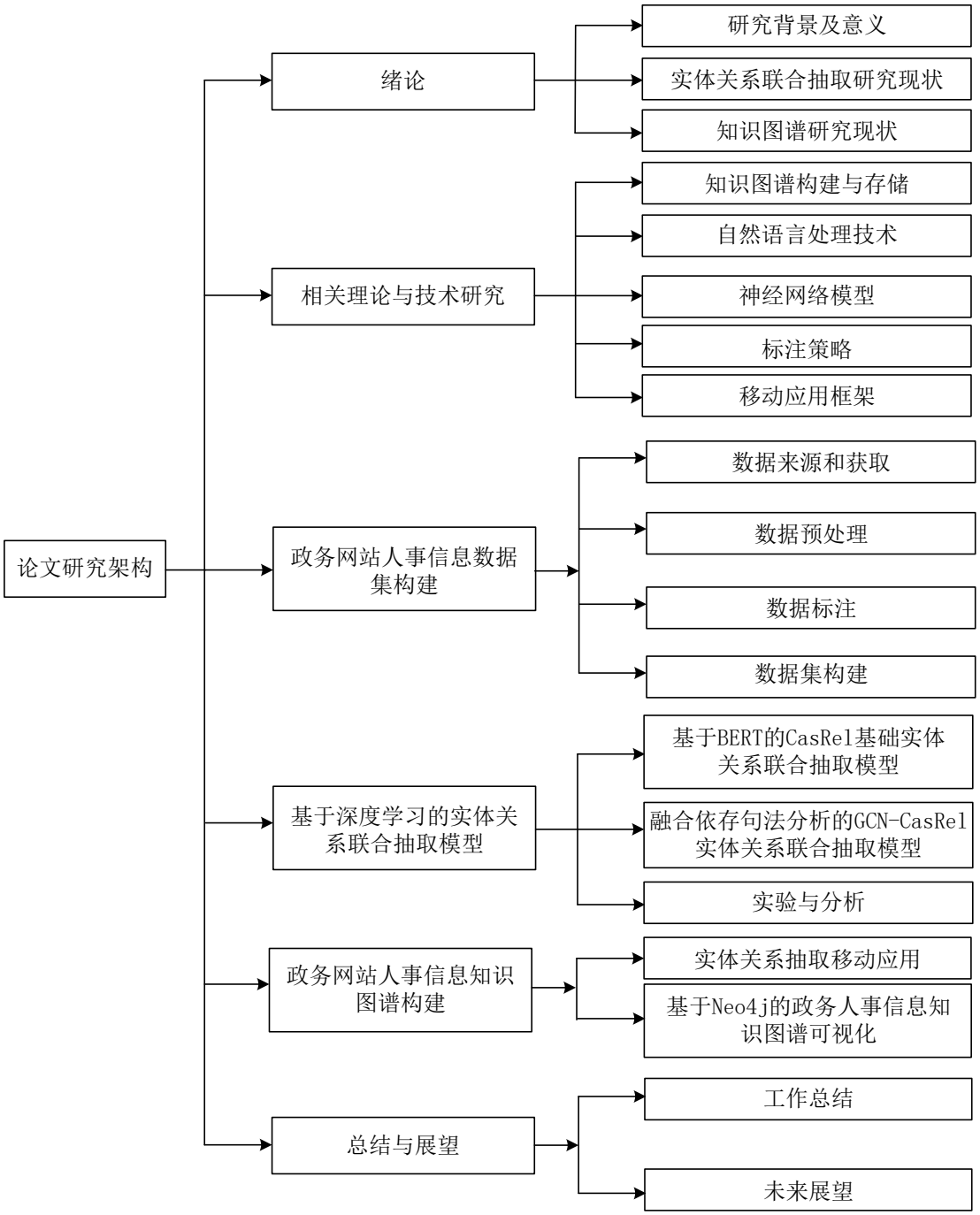

图 1.1 研究组织结构

1.5 本章小结

本章首先阐述了论文的研究背景及意义，接着总结了相关内容的国内外研究现状，详述了信息抽取的核心技术——实体关系抽取的各个技术特点和优劣，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

2.1 知识图谱构建与存储

2.1.1 知识图谱构建框架

知识图谱的构建方式分为两种：自底向上的构建和自顶向下的构建。自底向上的方法是在开放的数据集中抽取出实体、关系、属性等信息，完善相应的数据层，再将抽取出来的信息添加到数据层中，继而构建一个完整的知识图谱，目前绝大部分的知识图谱都是采用自底向上的构建方法。自顶向下的方法需要提前定义好本体模型和数据模式，根据概念和关系，以原始数据为基础进行实体抽取，将抽取好的实体添加进知识库中，这种方法需要一个现成的知识库，通常被应用于领域知识图谱的构建中。基于政务网站文本构建人事信息知识图谱的主要流程包括：数据获取、数据聚合、信息抽取、知识可视化。在数据获取环节中，通过爬虫获得政务网站的文本数据；在数据聚合环节中，对文本数据进行数据清洗、数据预处理等；在信息抽取环节中，通过训练数据集获得信息抽取模型；在知识可视化环节中，通过图数据库构建存储知识图谱，完成所有流程。知识图谱构建框架如图 2.1 所示：

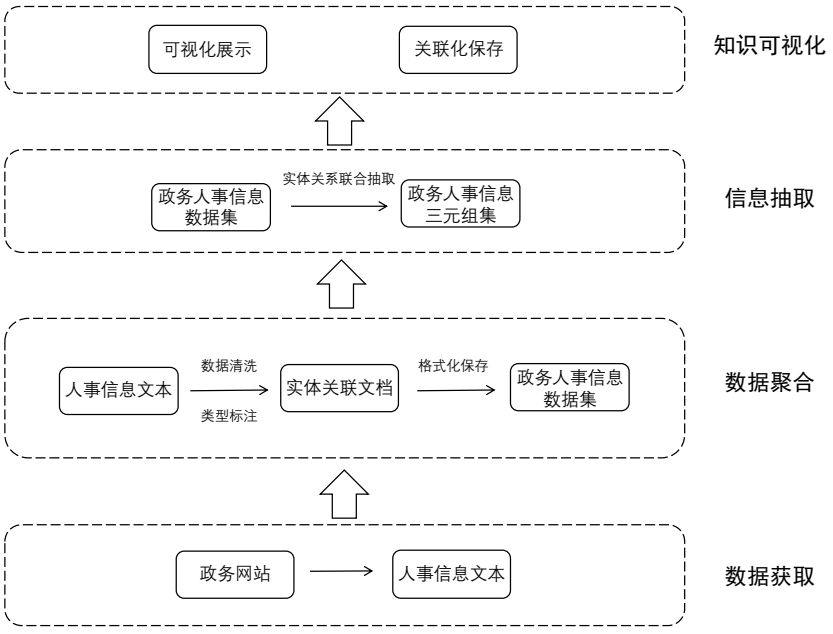

图 2.1 政务人事信息知识图谱构建框架

2.1.2 知识图谱存储

目前主流的知识图谱存储方式主要是 RDF 存储和图数据库存储。资源描述框架(Resource Decription Framework，RDF)[63]是用于描述网络资源的 W3C 标准，其描述的表现形式为 RDF 三元组，三元组由主体(Subject)、谓语(Predicate)、客体(Object)构成，通过此形式描述资源之间的关联，RDF 存储就是将数据以三元组的形式进行存储。图数据库存储是以边和节点表示数据的存储方式，Neo4j[64]是一种主流的图数据库，与其他存储方式相比较，Neo4j 有以下优点：

（1）具有很强的灵活性和适应性。不同于其他的存储方式，在 Neo4j 数据库中，数据是以图的结构进行存储的，节点的属性、类型都是分开存储，这样的设计使得数据库可以快速的应对任务的变化，灵活的对数据库的数据进行增删改。并且 Neo4j 既可以对立的展示数据，又可以嵌入到应用程序中，适应性较强。

（2）拥有更强的查询性能。Neo4j 是原生的图数据库，集成了常用的图论算法。由于图数据库能够直接存储图数据，可以直接使用图的遍历算法，能够就近的搜索查询节点邻近有限数据，不需要对所有数据进行搜索，因此无论数据本身的大小，Neo4j 数据库都能够快速的查询需要的节点。并且随着数据库中数据的增长，数据库性能也不会随之衰减。

（3）支持多语言开发。Neo4j 数据库是用 Java 语言和 Scala 语言编写而成的一种 NoSql 数据库，本身支持 Java 语言开发的需求。在此基础上，数据库可以通过内在驱动 py2neo 进行 Python 语言的开发，具有较强的开发性和丰富的开发场景。

2.2 自然语言处理技术

2.2.1 数据爬取

数据爬取普遍采用网络爬虫的方式从互联网网站上获取特定的数据，数据类型可以是链接、文字、图片等，网络爬虫的主要流程包括发送请求、获取网页、解析网页、存储数据等步骤。Python 语言经常被用于爬虫算法的开发，主要原因在于 Python 拥有强大丰富的库[65]，许多功能只需要调用库函数就能实现。在发

送请求的环节，通过调用 Requests 库对网页进行 get 请求，接着使用 Response 接受 get 请求的结果。在解析网页内容的环节，Python 语言有 BeautifulSoup、Re、Xpath 三个库供选择。BeautifulSoup 方法是将网页内容转换成特定结构，再利用查询命令对网页进行解析。Re 方法是通过正则表达式的创建解析网页内容。Xpath 方法通过路径匹配完成对内容的提取。除了编写程序进行爬虫外，越来越多的第三方软件也实现了对网页内容自动化获取，这些工具包括八爪鱼、WebCopy 等，这些工具不仅可以使用内置模板进行网页内容的爬取，而且可以自定义模板爬取特定网站的网页信息，并且能够在数据获取过程中完成对数据的简单清洗，简化数据预处理流程。

2.2.2 Transformer 模型

Transformer[66]于 2017 年由 Vaswani 等人提出，最初该模型被应用于处理文本序列的编码问题。以往在进行自然语言处理任务中的特征编码时，研究者普遍运用循环神经网络或卷积神经网络解决编码的问题，但是这两种模型以及他们的变体都有着明显的缺陷。比如在使用循环神经网络模型时，梯度消失问题严重的限制了模型提取语义特征的能力；在使用卷积神经网络时，卷积核的大小也会影响模型提取语义特征的能力。Transformer 模型创新性地使用了注意力机制组建网络结构，组成了多个编码-解码结构(Encoder-Decoder)进行语义特征提取，舍弃了以往使用的循环神经网络和卷积神经网络，在提升模型性能的同时还提升了响应速度。Transformer 模型结构如图 2.2 所示：

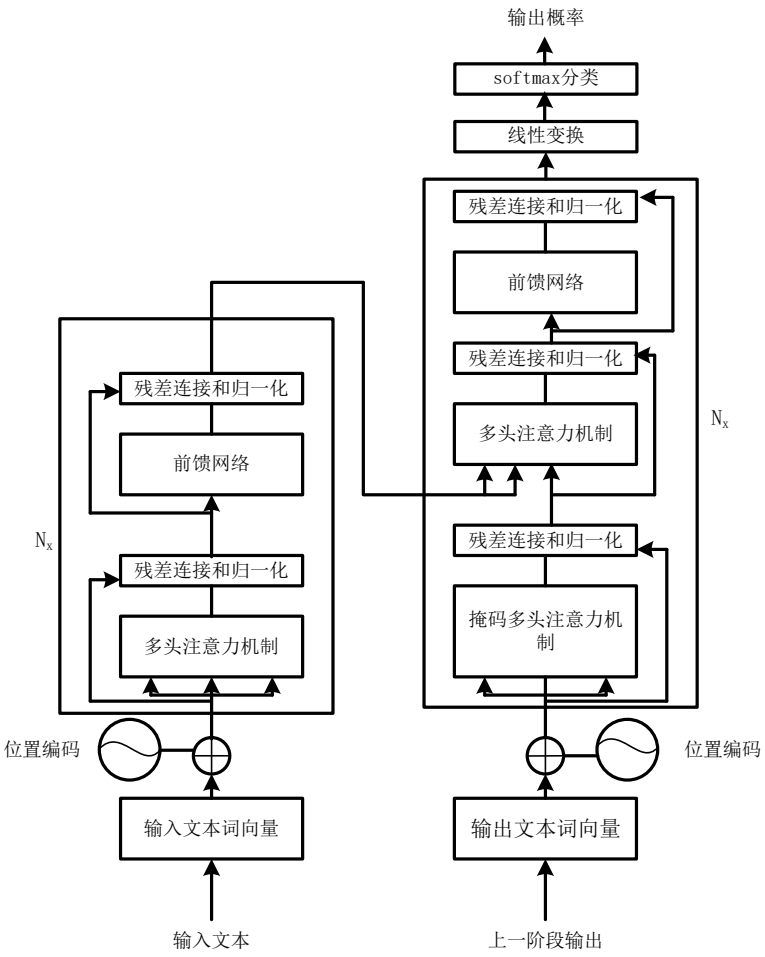

图 2.2 Transformer 模型结构

（1）注意力机制

注意力机制(Attention)最早由图像领域研究者提出，主要目的是在大量的信息中筛选出重要的或显著的信息，其原理是模仿了人类视觉注意力机制，人类在观察图像时会快速对整体进行扫描并且把注意力聚焦于图像的重点位置，给予显著的区域更多的关注，忽略不重要的区域。注意力机制通过对输入的信息赋予不一样的权重，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，并且可以根据不同的情况动态的调整权重。比如在自然语言处理的情感分类任务中，注意力机制就可以给予“好吃”“好看”“好心情等”词汇更高的权重，以此判别语句的情感。

注意力机制包含 Query 向量、Key 向量和 Value 向量，Query 表示 Target 的元素，Key-Value 对表示 Source 中的元素，Attention Value 表示计算出的 Attention

值。注意力机制的计算过程为：先输入 Query，将 Query 与 Key 进行对比，计算 Query 和每一个 Key 的相似度，进行归一化处理得到 Value 权重系数，最后综合相似度加权求和得到最后的值。计算注意力机制的计算公式如（2-1）下：

$$Attention(Query, Key, Value) = \sum_{i=1}^{n} similiarity(Query, Key_i) \cdot Value_i \tag{2-1}$$

其中n为 Key-Value 对的数量，常用的相似度函数有点积相似度、余弦相似度和多层感知机。对相似度进行归一化处理的 softmax 函数公式如（2-2）所示：

$$\alpha_i = Softmax(Sim_i) = \frac{e^{Sim_i}}{\sum_{j=1}^{n} e^{Sim_j}} \tag{2-2}$$

其中α_i为Key_i对应的$Value_i$权重系数，进行加权求和可以得到 Attention 值，公式如（2-3）所示：

$$Attention(Query, Key, Value) = \sum_{i=1}^{n} \alpha_i \cdot Value_i \tag{2-3}$$

传统的注意力机制只计算 Source 和 Target 之间的联系，自注意力机制(Self-Attention)在此基础上还会计算 Source 和 Target 端内部的 Attention 联系。Self-Attention 的计算获取 Source 和 Target 两端内部词与词之间的关联度，再接着计算两端之间词语的关联度。计算公式如（2-4）所示：

$$Attention(Query, Key, Value) = Soft\max(\frac{QK^T}{\sqrt{d_k}}) \cdot Value \tag{2-4}$$

其中d_k表示 Query、Key 和 Value 的向量维数。

多头注意力机制(Multi-Head Attention)包含多个 Self-Attention 层，将输入向量分成h组，分别对每组进行 Self-Attention 计算，最后将计算出的结果进行拼接得到 Multi-Head Attention 值，扩展了模型关注不同位置的能力。多头注意力机制的计算公式如（2-5）和（2-6）所示：

$$head_i = Attention(QueryW_i^Q, KeyW_i^K, ValueW_i^V) \tag{2-5}$$

$$MultiHead(Query, Key, Value) = Concat(head_1, head_2 ... head_h) \cdot W^Q \tag{2-6}$$

其中$head_i$表示并行的自注意力模块，W^Q为参数矩阵。

（2）Transformer 编码器

Transformer 编码器的作用在于对输入进行指定的特征提取，为解码提供有效的语义信息。整体结构如图 2.3 所示，编码器由 6 个相同的单元组成，每个单元都有一个多头注意力模块和一个前馈神经网络模块，在这两个模块后面又都分别增加了残差连接和归一化操作，这样可以避免梯度消失，加快模型收敛。

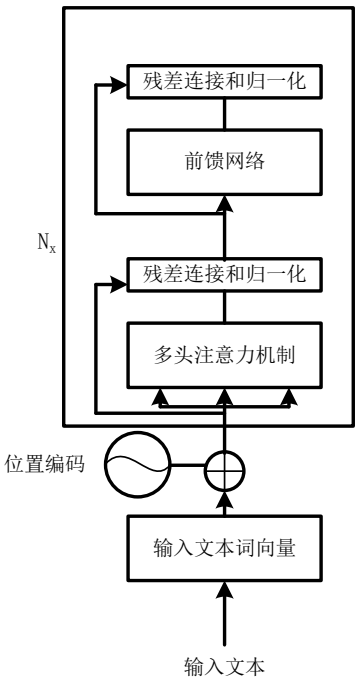

图 2.3 Transformer 编码器

编码器将词向量矩阵和位置嵌入矩阵转化为 Q、K 和 V 矩阵，经过多头注意力机制计算得到矩阵 X，再将新得的矩阵经过残差连接和归一化操作得到输出，公式如（2-7）所示：

$$x = Layernorm(X + sublayer(X)) \tag{2-7}$$

由多头注意力模块处理而来的输出作为前馈神经网络模块的输入，公式如（2-8）所示：

$$FFN(x) = \max(0, xW_1 + b_1)W_1 + b_2 \tag{2-8}$$

其中 *FFN* 表示前馈网络。

（3）Transformer 解码器

Transformer 解码器的作用在于从编码的表示中检索信息，其结构与编码器的结构类似，如图 2.4 所示，同样有 6 个功能相似的单元，都用到了自注意力机制和前馈神经网络。与编码器不同的地方在于，自注意力机制的计算结果还需要和编码器的输出再经过一次自注意力机制的计算，之后再进入前馈神经网络。而且由于解码器的输出带有时序效果，还需要在模型中加入掩码注意力机制，防止文本序列的某个位置提前得到后面位置的信息。

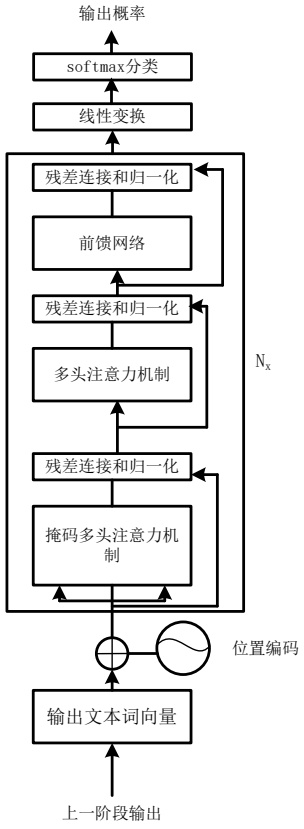

图 2.4 Transformer 解码器

随着技术和需求的发展，Transformer 模型广泛的应用于智能问答、情感分类、实体识别等自然语言处理领域。同时，在计算机视觉领域，Transformer 模型也大范围的被使用，比如图像分类、分割等任务，均取得了较好的效果。除此之外，Transformer 模型也是各种预训练模型的重要组成部分，基于 Transformer 的预训练模型如 BERT、ERNIE 等，都大幅了提升了在处理相关任务时的模型性能。

2.2.3 BERT 预训练模型

词嵌入(Word Embedding)技术是自然语言处理任务中一个关键技术，通过将输入的文本信息转化成低维稠密向量，相似的词语具有相近的词向量，这样在计算机能够识别文本信息的同时，也避免了维度灾难的问题。目前常用的词嵌入模型有 Word2vec[66]、 GloVe[68]等。但是传统的词向量模型都是固定表征的，无法解决一词多义的问题，如“算账”一词既有计算的意思又有与他人计较的意思。在这种情况下，BERT 模型的应用有效的解决此类问题。

2018 年 Google 公司提出了 BERT 预训练模型[69]，随后被研究者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 2.5 所示，基础架构由多个 Transformer 的 Encoder 部分组成，在文献[69]中提出了 BERT_base 和 BERT_large 两个模型，其中 BERT_base 由 12 层的 Encoder 组成，BERT_large 由 24 层 Encoder 组成。

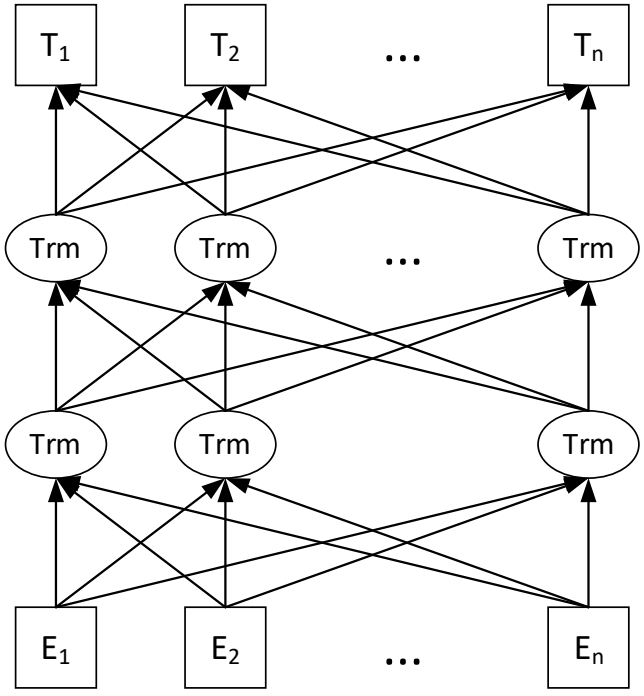

图 2.5 BERT 模型结构

BERT 将输入句子中的每个字符都用三个嵌入(Embedding)表示：单词嵌入(Token Embedding)、类型嵌入(Segment Embedding)和位置嵌入(Position

Embedding)。 Token Embedding 中[CLS]字符用于分类，对应的输出向量包含句子的关系信息，[SEP]字符用于识别句子的边界。Segment Embedding 用于区分句子的类型。Position Embedding 将字符的位置信息表示成特征向量，用来编码输入序列的顺序性，BERT 输入如图 2.6 所示。

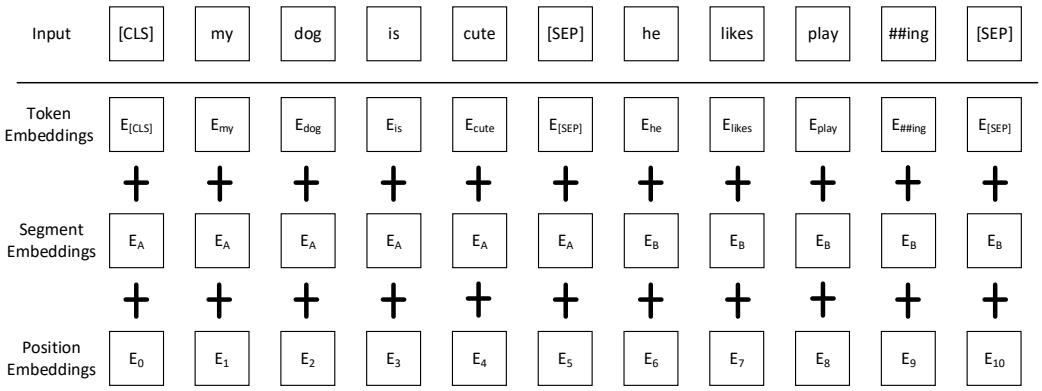

图 2.6 BERT 模型输入表示

BERT 预训练模型采用 Transformer 的 Encoder 部分，实现序列文本的双向编码，相比于其他模型使用 Transformer 中 Decoder 部分实现单向编码，这样的机制使得模型获得整个句子的语义，能够更好的完成特征提取任务。BERT 预训练模型在训练过程中同时结合了两种机制，分别是掩码语言模型(Masked Language Model，MLM)和下句预测(Next Sentence Prediction，NSP)。

（1）掩码语言模型

掩码语言模型借鉴了完形填空任务的思想，在单词序列输入之前，随机挑选15%的单词被[MASK]标记替换，然后模型结合其他没有被掩码的单词上下文预测被掩码的原单词。而在被掩码的 15%单词中，有 80%的单词被[MASK]标记替换，10%的单词被随机替换成别的单词，剩下的 10%的单词保持不变。掩码机制打破了原本的文本信息，在训练过程中让模型从上下文获取信息，使预测出的词汇无限逼近原本的词汇，这种方式让模型有着更深的语义理解。

（2）下句预测

在自然语言处理领域，有很多任务与判断两个句子间逻辑关系有关，而一般的语言模型并不具备判断句子间关系的能力。在 BERT 预训练模型中，有 50%的概率会将两个连续的语句作为训练样本，50%的概率将完全随机抽取的两个句

子作为训练样本，模型根据输入的两个句子，判断它们是否属于真实的连续语句。

掩码语言模型和下句预测这两个策略让 BERT 预训练模型在语义信息和句子间逻辑关系取得更好的训练效果，虽然会增加模型训练的时间，但是训练好的模型拥有优秀普适性，能够广泛运用到下游任务中。

2.2.4 依存句法分析

依存句法分析是自然语言处理中一项关键的技术，在命名实体识别、文本情感分析、智能问答等领域有着广泛的使用。依存句法分析能够表示句子的语法结构和句子中词语的依存关系，比如主谓关系、动宾关系等。一个依存关系中包含两个词语，一个核心词和一个依存词。依存句法分析以句子中的关键动词为中心，向下发散形成一个树状结构，这个结构就是依存句法树。依存句法树能够形容词语之间的逻辑关系，在一个句子中，有逻辑关系的两个词不论处于何种位置，在树状结构中，他们总是保持较近的距离，这样的表示方法能够很好的弥补文本语句表达不规范的问题。在对语句进行依存句法分析时，通常使用弧线表示依存关系，弧线两端分别连接着核心词和依存词。如图 2.7 所示，在句子“成都市市长宣布第一届中美创客比赛开幕”中，“宣布”这个动词是核心词，“市长”与“宣布”构成了主谓关系(SBV)，“宣布”与“开幕”之间构成了动宾关系(VOB)，依存弧从核心词出发，指向依存词。其他的词语与词语之间组成了不同关系的连接词，通过依存弧连接，依存弧上标注了词语之间的关系。

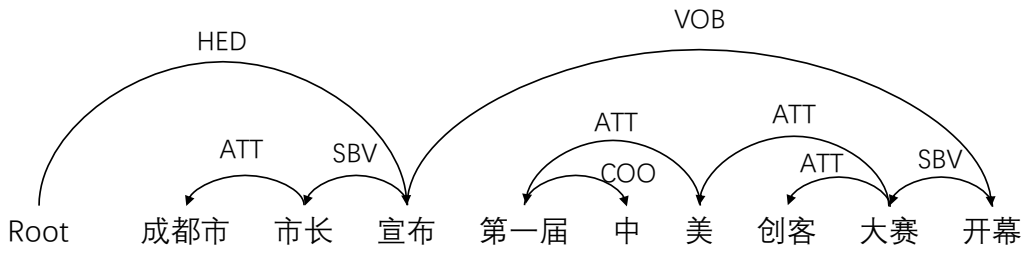

图 2.7 依存句法分析示例

目前，常用的依存句法分析工具有斯坦福的 Stanford CoreNlp 和哈工大的

LTP。依存句法分析工具采用的依存句法标签体系如表 2.1 所示：

表 2.1 依存句法分析关系标注表

关系类型	标注	描述	举例
主谓关系	SBV	subject-verb	他担任单位书记（他←担任）
动宾关系	VOB	直接宾语, verb-object	他担任单位书记（担任→书记）
间宾关系	IOB	间接宾语, indirect-object	他担任单位书记（担任→单位）
前置宾语	FOB	前置宾语, fronting-object	他什么事都做（事←做）
兼语.	DBL	double	大家称他为楷模（称→他）
定中关系	ATT	attribute	他是个大领导（大←领导）
状中结构	ADV	adverbial	不太聪明（不太←聪明）
动补结构	CMP	complement	读两遍报纸（读→两遍）
并列关系	COO	coordinate	书记和市长（书记→市长）
介宾关系	POB	preposition-object	在社区内（在→内）
左附加关系	LAD	left adjunct	书记和市长（和←市长）
右附加关系	RAD	right adjunct	同志们（同志→们）
独立结构	IS	independent structure	两个单句在结构上相互独立
核心关系	HED	head	句子的核心词

2.3 神经网络模型

2.3.1 卷积神经网络

卷积神经网络最初针对计算机视觉领域提出，与全连接神经网络相比，卷积神经网络的神经元不是每个之间都有连接的，神经元只和部分前层神经元相连，是非全连接的，所以在卷积神经网络的参数设置较少，整个模型的收敛速度大大提高。作为一种前馈神经网络，卷积神经网络的学习能力、并行能力都比较强，能够较好的提取局部信息，这些优点也使得研究者们逐渐的将卷积神经网络应用到自然语言处理领域，比如 Kim 等人[70]就将卷积神经网络应用于文本分类中。

卷积神经网络模型由输入层(Input Layer)、卷积层(Convolutional Layer)、池化层(Pooling Layer)、全连接层(Fully Connected Layer)构成。

输入层即数据的输入，卷积神经网络的输入可以是图像或者是文本，在处理文本任务的时候，输入层输入的是以矩阵形式存在的文本内容，内容以字为划分，矩阵的每一行就是一个字的字向量，如果是以词为单位的话，需要首先对文本内容进行分词处理，处理完后的矩阵即表示词单位的组合；卷积层是构建卷积神经网络的核心层，由滤波器和激活函数构成，主要作用是提取特征，在处理图像问题的时候，模型使用二维卷积进行特征提取，在处理文本内容的时候，因为一行代表一个字词的语义，所以模型使用的是一维卷积，通过卷积操作，模型可以识别输入内容的局部信息特征；池化层的作用是对提取到的特征进行降维处理，通过降低数据的空间尺寸，将网络参数的数量，提高计算资源的利用率，同时能够控制过拟合，常见的池化方式有最大池化(max-pooling)、平均池化(mean-pooling)和随机池化(Stochastic-pooling)，最大池化是对局部的特征取最大值，平均池化获取特征值平均值，随即池化根据概率最局部的特征值进行采样，元素值越大的概率越大；全连接层将前层计算得到的特征空间映射样本标记空间，在整个卷积神经网络模型中起到“分类器”的作用，降低特征位置的影响，提高模型稳定性。

与其他模型相比，卷积神经网络模型通过卷积层捕捉文本特征，再经过池化层选取具有代表性的特征，能够充分的表达出文本语义信息，卷积神经网络及其改进的网络在自然语言处理领域里面能够发挥重要的作用。

2.3.2 图卷积神经网络

卷积神经网络因其特殊的结构拥有良好的特征提取能力，自问世以来被广泛的应用于图像处理任务、自然语言处理任务等。卷积神经网络的参数共享和池化带来了平移不变性，即所识别的目标出现在不同的位置仍能得到相同的标签，这使得卷积神经网络在处理图像或文字等欧氏空间数据时具有良好的效果。但是在面对非欧氏空间数据，例如图数据，数据的不规则性会让具有有限维度矩阵的卷积神经网络失去效果。

日常生活中充满着大量的图结构数据，比如人际关系图谱、语句中的依存句法关系等，这些图结构数据里的节点连接各不相同，有的节点有一个连接，有的

节点有三个连接，这样的数据具有不规则性。图结构数据中，每个节点都有自己的特征，即图结构数据的节点特征；节点与节点之间也都存在联系，即图结构数据的结构特征。面对图结构数据的空间特征，普通的卷积网络效果不尽人意，这种情况下，研究者们将神经网络应用到图结构中，提出了图神经网络(Graph Neural Network，GNN)结构，在图神经网络结构中，图卷积神经网络是一个重要的研究方向。目前图卷积神经网络主要分别两种：基于谱域的图卷积神经网络和基于空间的图卷积神经网络。

（1）基于谱域的图卷积神经网络

基于谱域的图卷积神经网络以图信号处理为基础，首先利用拉普拉斯矩阵的特征向量作为正交基函数，接着对两个傅里叶变换后的图信号进行乘积计算，再将相乘的信号进行傅里叶逆变换操作。给定一个无向连接图$G=\{V,E\}$，其中 V 表示图的节点集合，E 表示连接节点的边的集合，|V|表示节点的个数，这里用 N 来表示。无向图由归一化图拉普拉斯矩阵表示，公式如（2-9）所示：

$$L=I_N-D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \tag{2-9}$$

其中D为图的度矩阵，A为图的邻接矩阵，拉普拉斯矩阵L为半正定矩阵，所以对其进行特征分解成公式（2-10）：

$$L=U\Lambda U^T \tag{2-10}$$

其中U 为特征向量，Λ 为对角矩阵。傅里叶域定义了图卷积算子，它是信号x和滤波器G的乘积，公式如（2-11）：

$$x*G_g=U\left(\left(U^Tx\right)\odot\left(U^TG\right)\right) \tag{2-11}$$

其中$\odot$表示 Hadmard 算法，可以将滤波器表示为：

$$g_\theta=diag(U^Tg) \tag{2-12}$$

则图卷积的公式可以简化为：

$$x*G_g=Ug_\theta U^Tx \tag{2-13}$$

基于谱域的卷积过程复杂，执行图卷积时需要加载整个图，这导致处理的图数据庞大时运行的效率不高。同时在基于谱域的方法中，图的傅里叶变换只能在无向图上生效，许多的图谱，比如交易网络图谱，都是有向图，这样的图结构数

据是无法进行特征提取的。除此之外，基于谱域的方法只能够对固定的数据进行卷积，因此图的节点不能发生改变，这使得该方法无法适应实际生活中大部分的动态数据。

（2）基于空间的图卷积神经网络

基于空间的图卷积神经网络通过寻找图数据节点的邻居节点，运用聚合函数获取节点特征，其思想来源就是卷积神经网络对图像的卷积运算，针对图结构数据的空间特征，通过对应邻居节点的信息更新节点依赖信息，基于空间关系定义图卷积，这使得模型能够获得图结构的任意位置有效信息。给定一个无向连接图，其中 V 表示图的节点集合，E 表示连接节点的边的集合，基于空间的图卷积操作如公式（2-14）所示：

$$H^{(l+1)} = (A+I)H^{(l)}W \tag{2-14}$$

其中 $H^{(l+1)}$ 是 $l+1$ 层图卷积输出表示，A 是图 G 的邻接矩阵，$H^{(l)}$ 是 l 层图卷积输出，W 是待训练参数。

基于空间的方法通过不断聚合邻居节点的信息对中心节点特征进行更新，可以有效的处理大规模的图结构数据，其灵活性也符合实际任务中的动态需求。基于空间的图卷积受到越来越多的关注，图注意力网络(Graph Attention Network，GAT)等模型都广泛的运用到研究任务中。

2.4 标注策略

在实体关系抽取任务中有一个重要环节，就是在基础文本信息基础上生成标注序列，这有助于实体识别，提升三元组抽取的准确率。本节对目前主要的标注策略进行归纳总结，包括以下几种：序列标注方法、指针标注方法和片段排列方法。

2.4.1 序列标注方法

在自然语言处理任务中，序列标注是最经典的标注策略，其工作原理是为文本信息中的每个字符打上标签，基于字符级别进行多分类，通常结合条件随机场(Conditional Random Field，CRF) 处理分类问题，例如 LSTM+CRF/BERT+CRF。

这种标注策略方法简单成熟，在简单文本中正确率也较高。但是因为序列标注的方法只形成一个标注序列，难以有效解决实体嵌套的问题，比如在实体“北京天安门”中还存在“北京”这个实体，序列标注的方法只能对其中一个实体进行标注，因此该方法适合长度较短，实体简单分明的文本信息。表 2.2 列举了主流的序列标注方法及其含义说明。

表 2.2 主流的序列标注方法及说明

序列标注方法	说明
BIO	B-begin：实体的开头部分 I-inside：实体的中间或结尾部分 O-outside：非实体部分
BMES	B-begin：实体的开头部分 M-middle：实体的中间部分 E-end：实体的结尾部分 S-single：表示实体的单个字符
BIOES	B-begin：实体的开头部分 I-inside：实体的中间部分 O-outside：非实体部分 E-end：实体的结尾部分 S-single：表示实体的单个字符

2.4.2 指针标注方法

由于序列标注只有一条标签序列，无法解决三元组问题，因此更多的标注策略被研究者们应用。序列标注方法需要标注所有的实体部分的头部字符、尾部字符和中间字符，并且用 O 来表示非实体部分，与序列标注不同，指针标注方法用 0 和 1 对文本信息进行标注。假设词向量长度为 $sentence\_len$，在指针标注方法中，会使用两个标签序列构建成一个 $sentence\_len \times 2$ 的矩阵，矩阵的一行表示实

体开头部分的标签，另一行表示实体尾部的标签，实体的开头和结尾部分都用 1 来表示，实体头部的 1 即头部指针，实体尾部的 1 即尾部指针，实体的中间部分和非实体部分都用 0 来表示，头部指针和尾部指针之间的部分即一个实体的序列表示，通常通过两个标签序列形成的矩阵既能够表达一个实体，指针网络的结构如图 2.8 所示。为了解决文本信息中的三元组重叠问题，可以层叠多个类似的指针网络，因为每个指针网络都是彼此独立的，所以在有三元组重叠的语句中，层叠指针网络也能够识别出多个实体。

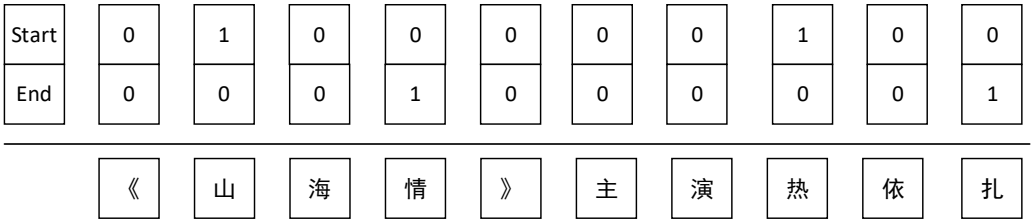

图 2.8 指针标注示例

2.4.3 片段排列方法

不同于其他的标注方法基于字词级别进行实体识别，继而间接的提取实体特征，片段排列的方法是直接对整个文本信息中的实体进行特征提取。片段排列的方法是基于枚举的思想，通过枚举出所有可能的排列片段，接着对所有的排列片段进行分类判定，最后存在的实体进行标注。假设文本的长度为n，将以步幅为 1 按照1~n进行滑动窗口操作，对文本按照顺序进行划分，如图 2.9 所示，长度为 1 的片段有n个，长度为 2 的片段有$n-1$个，长度为 3 的片段有$n-2$个，长度为n的片段有一个，可以得出对于长度为n的文本信息共有$\frac{n(n+1)}{2}$个排列片段，因此如果需要标注的文本信息过长时，会产生大量的冗余，影响模型标注效率。

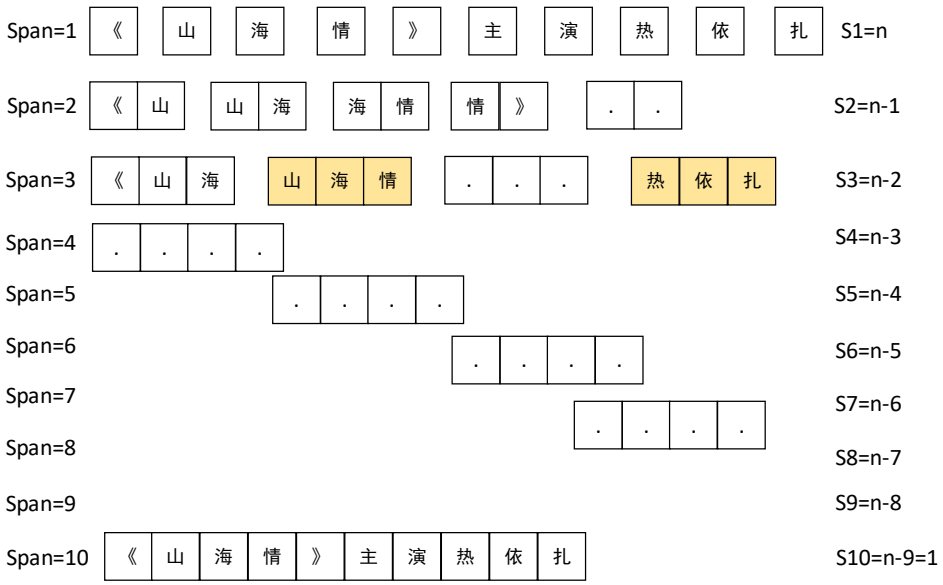

图 2.9 片段排列示例

2.5 移动应用框架

2.5.1 Flask 框架

Flask 是一个使用 Python 编写的轻量级 Web 应用框架，具有简单和易扩展的特点，所有又被称为微框架。Flask 框架依赖 WSGI 工具集 Werkzeug 和 Jinja2 模板引擎，其中 Werkzeug 是一个 WSGI 工具包，可以作为一个 Web 框架的底层库，里面封装了如 Request 等 Web 框架的指令，在 Flask 框架中 Werkzeug 实现了 Socket 服务端的功能，接收 http 请求并对请求进行预处理，进而触发 Flask 框架；Jinja2 模板引擎是 Python 下一个被广泛应用的模板引擎，是由 Python 编写的模板语言，他的设计思想来源于 Django 的模板引擎，并扩展了其语法和一系列强大的功能，是 Flask 内置的模板语言，其主要功能是对模板和数据进行渲染，将渲染后的字符串返回给用户浏览器。Flask 框架轻巧、简洁，不需要复杂的配置，能够基于 MVC 模式进行开发，对于开发者而言能够有良好的开发体验。

2.5.2 微信小程序

微信小程序是基于微信开发平台的新型开发软件，相比于传统 APP 需要下载才能使用不同，微信小程序只要通过二维码扫描或者通过名称搜索就可以查询使用。微信小程序主要是 B/S 架构，即浏览器/服务器架构，用户通过浏览器向服务器发送请求即可获取服务，这种架构的优点是维护部署方便、对硬件要求低。在前端开发方面，微信小程序有 WXML(WeiXin Markup Language)，WXSS(WeiXin Style Sheets)，在功能上分别对应传统前端开发的 HTML 和 CSS，使用 JavaScript 进行开发。服务器端兼容 Java、Python 等主流开发语言。数据库方面，小程序内置 MySQL 数据库，很好的适配了服务器端需求。微信小程序在架构上分为视图层和逻辑层，视图层由 WXML 和 WXSS 编写，主要功能是修改表达页面样式，决定组建在小程序页面的排列位置；逻辑层由 JS 语言编写，主要功能是页面逻辑处理和服务器端的网络请求。逻辑层将数据进行处理后发送给视图层，同时接收视图层的事件反馈。视图层将逻辑层的数据展示到界面，同时将视图层的事件发送给逻辑层。除此之外，在微信小程序开发里，JSON 用于状态栏、标题栏等的配置。

2.6 本章小结

本章主要对知识图谱构建和移动应用的相关理论和技术做出介绍，首先对知识图谱的构建流程和存储方式进行了概述，接着对爬虫技术、Transformer 模型、BERT 预训练模型、依存句法分析等自然语言处理相关技术进行了详细的阐述，然后介绍了知识图谱构建涉及的深度学习模型，对其原理进行了讲解，最后对实体关系抽取中的标注策略和构建移动应用框架的技术进行了说明。本章节的理论研究对下文实体关系抽取模型的提出、知识图谱构建和移动应用的开发奠定了强有力的理论基础。

3 政务网站人事信息数据集构建

自然语言处理领域的数据集分为公开数据集和自制数据集，面向政务人事信息三元组抽取这一特定任务目前没有质量较好的公开数据集，因此需要针对本文的实验自行构建文本数据集。本章围绕政务网站人事信息数据集的构建展开阐述，主要内容包括数据来源和获取、数据预处理、数据标注和数据集构建等内容。

3.1 数据来源和获取

本文的任务是在政务网站的公开文本上提取人物相关的实体和关系，因此为了保证数据的准确性和有效性，本文数据主要来自多个省市各级政府部门网站信息公开栏目中文本，这些文本涉及省、市、区各级机关和各个部门中有关人事信息的文本数据。为了让数据种类更加丰富和多元，除了以上文本数据，在原始数据中还包括了学校、医院等事业单位的人事信息数据以及部分公开数据集中与本文实验有关联的人物数据。

政务网站中涉及人物信息的公开文本包括人事任免信息、相关人物日常新闻报道等，这些文本数据数量庞大，对每一条文本分别进行下载会耗费大量时间，因此本文使用 Python 编写爬虫代码，自动化导出相关的文本数据。首先查看 HTML 网页代码，通过查看网页结构定位对应的文本目标标签，再使用 Python 语言编写爬虫代码，利用爬虫程序下载文本数据所在页面的文本。

3.2 数据预处理

成功获取原始数据后需要对原始数据进行预处理操作，预处理主要分为三个步骤：文本筛选、数据清洗、文本过滤。

（1）文本筛选。在下载的文本中筛选与本文所需数据关联较大的文本内容，去除关联较小的内容，保证文本数据的质量。

（2）数据清洗。检查筛选出来的文本内容，对错别字和语义不清的内容进行修改，同时删除文本中不必要的空格和标点。

（3）文本过滤。将每条文本的长度控制在 100 字符内，对超过 100 字符的内容进行判断，如果仍是与本文所需相关的内容，则将其保留并保存为新的一条

文本，如果是与本文所需无关的内容则将其去除，以此提高模型训练的效率。

3.3 数据标注

在深入研究政务人事信息文本内容的特点并结合了本文实验需求后，确定了所构建数据集中的实体类别和对应的关系类别，其中实体类别 10 种，对应的关系类别 10 种，作为构建政务网站人事信息数据集实体关系分类体系及标注规范，具体如表 3.1 所示。

表 3.1 人事信息实体类别及对应关系类别

实体 1	关系	实体 2
人物	民族关系	民族
人物	出生地	籍贯
人物	教育经历	学历
人物	所在单位	组织机构
人物	担任职务	职务
人物	政治面貌	面貌
人物	出生日期	日期
人物	毕业院校	学校
组织机构	单位所在	地点
组织机构	职位设置	职务

经过数据预处理后的文本需要标注文本中各自的实体和对应的关系，因为数据量较大，人工标注费时费力，为了提高标注效率，本文采用中文多元组联合标注工具 LAnn。LAnn 是一款基于浏览器的开源标注工具，不需要进行配置，下载源码后点击 html 文件即可进行实体关系标注，并可在 entity_type.js 和 relation_type.js 文件中根据数据集的要求修改实体类别和关系类别。LAnn 可以直接导入文本文件，在标注界面对文本进行实体关系标注后可以保存语料，系统会生成一个包含标注三元组的结果文件，将文件导出即可用于制作数据集。LAnn 标注界面如图 3.1 所示。

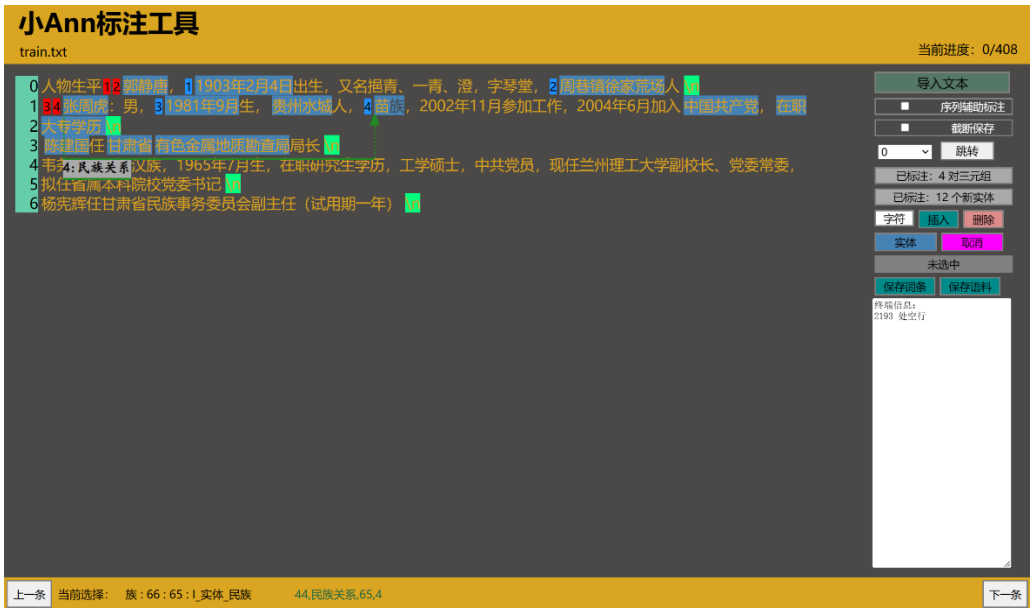

图 3.1 LAnn 标注界面

最后将标注过的数据转换成 SPO 三元组(Subject-Predicate-Object)的形式，每一条数据都包含原文和相应的三元组序列，并以 JSON 文件进行保存用于后续的实验开发，三元组形式如表 3.2 所示。

表 3.2 三元组数据示例

序号	三元组数据
示例 1	{"text": "人物生平郭静唐，1903 年 2 月 4 日出生，又名挹青、一青、澄，字琴堂，周巷镇徐家荒场人", "spo_list": [{"predicate": "出生日期", "object_type": "日期", "subject_type": "人物", "object": "1903 年 2 月 4 日", "subject": "郭静唐"}, {"predicate": "出生地", "object_type": "籍贯", "subject_type": "人物", "object": "周巷镇", "subject": "郭静唐"}]}
示例 2	{"text": "蔡辉同志一年试用期已满，经考察，甘肃省人民政府决定：蔡辉任甘肃省人民医院院长", "spo_list": [{"predicate": "担任职务", "object_type": "职务", "subject_type": "人物", "object": "院长", "subject": "蔡辉"}, {"predicate": "所在单位", "object_type": "组织机构", "subject_type": "人物", "object": "甘肃省人民医院", "subject": "蔡辉"}, {"predicate": "职位设置", "object_type": "职务", "subject_type": "组织机构", "object": "院长", "subject": "甘肃省人民医院"}, {"predicate": "单位所在", "object_type": "地点", "subject_type": "组织机构", "object": "甘肃", "subject": "甘肃省人民医院"}]}

表中 text 表示原来的文本内容，sop_list 表示三元组的内容，每一个三元组内容都用一个括号表示，predicate 表示预测的关系，subject 和 object 分别表示两个实体，subject_type 和 object_type 分别表示对应的实体类型。

3.4 数据集构建

通过上文对原始数据进行获取并且完成预处理后，数据集的文本数达到 4500 条，每条数据集都包含有若干个 SPO 三元组，总共有十个实体类别和十个关系类别。政务网站人事信息数据集关系类别信息如表 3.3 所示。

表 3.3 数据集关系类别信息

关系类别	实体 1	实体 2	关系个数
民族关系	人物	民族	744
出生地	人物	籍贯	633
教育经历	人物	学历	520
所在单位	人物	组织机构	3508
担任职务	人物	职务	3549
政治面貌	人物	面貌	627
出生日期	人物	日期	645
毕业院校	人物	学校	336
单位所在	组织机构	地点	3468
职位设置	组织机构	职务	3479

3.5 本章小结

本章详细介绍了政务网站人事信息数据集的构建过程，包括数据来源和获取、数据预处理、数据标注等步骤。在数据标注部分介绍了数据实体关系分类标准、数据标注方法以及最终的数据呈现形式。本章构建的数据集将用于下文的深度学习实验和知识图谱构建。

4 基于深度学习的实体关系联合抽取模型

通过第二章对知识图谱构建相关技术的介绍可知，实体关系抽取是构建知识图谱的关键任务。政务网站上的文本内容往往以非结构化数据的形式呈现，利用深度学习模型可以在这些非结构化数据中抽取出结构化三元组信息，只有以三元组信息为基础才能构建相关的知识图谱。本章的主要研究内容为基于CasRel模型构建实体关系联合抽取模型GCN-CasRel，并在第三章构建的数据集上进行实验对比。

4.1 基于 BERT 的 CasRel 基础实体关系联合抽取模型

实体关系抽取是信息抽取任务中关键的步骤，其抽取的结果以<Subject，Predicate，Object>即 SPO 三元组的形式呈现，而三元组是构建知识图谱的基本组成部分，所以实体关系抽取的性能也决定着知识图谱的质量。基于深度学习的实体关系抽取方法分为两种：流水线法和联合法。流水线法是一种传统的实体关系抽取方法，它的主要流程是先利用命名实体识别模型识别出文本中的实体，接着在已经识别好的实体基础上利用关系抽取模型抽取文本中的三元组信息，这种方法存在着实体识别效果带来的误差积累问题，同时流水线法无法解决三元组重叠问题。

三元组重叠是实体关系抽取任务中一个常见的问题，主要有 SEO 和 EPO 两种重叠类型。SEO(Single Entity Overlap)是指多个实体跟同一个实体存在关联关系；EPO(Entity Pair Overlap)是同一对实体之间存在多种关系。以“蔡辉任甘肃省人民医院院长”为例，其中能够抽取出<蔡辉，担任职务，院长><蔡辉，所在单位，甘肃省人民医院><甘肃省人民医院，单位所在，甘肃><甘肃省人民医院，职位设置，院长>等几个三元组，其中“蔡辉”这个人物就与多个实体有关联关系，即存在着 SEO 的重叠问题。本文面向的文本是人事信息文本，以人物为核心必然会存在着较多的重叠三元组，因此传统的流水线方法效果有限。而实体关系联合抽取方法是对实体和关系同步抽取，能够较好的解决了三元组重叠的问题。

传统的关系抽取模型将实体间的关系(relation)看作是分配给主体(subject)和客体(object)的离散标签，即 $f(s,o)\rightarrow r$，但是这样的方法面对存在重叠数据的文

本时无法准确判断多个实体对之间的关系。而 CasRel 模型提出一种端到端的级联二进制标记框架，将关系建模为主体映射到尾实体的函数，即 $f_r(s) \to o$，在这种框架下三元组的提取分为两步，首先通过预训练 BERT 模型获取所有可能的主体；接着针对所有的主体，利用特定的关系识别器识别所有可能的关系，并找到相对应的客体。

CasRel 模型包含两个部分：编码层和级联解码器。编码层负责给文本信息编码信息特征，解码端由主体标注器和若干个特定关系的客体标注器组成，主体标注器用于识别所有可能存在的实体，特定关系的客体标注器用于识别与关系相对应的客体并提取三元组信息，模型结构如图 4.1 所示。

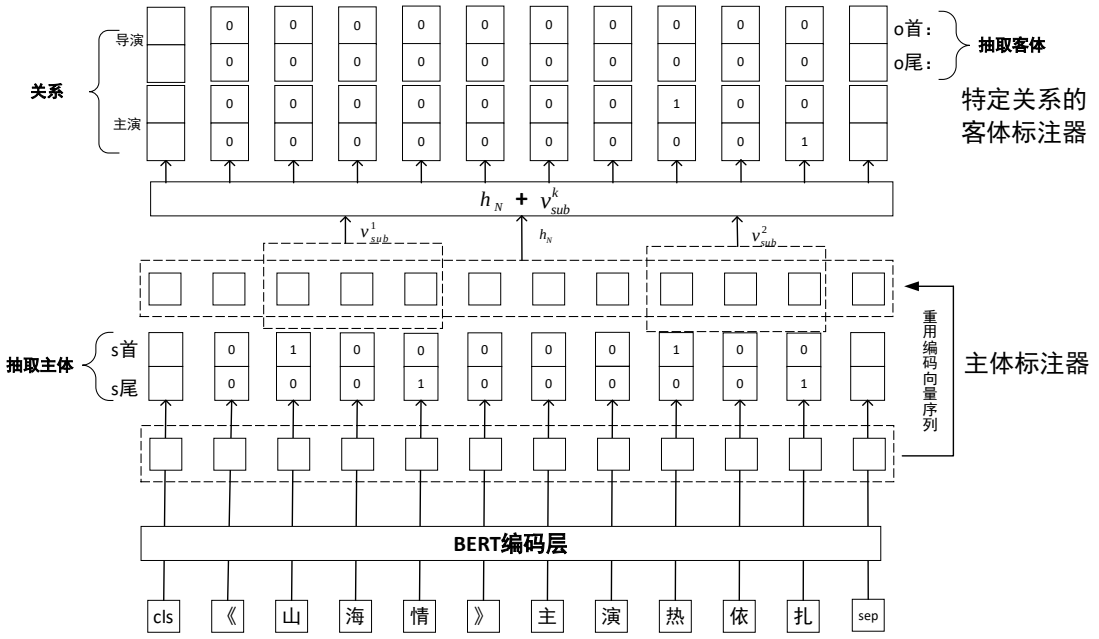

图 4.1 CasRel 模型架构

编码层的主要任务是对输入的文本信息完成编码操作，通过将信息编码成特征向量提取文本信息特征。CasRel 模型中采用 BERT 模型进行编码，用Tran(X)来表示 Transformer 模块，X表示文本的输入向量，具体公式如下所示。

$$h_o = SW_s + W_p \tag{4-1}$$

$$h_\alpha = Trans(h_{\alpha-1}), \alpha \in [1, N] \tag{4-2}$$

其中，S 表示输入文本的独热编码矩阵，W_s 表示文本的词嵌入矩阵，W_p 表

示文本的位置嵌入矩阵，h_α 表示隐藏状态向量，N 表示 Transformer 模块的数量，因为预训练模型由 12 层 Transformer 构成，所以此处N=12。

通过编码层后需要对编码信息进行解码，CasRel 模型构造了一个级联解码器，通过两个级联步骤对三元组进行提取。级联解码器由主体标注器和若干个特定关系的客体标注器构成，通过主体标注器尽可能地查找所有存在的主体，再针对每一个查找到的主体寻找其可能对应的关系，最后根据实体和关系找到相关联的客体。通过以上步骤完成级联解码，抽取文本中的三元组信息。

（一）主体标注器

主体标注器对编码层输出的编码向量h_N进行解码来识别所有可能存在的主体。针对主体的起始和主体的结尾分别建立一个二进制分类器来识别主体的起始位置和结束位置，构造一个$sentence_len * 2$的矩阵组成指针标注网络，通过为h_N中每一个向量分配二进制标记，当预测到主体的起始位置或者结束位置时，会在对应的二进制分类器中标记“1”，其余部分标记为“0”。具体过程是对每个输入子词计算主体起始位置和结束位置的可能性，设定一个阈值，超过这个阈值的就判定为“1”，低于这个位置的判定为“0”，计算过程公式如下所示。

$$p_i^{start\_s} = \sigma(W_{start} x_i + b_{start}) \tag{4-3}$$

$$p_i^{end\_s} = \sigma(W_{end} x_i + b_{end}) \tag{4-4}$$

其中 $p_i^{start\_s}$ 表示输入的文本序列中第i个位置是主体起始位置的概率，$p_i^{end\_s}$ 表示输入的文本序列中第i个位置是主体结束位置的概率，经过计算如果概率超过某个阈值，就会判定为主体的起始或者结尾，被标记为“1”，否则标记为“0”。x_i 表示输入文本的第i子词的编码表示，即$h_N[i]$，W_{start} 和W_{end} 表示可训练的参数，b_{start} 和b_{end} 表示偏置，σ 表示 sigmoid 激活函数。

如果一个输入文本中包含多个主体，则采用最近原则来匹配主体，即将识别出来的主体起始位置和结束位置按照最近的距离组合成一个主体的头和尾，然后根据起始位置和结束位置确定主体的跨度。并且如果确定了一个主体的起始位置，那么起始位置之前的内容将被屏蔽，这样能够保证主体的完整性和前后连续性，杜绝前一个主体的结尾和另一个主体的开头组成主体的可能性。如图 4.1 所示，

距离“山”最近的标记为“1”的字是“情”，二者作为主体的起始位置和结束位置，检测出一个完整主体“山海情”，同理也检测出另一个完整的主体“热依扎”，而在确定了“热依扎”这个主体的起始位置“热”时，将会屏蔽“热”以前的词，避免产生一个无法构成词汇的主体。

（二）特定关系的客体标注器

特定关系的客体识别模块由多个特定关系的客体标注器组成，每一个关系都会存在一个对应该关系的客体标注器。特定关系的客体标注器与主体标注器的结构类似，也是通过标记实体的起始位置和结束位置来确定一个完整的实体的。与主体标注器不一样的是除了经过编码层输出后的编码向量h_N，还要加入对主体标注器的主体特征，这种共享参数的方法加强了两个实体识别模块的联系。在遍历所有与主体相关联的关系类型后，通过关系的映射寻找客体，如果对应的客体不存在，就用“null”来表示。如图 4.1 所示，对主体“山海情”进行关系遍历，对于“导演”这个关系无法找出特定的客体，对于“主演”这个关系能够找出特定的客体“热依扎”。同理，对于主体“热依扎”也会有同样的客体标注器，当遍历完所有的关系后发现“热依扎”没有对应关系的客体，就用“null”来表示。在特定关系的客体标注器中，同样是通过二进制分类来标记起始位置和结束位置，计算公式如下所示。

$$p_i^{start\_o} = \sigma(W_{start}^r(x_i + v_{sub}^k) + b_{start}^r) \tag{4-5}$$

$$p_i^{end\_o} = \sigma(W_{end}^r(x_i + v_{sub}^k) + b_{end}^r) \tag{4-6}$$

其中 $p_i^{start\_o}$ 表示输入的文本序列中第i个位置是针对关系r客体起始位置的概率，$p_i^{end\_o}$ 表示输入的文本序列中第i个位置是针对关系r客体结束位置的概率。x_i 表示输入文本的第i子词的编码表示，即$h_N[i]$，W_{start}^r 和 W_{end}^r 表示可训练的参数，b_{start}^r 和 b_{end}^r 表示偏置，σ 表示 sigmoid 激活函数。v_{sub}^k 表示在主体标注器上识别到的第k个主体的编码表示，主体通常是由多个字符组成的，v_{sub}^k 与 x_i 的向量维度不一致无法直接相加，因此需要将第k个主体的起始位置和结束位置之间的平均向量表征为v_{sub}^k。

除了标记客体，特定关系的客体识别模块也会输出关系，如果检测到的主体

和客体之间不包含某种关系，则该客体的起始位置和结束位置都会标记为“0”，如果检测到的主体和客体之间包含某种关系，则会输出完整的三元组信息。

4.2 融合依存句法分析的 GCN-CasRel 实体关系联合抽取模型

CasRel 模型能够有效的解决了三元组重叠的问题，但是在进行关系抽取时，文本中的 token 往往具有复杂的语法关联，CasRel 模型在进行关系抽取时没有将此考虑进去，本文提出 GCN-CasRel 实体关系联合抽取模型，该模型在 CasRel 模型引入图卷积神经网络捕获依存句法关系信息，进一步提升实体关系抽取模型的性能。GCN-CasRel 模型总体结构如图 4.2 所示，相比 CasRel 基础模型，GCN-CasRel 模型主要有两点改进：

（1）GCN-CasRel 模型将 CasRel 使用的 BERT_base 模型进行替换，结合中文文本的特点选择 BERT-wwm-ext 预训练模型作为模型的编码器。

（2）利用图卷积神经网络对依存句法关系进行建模，将其作为先验知识融入到模型中，同时利用注意力机制对构建的依存句法树进行剪枝，保存有效的依存句法树信息，提高模型的实体关系抽取性能。

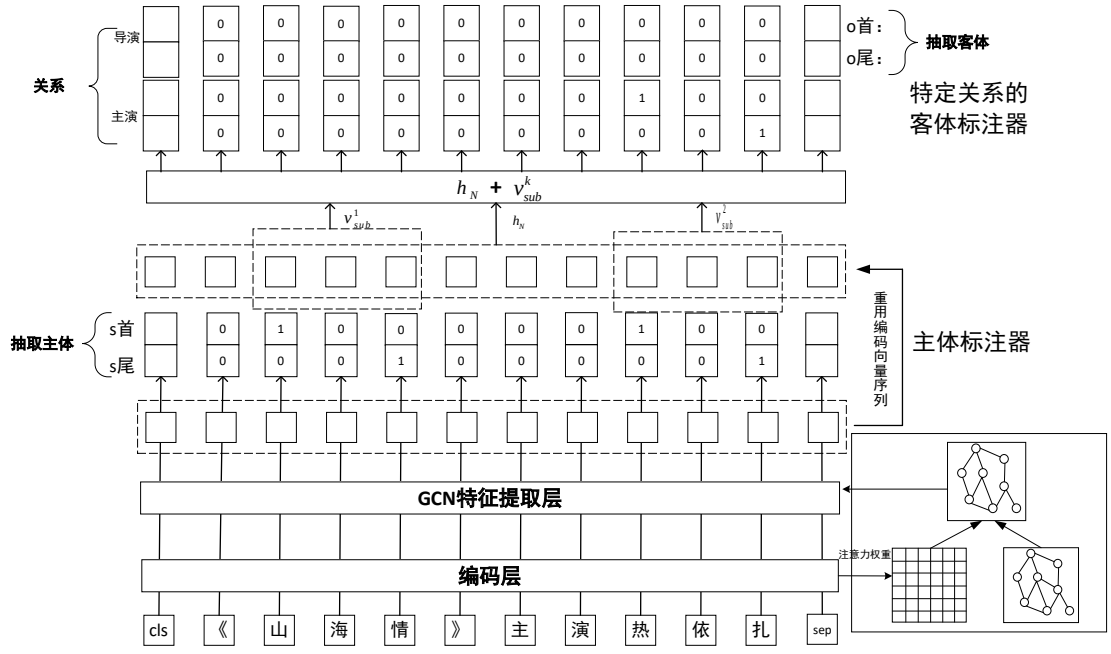

图 4.2 GCN-CasRel 模型架构

4.2.1 编码层

编码层采用 BERT-wwm-ext 预训练模型作为编码器，相比于 CasRel 模型中采用的 BERT_base 基础模型，BERT-wwm-ext 预训练模型更改了预训练样本的生成策略，BERT_base 基础模型使用[MASK]标签替换字，而 BERT-wwm-ext 模型用[MASK]标签替换一个完整的词。中文的最小 token 是字，一个词由多个字组成，包含着更多的信息，所以对全词进行掩码操作能够更好的获取上下文信息，同时 BERT-wwm-ext 模型还增加了训练数据集和训练步数，因此在中文语料训练中，BERT-wwm-ext 预训练模型拥有更好的效果。BERT-wwm-ext 预训练模型与 BERT_base 模型采用一样的构型，均由 12 层 Transformer 构成。

4.2.2 GCN 特征提取层

在中文实体关系抽取任务中，文本中的字词间往往具有复杂的语法联系，如果能够提取相关特征那么将会进一步提高实体关系抽取模型的性能。在关系抽取任务中，依存句法关系能够提供丰富的上下文信息，引导模型更好的实现实体关系抽取[70]。本层将输入文本的依存句法关系作为先验知识，利用图卷积神经网络模型进行建模，将其融入到模型之中。根据第二章的内容可以得知，依存句法关系可以通过分析输入文本的结构信息揭示词与词之间的依赖关系，以词为节点，词与词之间的结构关系作为弧，这样就能够将文本数据转化成树状结构。树是一种特殊的图，因此可以采用图卷积神经网络对其结构信息进行捕获。首先使用 Stanford CoreNlp 对输入文本进行依存句法分析，充分学习文本句法结构信息，将其转化为图的形式，图中的节点就是文本中的词，节点之间的关系就是图的边，并通过构建邻接矩阵A来表示依存句法结构，如果矩阵中第i个词与第 j 个词存在关联，则 $A_{i,j}=1$，否则 $A_{i,j}=0$。

依存句法分析虽然有助于模型识别实体之间的关系标签，但是并非所有的依存句法信息都是有用的，有些信息会成为噪声，影响模型对于关系的识别[72]。为了过滤这些噪声，本文利用注意力机制区分文本特征的重要性，对A中的边权重进行过滤，构建新的邻接矩阵 $\tilde{A}$。具体过程为：首先将多个自注意力机制头的注

意力权重矩阵平均，再设置两个阈值α和β，将第i个词到第j个词的平均注意力权重与阈值相比，大于阈值β的将在依存句法树中添加第i个词到第j个词的边，小于阈值α的将删除第i个词到第j个词的边，若平均注意力权重在二者之间的将保留依存句法树原有状态。计算公式如下所示。

$$\bar{A}^{att} = \frac{1}{h}\sum_{t=1}^{h} A_t^{att} \tag{4-7}$$

$$\tilde{A}_{i,j} = \begin{cases} 1, \bar{A}_{i,j}^{att} \geq \beta \\ A_{i,j}, \alpha < \bar{A}_{i,j}^{att} < \beta \\ 0, \bar{A}_{i,j}^{att} \leq \alpha \end{cases} \tag{4-8}$$

其中$\bar{A}^{att}$表示h个自注意力机制头的注意力权重矩阵平均，A_t^{att}表示第t个自注意力机制头的注意力权重矩阵，$\bar{A}_{i,j}^{att}$表示$\bar{A}^{att}$矩阵中第i个词到第j个词的权重，$\tilde{A}_{i,j}$表示新构建的邻接矩阵中第i个词到第j个词的权重，$A_{i,j}$表示原有依存句法树矩阵中第i个词到第j个词的权重。

在获得新的邻接矩阵后，将矩阵输入到多层图卷积神经网络中进行语义特征信息提取。以融合上下文信息的词向量作为初始节点特征，在图卷积神经网络中进行迭代，上一层的输出为下一层的输入，根据事先设置好的层数，在最后一层输入最终的节点表示，具体计算公式如下所示。

$$h_i^l = \sigma\left(\frac{1}{d_i}\sum_{j=1}^{n} \tilde{A}_{ij} W^l h_j^{l-1} + b^l\right) \tag{4-9}$$

其中，h_i^l为第l层图卷积神经网络第i个节点的表示，h_j^{l-1}为l层上一层节点的临时表示，d_i表示第i个节点的度，σ为ReLU激活函数，W^l为第l层权重参数矩阵，b^l为第l层偏置量。

4.2.3 实体关系抽取层

GCN-CasRel模型中的实体关系抽取层采用CasRel模型中的级联解码器，其中包括主体标注器和若干个特定关系的客体标注器，主体标注器的任务是识别所有可能存在的主体，特定关系的客体标注器任务是遍历各个关系找出与主体对应

的客体并抽取完整的三元组信息，具体内容见前文所述。

4.2.4 损失函数

本模型定义的损失函数为主体标注器和特定关系的客体标注器损失函数之和。给定一个句子 x，主体标注器寻找主体 s 的过程可以表现为优化似然函数，计算公式如下所示。

$$p_\theta(s \mid x) = \prod_{t \in \{start\_s, end\_s\}} \prod_{i=1}^{L} (p_i^t)^{I\{y_i^t=1\}} (1-p_i^t)^{I\{y_i^t=0\}} \tag{4-10}$$

其中L表示句子的长度。如果z为真，则$I\{z\}=1$，如果z为假，则$I\{z\}=0$。$t=start_s$时，y_i^t表示第i个词的主体起始位置，$t=end_s$时，y_i^t表示第i个词的主体结束位置。$\theta=\{W_{start}, W_{end}, b_{start}, b_{end}\}$表示可训练参数。

给定句子 x 和主体特征 s 的情况下，特定关系的客体标注器寻找与关系对应的客体 o 的过程为优化似然函数并抽取三元组，计算公式如下所示。

$$p_{\phi_r}(o \mid s, x) = \prod_{t \in \{start\_o, end\_o\}} \prod_{i=1}^{L} (p_i^t)^{I\{y_i^t=1\}} (1-p_i^t)^{I\{y_i^t=0\}} \tag{4-11}$$

其中L表示句子的长度。如果z为真，则$I\{z\}=1$，如果z为假，则$I\{z\}=0$。$t=start_o$时，y_i^t表示第i个词对应客体的起始位置，$t=end_o$时，y_i^t表示第i个词对应客体的结束位置。用“null”标记的没有存在对应关系的客体用o_ϕ来表示，$\phi_r=\{W_{start}^r, W_{end}^r, b_{start}^r, b_{end}^r\}$表示可训练参数。

整体模型的最终损失函数将主体标注器和特定关系的客体标注器的损失函数对数化后相加，如下所示，使用 Adam 梯度算法加速优化训练过程。

$$\sum_{j=1}^{|D|}\left[\sum_{s \in T_j} \log p_\theta\left(s \mid x_j\right) + \sum_{r \in T_j \mid s} \log p_{\phi_r}\left(o \mid s, x_j\right) + \sum_{r \in R \setminus T_j \mid s} \log p_{\phi_r}\left(o_\phi \mid s, x_j\right)\right] \tag{4-12}$$

其中 $s \in T_j$ 表示出现在三元组中的主体，$r \in T_j \mid s$ 表示主体引导的三元组中的关系，$r \in R \setminus T_j \mid s$ 表示除了主体引导的三元组关系外的其余关系。

4.3 实验与分析

4.3.1 实验数据

在训练模型时，已经进行数据预处理的数据集按照 4:1 的比例划分为训练集和测试集，3600 条文本用于训练，900 条文本用于测试。

4.3.2 实验评价指标

评价指标能够更好地反映出评价实体关系联合抽取模型对于抽取任务的影响，清晰的展示实验文本中正确抽取出来的三元组数量。在评价实验的指标里，一般采用查准率(Precision)和查全率(Recall)。

查准率是在所有预测为正样本中实际为正样本的比例，计算公式如（4-13）所示。

$$Precision=\frac{TP}{TP+FP} \tag{4-13}$$

查全率是在所有实际为正样本中被正确预测的比例，计算公式如（4-14）所示。

$$Recall=\frac{TP}{TP+FN} \tag{4-14}$$

在以上公式中，TP表示实际为正的样本在模型预测结果中也为正样本，FP表示实际为正的样本在模型预测结果中为负样本，TN表示实际为负的样本在模型预测结果中也为负样本，FN表示实际为负的样本在模型预测结果中为正样本。理论上查准率和查全率数值越高代表模型的性能越好，但是在实际计算中，查准率和查全率会在某些情况下产生矛盾。而在评价模型时，只考虑查准率或只考虑查全率都无法客观评价模型的性能优劣，因此为了平衡这两个评价指标，在评价模型时往往会引入一个新的评价指标$F1$值，$F1$值的计算公式如（4-15）所示。

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{4-15}$$

由公式得出，$F1$值最大值为 1，最小值为 0，在 0~1 的值域里，$F1$值越大，模型的性能越好。

4.3.3 实验环境及参数设置

本实验训练通过云算力平台 AutoDL，所用的服务器配置为：系统为 Ubuntu 20.04 系统、CPU 为 AMD Ryzen 5 4600H、显卡为 RTX 3060、12G 显存。深度学习配置的环境为 PyTorch 1.10、Python 3.8，编码格式统一为 utf-8，详细参数设置如表 4.1 所示。

表 4.1 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	0.5

4.3.4 实验结果与分析

为了验证本模型的效果，选取了主流的几种实体关系联合抽取模型在本章构建的政务人事信息数据集上进行试验。

（1）NovelTagging[73]：模型通过序列标注方法将实体关系抽取问题转换成标记问题，同时采用端到端的模型生成标注序列实现提取实体及其关系。

（2）GraphRel[74]：模型通过关系加权的图卷积神经网络考虑实体和关系之间的交互，基于给定单词特征建立完整关系图来整合实体和关系。

（3）ETL-Span[75]：模型提出的基于 span 的标记方案，将实体识别和关系抽取两个任务进一步分解为多个序列标记问题，并采用分层边界标记和多跨解码算法解决三元组重叠问题。对比结果如表 4.2 所示。

表 4.2 对比实验结果

模型名称	P	R	F_1
NovelTagging	51.28	24.62	33.27
GraphRel	63.37	53.67	58.19
ETL-Span	80.18	66.73	72.84
GCN-CasRel	85.38	83.99	84.68

通过对比可以得知，NovelTagging 的效果最差，主要是因为该模型采用序列标注的方法进行实体关系抽取时只能为每一个文本单词打上一个标签，基于序列标注的方法无法解决三元组重叠问题，而本数据集中包含有大量的重叠数据，因此在数据集上的表现较差。GraphRel 引入图卷积神经网络获取文本结构特征，基于图的方法让关系抽取性能取得了进步，尤其是查准率有了很大的提高。ETL-Span 的思路与本文类似，都是先识别主体，再根据抽取关系和对应的客体，性能远高于其他两个模型，说明指针网络在三元组重叠问题上有更好的表现，但该模型使用的是序列标注并且采用一个序列来抽取关系和客体，而本模型针对每个特定关系都有标注器，在面对关系复杂的文本时抽取性能更强。

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，主要验证编码层和 GCN 特征提取层对模型性能的影响，实验结果如表 4.3 所示。对比模型为：（1）在 CasRel 模型的基础上使用 LSTM 作为编码层；（2）CasRel 基础模型，采用 BERT_base 预训练模型作为编码层；（3）在 CasRel 模型的基础上使用 BERT-wwm-ext 预训练模型作为编码层；（4）在 CasRel 模型的基础上使用 BERT-wwm-ext 预训练模型作为编码层，并且增加 GCN 特征提取层，即本文的模型。

表 4.3 消融实验结果

模型名称	P	R	F_1
$\text{CasRel}_{\text{LSTM}}$	75.45	73.07	74.24
$\text{CasRel}_{\text{BERT}}$	80.66	80.43	80.54
$\text{CasRel}_{\text{BERT-wwm-ext}}$	82.11	81.03	81.57
GCN-CasRel	85.38	83.99	84.68

表中数据表明可以看出，使用 BERT 及其衍生预训练模型的三个模型都高于使用 LSTM 作为编码层的模型，说明 BERT 预训练模型在处理文本任务时对模型性能的提升更加明显。而使用 BERT-wwm-ext 作为编码层的 CasRel 模型在三个指标上也比使用 BERT_base 的 CasRel 基础模型有一定的提升，这说明 BERT-wwm-ext 相比 BERT_base 模型更适用于数据集为中文的任务。本文提出的 GCN-CasRel 模型相比 CasRel 基础模型在三个指标上都得到了明显的提高，并且也高于使用 BERT-wwm-ext 作为编码层的 CasRel 模型，这说明通过图卷积神经网络对依存句法树进行建模捕获句法结构信息，对实体关系抽取性能提升起到了很大的作用。

4.4 本章小结

本章基于 CasRel 模型提出了 GCN-CasRel 实体关系联合抽取模型，CasRel 是一个端到端的级联二进制标记框架，通过将关系建模为主体映射到客体的函数，解决了政务人事信息文本存在的三元组重叠问题。在 CasRel 模型的基础上，本文引入依存句法分析，利用图卷积神经网络模型对依存句法树进行建模，使模型更好的理解文本的语法信息，同时利用注意力机制过滤依存句法树的噪声提升模型的三元组抽取准确性。通过在前文构建的数据集上进行对比实验和消融实验，实验结果证明 GCN-CasRel 模型相比经典的联合抽取模型在查准率、查全率和 F1 值上都得到了明显的提升，构建的特征提取层对模型性能提高有明显作用。

5 政务网站人事信息知识图谱构建

随着互联网的发展和相关条例的公布，各级政府部门、事业单位都在政务网站公开人事信息以及包含人事信息在内的相关新闻，这些文本信息中包含着大量的关联信息，从这些信息中可以挖掘出相关的三元组信息。人事信息三元组数据中实体与关系存在着大量的关联，有人物与非人物实体的关联，也有非人物实体之间的关联，目前政务网站以及相关平台缺少对这些关联实体的直观展示，也没有可供查询管理的工具。本章将针对这一问题基于 Neo4j 图数据库进行知识图谱可视化，用户可以通过知识图谱挖掘出实体之间的关联，更加直接有效地利用人事信息数据，构建的知识图谱也能应用于智能问答、信息搜索等领域。

知识图谱构建流程包括：数据获取、数据聚合、信息抽取、知识可视化。首先编写爬虫代码进行文本获取，接着对文本信息进行预处理，随后将处理好的文本通过实体关系联合抽取模型进行三元组的提取，最后一步将三元组导入到图数据库中进行图谱可视化的操作。本文针对甘肃省部分政务网站公开的人事信息进行知识图谱可视化，通过前文的实体关系联合抽取实验，能够从政务人事信息文本中抽取相关的三元组信息，将非结构数据化的文本信息转换成结构化数据，在将数据导入图数据库中构建知识图谱。

本文第二章介绍了移动应用开发的相关技术，同时第四章的研究表明基于深度学习的人事信息实体关系联合抽取模型能够快速准确的识别出文本中的实体以及实体之间的各种关系，能够针对相关的文本文档抽取人事信息三元组。为了进一步简化知识图谱构建的过程，提升知识图谱的可扩展性，本章基于上文中的实体关系抽取模型，设计并开发一种基于微信平台的人事信息实体关系抽取小程序，微信小程序能够让使用者更加便捷的输入文本内容并抽取三元组信息，三元组信息能够自动进入数据库中，在此基础上可以实现知识图谱的快速构建，对于已经建立好的知识图谱，仍可以通过小程序抽取三元组信息并扩展现有的知识图谱。整体构建过程如图 5.1 所示。

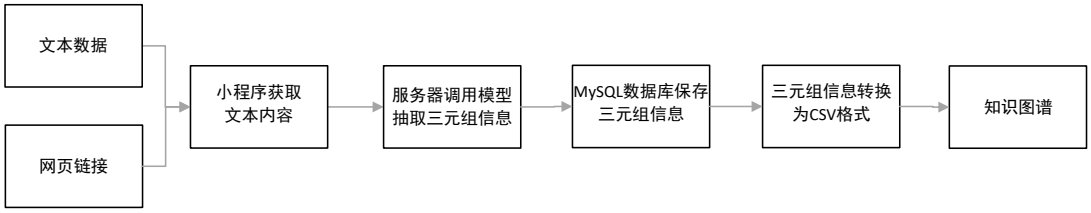

图 5.1 知识图谱构建流程图

5.1 实体关系抽取小程序设计与开发

5.1.1 需求分析

如前文所述，为了提高人事信息实体关系三元组抽取的速度和便捷性，本章基于微信平台开发针对政务相关文本的人事信息实体关系抽取小程序。通过该小程序，可以对包含政务相关人事信息的长文本进行相关实体识别和实体关系的抽取。下面将对小程序的系统功能、开发逻辑、性能需求等相关开发需求进行介绍和阐述。

功能部分，小程序需要实现以下几个功能：（1）文本的获取功能：文本的获取分别两种方式，一种是直接在文本框中输入相应的文本或者是复制粘贴相关文本到文本框中，然后开始实体关系抽取，另外一种是通过网址获取文本，在文本框中输入特定类型的网址，如果网址中包含相关的文本信息，系统会自动识别相关的信息。（2）文本的修改功能：如果输入的文本或者是网址链接出现误差，能够进行局部手动修改或者通过相应的按钮进行文本框的重置使其处于空白状态。（3）实体关系抽取功能：实体关系抽取功能是小程序的核心功能，通过部署在服务器端已经训练好的实体关系抽取模型对小程序获取的文本进行实体关系抽取，输出结果为实体关系三元组。（4）结果展示功能：在主页面应有返回实体关系抽取结果的界面，如果实体关系三元组抽取成功则在此界面中显示结果，如果实体关系三元组抽取失败，则要在界面出现相应的提醒内容。

在小程序的开发逻辑需求上，整个小程序系统应该包含两个前端和后端两个模块，前端负责页面显示以及文本内容的获取，后端负责整体功能的运行，通过预留的接口，让前端可以调用后端功能，实现对文本内容的实体关系三元组的抽

取。

在性能需求方面，人事信息实体关系抽取小程序需要满足及时性、可靠性和准确性方面的需求。（1）及时性：小程序应有较快的反应速度，在用户的网络速度和设备硬件都满足基本使用条件时，小程序从文本的获取到输出结果的返回应在三秒内（2）可靠性：无论输入的文本信息的字数多少以及文本是否与人事信息相关，小程序都应该予以反馈，如果输入的内容中含有人事信息相关的实体关系，小程序应该返回对应的结果，如果输入的内容不包含，小程序也要能够准确识别并做出相应反馈。除此之外，小程序的使用应该总体流畅，较少出现卡顿甚至崩溃的现象。（3）准确性：在文本框获取的文本内容包含人事信息或者网址链接返回的网页内容包含相应的人事信息，小程序返回的实体关系三元组抽取正确率应该处于较高水准。

5.1.2 系统设计

本节内容是设计人事信息实体关系抽取小程序的架构，让用户能够通过小程序的界面进行操作，调用实体关系联合抽取模型，对文本进行信息抽取，服务器端再将结果传送给数据库。为了满足小程序的要求，首先需要将第三章研究得出的实体关系联合抽取模型部署到云服务器上，在小程序界面用户可以选择输入文本或者输入网站完成模型的输入，小程序的内容会经过 Flask 框架输入到服务器端的实体关系联合抽取模型进行实体识别和关系抽取并将结果进行输出，输出的结果在传送给数据库的同时也会返回到小程序的结果展示框。整个小程序系统分为三个部分：前端页面、服务器端和数据库端。

前端页面负责界面访问、目标文本获取、三元组抽取结果展示，此页面包含文本输入对话框、网址输入对话框、结果展示框。服务器端承担小程序的核心功能，包括对前端页面操作请求的相应、文本的预处理、网站规则的识别、对前端页面输入的文本进行实体识别和关系抽取、结果的返回等，在服务器端部署着 Flask 微框架，该框架拓展性强、配置简单，服务器端功能的实现都要经过框架处理。数据库端主要是存储通过前端页面输入的文本内容和经过模型抽取出来的人事信息三元组，存储的数据经过处理可以作为模型训练集的补充，帮助模型提升性能。

5.1.3 系统开发

文本提出的政务人事信息实体关系联合抽取模型的开发环境如下，服务器端使用 Python 语言进行编写，通过接口连接前端页面，前端使用微信开发者工具开发，页面设计上采用微信小程序官方的 WeUI 样式库，实体关系联合抽取模型基于 PyCharm 开发工具设计并训练。具体系统开发环境配置如表 5.1 所示。

表 5.1 系统开发环境配置

选项	工具及参数
操作系统	Windows 64 位
CPU	AMD Ryzen 5 4600H with Radeon Graphics 3.00 GHz
内存	16G
开发语言	Python、WXML、WXSS、JavaScript
开发工具	微信开发者工具 Stable 1.06.2209190、PyCharm2021、MySQL

前端页面的主要功能是获取相关的文本信息，向服务器端发送请求。小程序的所有页面都保存在微信开发者工具平台 pages 文件夹下，pages 文件夹包括 index、link、text 三个子文件夹，分别包含着小程序主页、网址输入页面、文本输入页面三个页面，如图 5.2（a）~（c）所示。除此之外，小程序开发中 app.json 文件决定全局配置，包括页面文件的路径、窗口表现等；app.js 文件负责小程序的全局逻辑；app.wxss 定义小程序的整体样式，在这里小程序引入 weui.wxss，采用此样式库对小程序进行设计。

图 5.2 系统主页面：（a）主页面;（b）网址输入页面；（c）文本输入页面

小程序前端页面的主页由名称、简介和文本获取的两个功能列表组成，点击通过网址抽取信息进入网址抽取页面，如图 5.2（b）所示，在网址获取框中输入相应的网址进行提交，前端页面就会向服务器端发送 POST 请求。在主页中点击通过文字抽取信息进入文字抽取页面，如图 5.2（c）所示。在文本获取框中输入相应的文本或者网站链接进行提交，在发送 POST 请求后，如果成功提取将会返回抽取结果，如图 5.3(a)和 5.3(b)所示，如果不成功将会返回信息抽取失败提醒，如图 5.3(c)所示。

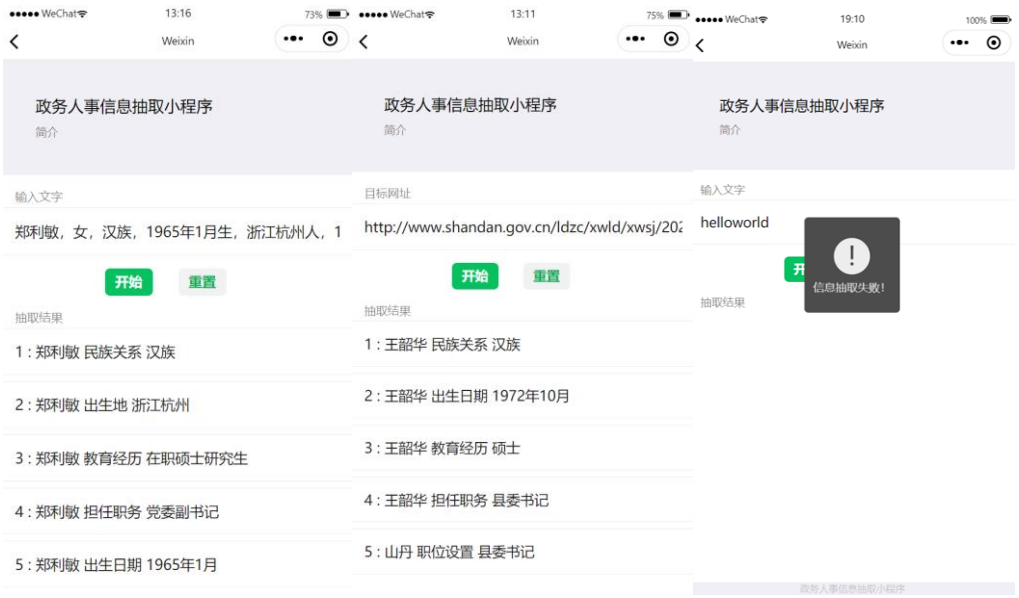

图 5.3 模型抽取反馈；（a）文本抽取成功;（b）链接抽取成功；（c）模型抽取失败

服务器端使用阿里云服务器，主要负责的功能包括实体关系联合抽取模型的部署和运行、响应小程序前端页面发出的请求操作。服务器端在获得文本内容后会保存文本内容，在模型抽取三元组成功后，经过 Flask 框架将抽取出的三元组结果返回到小程序前端页面，并将文本内容和抽取的三元组保存至本地数据库。

数据库负责将信息抽取成功的文本及人事关系三元组存储起来，用于训练集的扩充和实体关系联合抽取模型的迭代。数据库表单设置如表 5.2 所示。

表 5.2 数据库表单设置

字段	类型	备注
Id	int	主键
textcon	string	文本内容
spo	string	三元组

5.2 基于 Neo4j 的政务人事信息知识图谱可视化

鉴于 Neo4j 图数据库的界面友好、拓展性强等特点，本文采用 Neo4j 图数据库作为可视化的工具。根据图数据库的要求，在进行可视化操作前，经过小程序

抽取的实体关系三元组数据已经转换成csv格式，为将数据导入到数据库做准备。py2neo是用于连接Neo4j的一个库，通过引用这个库可以编写Python代码将csv文件导入到数据库中。知识图谱可视化的步骤包括：第一步将三元组信息按照要求转化成csv格式；第二步进入cmd在对应的bin文件夹中执行neo4j.bat console命令，成功启动Neo4j数据库后输入http://localhost://7474/browser进入到数据库浏览器页面；第三步运行Python代码，将文件中的实体节点和关系导入到数据库中完成可视化。本文基于甘肃省政务网站人事信息文本构建政务人事信息知识图谱，部分政务人事信息知识图谱如图5.4所示。

图 5.4 政务人事信息知识图谱部分图

在完成知识图谱可视化以后，可以通过Cypher语言对知识图谱进行相关的查询和修改。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5.3所示。

表 5.3 Cypher 语言常用命令

命令	说明	Cypher 语法
CREATE	创建实体和关系	CREATE (<node-name>:<label-name>)
MATCH	检索实体和关系	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	删除实体和关系	DELETE <node-name-list>
MERGE	CREATE 与 MATCH 命令组合	MERGE (<node-name>:<label-name> {<Property1-name>:<Pro<rty1-Value>})

以 MATCH 命令为例，在 Neo4j 搜索框中输入搜索担任职位为“校长”的指令：MATCH p=()-[r:'担任职务']->(校长) RETURN p LIMIT 25，执行命令后返回所有担任职务为校长的实体，结果如图 5.5 所示。

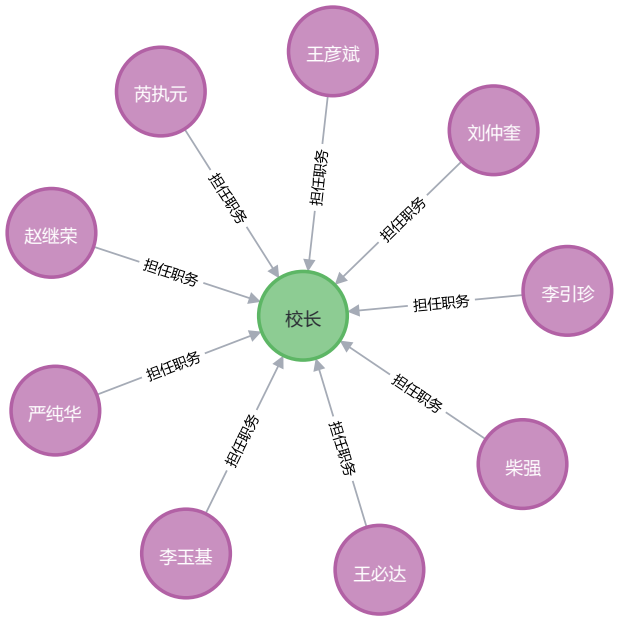

图 5.5 知识图谱查询结果

5.3 本章小结

为了提高知识图谱的构建效率，降低知识图谱的构建难度，本文基于微信平台开发了人事信息实体关系抽取小程序来对三元组信息进行抽取。首先对系统的需求进行分析，确定小程序需要实现的功能，接着对系统进行了设计，并且根据实际需求进行开发，最后展示了开发完成的系统界面和相关操作。在微信小程序上对甘肃省政务网站人事信息文本进行人事信息三元组抽取，在此基础上利用Neo4j 图数据库完成了政务人事信息知识图谱可视化。

6 总结与展望

6.1 工作总结

随着政务信息化发展水平的不断提高，政务网站也产生了越来越多的政务数据。在这些数据中，人事信息数据是非常重要并且具有较高研究价值的一种数据，但是目前的政务人事信息数据经常出现在人事任免、政事新闻等文本信息中，数据处于分散状态且缺少关联，这对于相关的学术研究、产业利用非常不便，因此提供一个直观有效的方案和工具具有较高现实意义。本文基于深度学习模型，针对政务人事信息数据实体关系抽取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）建立政务网站人事信息数据集。在实体关系抽取任务领域，目前还没有大型的公开数据集，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，所以本文选择自行构建数据集。本文采用 Python 爬虫技术，对政务网站上的文本进行信息爬取，确定实体和关系类别后对文本数据进行预处理，构建了政务网站人事信息数据集，为后续实验提供了基础。

（2）提出 GCN-CasRel 实体关系联合抽取模型。针对三元组重叠的问题，CasRel 模型提出一种端到端的级联二进制标记框架，通过将关系建模为主体映射到客体的函数解决了传统流水线法无法避免的三元组重叠问题。在此基础上，本文引入图卷积神经网络，通过对依存句法关系进行建模获得文本句法特征，同时利用注意力机制过滤依存句法树的噪声进一步提高了模型实体关系抽取的性能。

（3）构建政务网站人事信息知识图谱。目前关于实体关系抽取尤其是涉及政务人事信息的移动应用处于空白，鉴于微信的广泛性以及小程序开发的便捷性、兼容性等特点，本文基于微信开发者工具并结合深度学习模型部署，开发了一款政务人事信息实体关系抽取小程序，满足使用者对于快速抽取三元组信息需求，并在此基础上采用 Neo4j 图数据库对小程序抽取的人事信息三元组进行可视化，完成政务网站人事信息知识图谱的创建。

6.2 未来展望

本文针对政务网站人事信息数据进行实体关系三元组抽取，并在此基础上构建了政务人事信息实体关系抽取小程序以及政务人事信息知识图谱，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）扩充政务人事信息数据集。目前的数据集的文本数据量偏少，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，鉴于人事信息之间的复杂关联以及政府机关庞大的职位机构设置，后期可以进一步增加实体类别和关系类别以更好的适应实际需求。

（2）改善抽取模型。目前的模型虽然能够在面对实体三元组重叠问题时有较好的效果，但是面对实体嵌套问题性能不佳，后期会将重点放在解决实体嵌套问题上。除此之外，可以引入更加丰富的外部知识库，进一步提升关于政府机构等相关实体的识别准确率。

（3）完善微信小程序应用。小程序迭代更新包括 UI 界面的设计和优化、后台服务器性能提升、代码优化等方面。除此之外，小程序在功能方面也可以突破，比如与 Neo4j 图数据库互联直接构建知识图谱等。

参考文献

[1] 徐增林,盛泳潘,贺丽荣,王雅芳.知识图谱技术综述[J].电子科技大学学报,2016,45(4):589-606.

[2] 刘绍毓, 李弼程, 郭志刚,等. 实体关系抽取研究综述[J]. 信息工程大学学报,2016,17(5):7.

[3] 李冬梅,张扬,李东远,林丹琼.实体关系抽取方法研究综述[J].计算机研究与发展,2020,57(7):1424-1448.

[4] 张少伟,王鑫,陈子睿,王林,徐大为,贾勇哲.有监督实体关系联合抽取方法研究综述[J].计算机科学与探索,2022,16(4):713-733.

[5] Chinchor N, Marsh E. Muc-7 information extraction task definition[C]// Proceedings of the 7th message understanding conference (MUC-7), Appendices.1998:359-367.

[6] McDonald R, Pereira F, Kulick S, et al. Simple algorithms for complex relation extraction with applications to biomedical IE[C]//Proceedings of the 43rd annual meeting of the association for computational linguistics. Stroudsburg:ACL,2005:491-498.

[7] Aone C, Halverson L, Hampton T, et al. SRA: Description of the IE2 system used for MUC-7[C]//Proceedings of the 7th Message Understanding Conference (MUC-7). Stroudsburg:ACL,1998[2020-05-29]. https://www.researchgate.net/publication/2243565_Sra_Description_Of_The_Ie2_System_Used_for_MUC-7.

[8] Fukumoto J,Masui F, Shimohata M, et al. Oki electric industry: Description of the Oki system as used for MUC-7[C]//Proceedings of the 7th Message Understanding Conference (MUC-7). Stroudsburg:ACL,1998[2020-05-29]. https://core.ac.uk/display/21411891.

[9] 邓擘,樊孝忠,杨立公.用语义模式提取实体关系的方法[J].计算机工程,2007(10):212-214.

[10] 徐健, 张智雄. 典型关系抽取系统的技术方法解析[J]. 数字图书馆论坛, 2008(9):6.

[11] Aone C, Ramos-Santacruz M. REES: a large-scale relation and event extraction system[C]//Proceedings of the 6th Applied Natural Language. Stroudsburg:ACL,2000:76-83.

[12] Collins M, Duffy N. Convolution kernels for natural language[J]. Advances in neural information processing systems, 2001:625-632.

[13] Culotta A, Sorensen J. Dependency tree kernels for relation extraction [C]// Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. 2004:423-429.

[14] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. Stroudsburg:ACL,2004: 22-26.

[15] Xia S, Lehong D. Feature-based approach to Chinese term relation extraction [C]//2009 International Conference on Signal Processing Systems. IEEE, 2009:410-414.

[16] Jiang J, Zhai C X. A systematic exploration of the feature space for relation extraction[C]//Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. 2007:113-120.

[17] 车万翔,刘挺,李生.实体关系自动抽取[J].中文信息学报,2005(2):1-6.

[18] 甘丽新,万常选,刘德喜,钟青,江腾蛟.基于句法语义特征的中文实体关系抽取[J].计算机研究与发展,2016,53(2):284-302.

[19] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction[J]. Journal of machine learning research, 2003, 3(2): 1083-1106.

[20] Zhang X, Gao Z, Zhu M. Kernel methods and its application in relation extraction[C]//Proceedings of International Conference on Computer Science and Service System (CSSS). IEEE, 2011:1362-1365.

[21] 刘克彬,李芳,刘磊,韩颖.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(8):1406-1411.

[22]虞欢欢,钱龙华,周国栋,朱巧明.基于合一句法和实体语义树的中文语义关系抽取[J].中文信息学报,2010,24(5):17-23.

[23]Chapelle O, Scholkopf B, Zien A. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews][J]. IEEE Transactions on Neural Networks, 2009, 20(3):524-542.

[24]Brin S. Extracting patterns and relations from the world wide web[C]// Proceedings of the International Workshop on The World Wide Web and Databases. Berlin: Springer, 1998:172-183.

[25]Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections[C]//Proceedings of the 5th ACM Conference on Digital Libraries. 2000:85-94.

[26]Balcan F, Blum A, Yang K. Co-training and expansion: Towards bridging theory and practice[C]//Proceedings of the Advances in Neural Information Processing Systems. Cambridge, MA: Massachusetts Institute of Technology,2005:89-96.

[27]Zhu X, Ghahramani Z, Lafferty J D. Semi-supervised learning using gaussian fields and harmonic functions[C]//Proceedings of the 20th International conference on Machine learning. ACM, 2003:912-919.

[28]Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora[C]//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL,2004:415-422.

[29]Rozenfeld B, Feldman R. High-performance unsupervised relation extraction from large corpora[C]//Proceedings of the 6th International Conference on Data Mining (ICDM'06). IEEE, 2006: 1032-1037.

[30]Shinyama Y, Sekine S. Preemptive information extraction using unrestricted relation discovery[C]//Proceedings of Human Language Technology Conference of the NAACL. Stroudsburg: ACL,2006:304-311.

[31]秦兵,刘安安,刘挺.无指导的中文开放式实体关系抽取[J].计算机研究与发展,2015,52(5):1029-1035.

[32]Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504-507.

[33] 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.

[34] Xu K, Feng Y, Huang S, et al. Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling[J]. Computer Science, 2015,71(7):941-949.

[35] 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.

[36] Xu Y, Mou L, Li G, et al. Classifying relations via long short term memory networks along shortest dependency paths[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015:1785-1794.

[37] Zhang S, Zheng D, Hu X, et al. Bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. 2015:73-78.

[38] 鄂海红,张文静,肖思琪,程瑞,胡莺夕,周筱松,牛佩晴.深度学习实体关系抽取研究综述[J].软件学报,2019,30(6):1793-1818.

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

[40] Zheng S, Hao Y, Lu D, et al. Joint entity and relation extraction based on a hybrid neural network[J]. Neurocomputing, 2017, 257(12):59-66.

[41] Wei Z, Su J, Wang Y, et al. A novel cascade binary tagging framework for relational triple extraction[J]. arXiv preprint arXiv:1909.03227, 2019.

[42] Wang S, Zhang Y, Che W, et al. Joint extraction of entities and relations based on a novel graph scheme[C]//IJCAI. 2018:4461-4467.

[43] Zeng X, Zeng D, He S, et al. Extracting relational facts by an end-to-end neural model with copy mechanism[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018:506-514.

[44]刘辉,江千军,桂前进,张祺,王梓豫,王磊,王京景.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.

[45]Etzioni O, Banko M, Soderland S, et al. Open information extraction from the web[J]. Communications of the ACM, 2008, 51(12):68-74.

[46]Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011:1535-1545.

[47]Qiu L, Zhang Y. ZORE: A syntax-based system for chinese open relation extraction[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014:1870-1880.

[48]郭喜跃. 面向开放领域文本的实体关系抽取[D].华中师范大学,2016.

[49]姚贤明,甘健侯,徐坚.面向中文开放领域的多元实体关系抽取研究[J].智能系统学报,2019,14(3):597-604.

[50]AMIT S. Introducing the knowledge graph[R]. America: Official Blog of Google, 2012.

[51]李文鹏,王建彬,林泽琦,赵俊峰,邹艳珍,谢冰.面向开源软件项目的软件知识图谱构建方法[J].计算机科学与探索,2017,11(6):851-862.

[52]陈亚东,鲜国建,寇远涛,郭淑敏,刘现武.我国苹果产业知识图谱构建研究[J].中国农业资源与区划,2017,38(11):40-45.

[53]由丽萍,郎宇翔.基于商品评论语义分析的情感知识图谱构建与查询应用[J].情报理论与实践,2018,41(8):132-136,131.

[54]奥德玛,杨云飞,穗志方,代达劭,常宝宝,李素建,昝红英.中文医学知识图谱CMeKG 构建初探[J].中文信息学报,2019,33(10):1-9.

[55]陈璟浩,曾桢,李纲.基于知识图谱的“一带一路”投资问答系统构建[J].图书情报工作,2020,64(12):95-105.

[56]华斌,吴诺,李若瑄.基于知识图谱的电子政务项目评价方法研究与实践[J].情报理论与实践,2021,44(2):147-153,146.

[57]高晨翔,黄新荣.区域政务微博知识图谱构建及可视化研究[J].现代情报,2020,40(12):90-99,113.

[58]朱宗尧.政务图谱：框架逻辑及其理论阐释——基于上海“一网通办”的实践[J].电子政务,2021(4):40-50.

[59]黄贵辉,许正中.国内行政改革研究热点与发展趋势研究——基于CiteSpace知识图谱分析[J].长白学刊,2021(05):65-74.

[60]于娟,黄恒琪,席运江,朱正祥.基于图数据库的人物关系知识图谱推理方法研究[J].情报科学,2019,37(10):8-12.

[61]黄娟,陈崇成,叶晓燕,马腾.“民国清流”名人文化主题数据的组织和可视化方法[J].地球信息科学学报,2020,22(5):954-966.

[62]孙洪伟,司唯山,纪兆辉.基于本体的家谱知识图谱构建及信息检索系统的设计实现[J].计算机产品与流通,2020(9):156.

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

[64]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.

[65]潘晓英,陈柳,余慧敏,赵逸喆,肖康泞.主题爬虫技术研究综述[J].计算机应用研究,2020,37(4):961-965,972.

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

[67]Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[68]Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.

[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] Kim Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL,2014:1746-1751.

[71] Sun K, Zhang R, Mao Y, et al. Relation extraction with convolutional network over learnable syntax-transport graph[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(5):8928-8935.

[72] Tian Y, Chen G, Song Y, et al. Dependency-driven relation extraction with attentive graph convolutional networks[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021:4458-4471.

[73] Zheng S, Wang F, Bao H, et al. Joint extraction of entities and relations based on a novel tagging scheme[J]. arXiv preprint arXiv:1706.05075, 2017.

[74] Fu T J, Li P H, Ma W Y. Graphrel: Modeling text as relational graphs for joint entity and relation extraction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:1409-1418.

[75] Yu B, Zhang Z, Shu X, et al. Joint extraction of entities and relations based on a novel decomposition strategy[J]. arXiv preprint arXiv:1909.04273, 2019.

致 谢

三年研究生生活转瞬即逝，出入校门的场景还犹在昨日，转眼已经到了离别的时刻。三年的时光短暂而又充实，三年来的所有收获和成长都离不开每一个帮助我的人，在这里我真心对他们表示感谢！

感谢我的导师杨海军教授，感谢老师的谆谆教诲和耐心指导。入学时老师便为我指定了学习方向，每一次组会，杨老师都会认真听取我的汇报，为下一步的研究指明方向。无论是小论文还是毕业论文，老师都给我了很多的帮助和意见，让我得以顺利的毕业。杨老师不仅在学业方面给予了我很大的帮助，在生活中也给了我很多的鼓励，无论面临什么样的困难，杨老师温和的话语、豁达的态度都让我能够沉下心来迎接挑战，祝愿老师身体健康，心想事成！

感谢实验室的小伙伴们和我的舍友，三年时光里我们一起探讨学业、传递经验，正是你们高涨的学习热情和孜孜不倦的态度让我有了充足的前进动力，祝你们前程似锦！

感谢答辩组的所有专家和老师，在百忙之中抽出时间指导我的研究工作，为我的论文提出宝贵的意见，让我更加进步！

感谢父母，没有你们的默默支撑就没有我的今天。感谢我的父亲和母亲，谢谢你们给予了我你们全部的爱，为我提供了温暖的港湾，祝你们永远快乐、永远幸福！

感谢陪伴了我七年的女友，感谢你在快乐的时候与我分享喜悦，在困难的时候陪我共度难关、不离不弃，祝愿我们早日修成正果！

最后谢谢所有关心和帮助过我的人，谢谢你们！

攻读硕士学位期间发表的论文及科研情况

发表论文：

[1]秦伟德,杨海军,张晓蝶. 基于无监督聚类算法的多准则 ABC 分析应用[J]. 物流时代周刊,2022(12):27-29