密级：_____

河北地质大学硕士学位论文

基于深度学习的实体关系抽取研究

论 文 作 者：	刘鹏鹏	学 生 类 别：	全日制
一 级 学 科：	计算机科学与技术	学 科 专 业：	计算机应用技术
指 导 教 师：	亢俊健	职　　称：	教授

Dissertation Submitted to
Hebei GEO University
for
The Master Degree of
Computer Application Technology

Research on Entity Relationship Extraction Based on Deep Learning

by
Liu PengPeng
Supervisor: Prof. Kang JunJian

October, 2021

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文不包含任何他人或集体已经发表的作品内容，也不包含本人为获得其他学位而使用过的材料。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：刘鹏鹏　　日期：2021.12.12.

关于学位论文版权使用授权的说明

本人完全了解河北地质大学关于收集、保存、使用学位论文的以下规定：学校有权采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供本学位论文全文或者部分内容的阅览服务；学校有权将学位论文的全部或部分内容编入有关数据库进行检索、交流；学校有权向国家有关部门或者机构送交论文的复印件和电子版。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：刘鹏鹏　　日期：2021.12.12

导　师　签　名：元俊雄　　日期：2021.12.12

摘　　要

作为自然语言处理领域的一项重要基础工作，实体关系抽取是构建知识图谱的关键一环，其目的在于发现文本中的实体并确定实体间存在的语义关系，该研究有助于知识库构建，进而为智能搜索、智能推荐、智能问答等提供有力地服务跟支持。

早期的方法都非常依赖研究人员根据语料的特点去人为设计特定的模板，或人为设计相关特征，该过程需要付出很高的时间成本，且难以在各个领域上进行推广应用。由于近年来深度学习技术的不断突破和发展，研究者们开始研究用深度学习的方式来解决实体关系抽取任务并在该任务上不断取得新的突破。然而，在这些方法中依旧存在依赖领域知识库特征、对于数据的特征学习能力不足、对特征的优化跟组合缺乏更进一步的研究等问题。本文围绕以上三点进行研究，使得本文提出的新的实体关系抽取模型可以自动挖掘出数据本身所蕴含的丰富特征，不必依赖外部的各种知识库资源。本论文的研究内容和创新点如下：

（1）提出融合词特征与相邻词间特征的关系抽取模型

本文提出一种融合词特征与相邻词间特征的关系抽取模型，该模型可充分利用卷积神经网络以及双向长短期记忆网络的特点，并结合注意力机制来提取基于词的句子特征，基于相邻词间关系的句子特征，进一步挖掘出自然语言文本中潜在蕴含的语义信息。实验结果显示，本论文所设计的实验在公开的 SemEval-2010Task8 以及 Wiki80 数据集上进行训练和测试，测试结果对比于其它 3 个经典及主流模型均有所提高。

（2）提出结合实体相关信息的多特征组合的关系抽取模型

实体以及实体左右的上下文背景信息有利于确定实体在句中的语义关系，这对关系抽取任务起到非常重要的作用。因此，在不依赖外部知识的前提下，为了使模型能够更好地利用到实体信息，学习到句中与实体相关的重要上下文向量信息，本文构建了一个能学习句子连续特征，又能捕获到实体及实体在句中复杂背景关系的神经网络模型。最后，本文对模型学到的各类特征进行优化组合。实验结果表明，本文提出的融合结合实体相关信息的多特征组合的关系抽取模型有效提升了实体关系抽取效果并优于对比模型。

关键字：知识图谱；关系抽取；特征融合；深度学习；注意力机制

ABSTRACT

As an important basic work in the field of natural language processing, entity relationship extraction is a key part of the construction of knowledge graphs. Its purpose is to discover entities in texts and determine the semantic relationships between entities. This research is helpful to the construction of knowledge bases. And then provide powerful services and support for smart search, smart recommendation, smart question and answer, etc.

Early methods rely heavily on researchers to artificially design specific templates or artificially design related features according to the characteristics of the corpus. This process requires a high cost of time and is difficult to promote and apply in various fields. Due to the continuous breakthroughs and development of deep learning technology in recent years, researchers have begun to study the use of deep learning to solve the task of entity relationship extraction and continue to make new breakthroughs in this task. However, these methods still have problems such as dependence on domain knowledge base features, insufficient feature learning capabilities for data, and lack of further research on feature optimization and combination. This article focuses on the above three points, so that the new entity relationship extraction model proposed in this article can automatically mine the rich features contained in the data itself, without relying on various external knowledge base resources. The research content and innovations of this paper are as follows:

(1) Propose a relationship extraction model for fusion of word features and features between adjacent words

This paper proposes a relationship extraction model that combines word features and features between adjacent words. This model can make full use of the characteristics of convolutional neural networks and bidirectional long-term memory networks, and combine the attention mechanism to extract word-based sentence features. The sentence characteristics of the relationship between adjacent words can further dig out the semantic information latent in the natural language text. The experimental results show that the experiments designed in this paper are trained and tested on the public SemEval-2010 Task8 and Wiki80 data sets, and the test results are improved compared to the other three classic and mainstream models.

(2) Propose a relationship extraction model that combines multiple feature combinations with entity-related information

The entity and the contextual information around the entity help determine the semantic relationship of the entity in the sentence, which plays a very important role in the task of relationship extraction. Therefore, without relying on external knowledge, in order to enable the model to make better use of entity information and learn important context vector information related to the entity in the sentence, this paper constructs a sentence that can learn the continuous features of the sentence and capture A neural network model of the complex background relationship between entities and entities in sentences. Finally, this article optimizes the combination of various features learned by the model. The experimental results show that the multi-feature combination relationship extraction model proposed in this paper, which combines entity-related information, effectively improves the entity relationship extraction effect and is better than the comparison model.

KEYWORDS: Knowledge Graph; Relation Extraction; Feature Fusion; Deep Learning; Attention Mechanism

目　　录

第一章　绪论

1.1 研究背景及意义

在国家“十四五和 2035 年远景目标纲要”中，着重提及了发展新一代人工智能(Artificial Intelligence，AI)，并将其纳入国家发展战略。AI 的发展程度，也侧面反映出国家的综合国力。随着相关理论、算法的不断突破以及技术的不断成熟，我们逐渐迈入“万物智能”的新时代。

知识图谱在人工智能的研究中占据重要位置，其研究对于解决人工智能可解释性问题具有重要意义，知识图谱尤其在智能感知、语义理解方面有重要的研究价值，知识图谱与深度学习一起推动着人工智能的发展。近几年来知识表示学习与深度学习等技术的迅速发展，推动了知识图谱关键技术的不断发展和突破。在工业界，知识图谱被广泛关注并在搜索、推荐、问答等应用领域上并取得了显著的成果。谷歌、微软、苹果、百度、阿里、搜狗、华为、美团等众多国内外互联网企业先后构建了大规模通用或领域知识图谱，比如谷歌、百度等构建的知识图谱应用在搜索引擎上可为用户提供基于实体与关系查询的智能化语义搜索，这样可以更好地理解并满足用户的查询需要。除此之外，基于知识图谱技术理论构建的知识库还可为智能问答系统、智能决策系统、智能推荐系统提供知识服务。知识图谱的意义不仅在于其具有巨大的商业应用价值，而且还具有重要的理论价值[1]。目前，知识图谱发展的瓶颈，同时也是制约知识图谱快速应用的关键的热点与技术难点之一在于大规模知识库的构建，也就是从大规模自由文本中抽取出实体关系进行知识图谱的自动构建。本文研究的是基于深度学习的实体关系抽取方法，旨在提高模型的关系提取性能，从而为知识图谱的自动构建提供技术支持。

关系抽取的目标是从文本中抽取出实体之间存在的语义关系[2]。关系抽取一般是先识别出文本中的实体，然后进一步在实体所在文本中抽取出实体之间可能存在的关联关系[3]。简而言之，关系抽取的关键是准确找到文本中已标记实体间的语义关系。在知识图谱领域中，实体一般指的是具体的事物或者抽象的概念，如一个人、一本书，也比如数学、机器学习等概念实体，在实际应用中是根据具体任务不同而定义的。一般而言，实体通常如地点、时间、地名、人名、机构等等。如在句子“知识图谱是门多学科交叉的重要领域，属于人工智能的一个重要分支”中，“知识图谱”，“人工智能”都是抽象的概念。关系抽取是基于以上给定实体的条件下，在文本中识别出实体间存在的关系。如上述句中的实体“人工智能”与实体“知识图谱”存在“包含”

关系。构建关系抽取模型的目的是能自动识别出“人工智能”与“知识图谱”之间的“包含”关系。

关系抽取模型的目标是能自动且精准地抽取出实体间的语义关系。关系抽取的最终目的是将抽取到的实体关系三元组转化成有组织、有结构的知识库数据，进而为下游任务提供知识服务[4]。此外，从理论价值上讲，关系抽取相关技术同时也可为自然语言处理（Natural Language Processing, NLP）相关任务提供理论支持，因此具有重要的研究意义。

1.2 国内外研究现状

关系抽取这一概念早在二十世纪九十年代举办的第七届信息理解会议（Message Understanding Conference，MUC）中被提出，会议中提出将关系抽取纳入到信息抽取任务中[5]，此后在各大国内外 NLP 相关会议中不断发展完善，如自动内容抽取(Automatic Content Extraction, ACE)会议、文本分析会议(Text Analysis Conference Knowledge Base Population, TAC-KBP)以及全国知识图谱与语义计算大会（CCKS: China Conference on Knowledge Graph and Semantic Computing）等，各类评测会议会定义不同的测评任务同时发布相应的数据集。直至今日，有关关系抽取的研究仍被国内外科研人员青睐。目前，国内外研究人员就实体关系抽取任务的研究积累了各类方法，如果从其对标注数据的要求上进行分类，可将其划分为：无监督的抽取方法、半监督的抽取方法、远程监督的抽取方法以及有监督的抽取方法[6]。有监督的实体关系抽取方法需要人工去标注大量的数据用以模型的有监督学习，其在本质上属于关系分类的问题；半监督的方法相较于前者仅需少量人工标注的数据即可，将此作为初始的种子集，通过迭代来学习对未标注数据的识别；无监督的实体关系抽取方法相比而言更适合于开放领域，因为其不需预先标注任何数据，就能挖掘出实体对潜在的关系；远程监督的实体关系抽取方法是利用已有的知识库中蕴含的大量而丰富的关系事实，通过其与大量非结构化的文本对齐来自发的生成标注的数据[7]。本节后续内容是对上述四类关系抽取方法的展开介绍。

1.2.1 无监督的方法

基于无监督的方法是指语料在无人工标注的条件下进行，这时可以将问题看作是聚类问题[8]。其核心思想如下：第一步要标记出预处理后语料文本中的实体；然后统计文本句子中共现的实体对及其所在句中的背景信息，在此基础之上分别计算出实体到其对应上下文背景信息中的相似度，然后基于相似度来进行聚类操作；最终，在聚

类结果中，每一类都表示一种特定的关系[9]。无监督学习比较适用于开放领域，能够挖掘出实体对之间潜在的未知关系。

历史上 Hasegawa 等[10]首次提出并使用完全基于无监督的抽取方法，这为基于无监督的方法研究提供了基石，这类方法只需实体识别操作来完成语料中命名实体的标记，然后模型就会针对这些标注了实体的语料进行无监督学习，通过设置阈值来进行聚类，发掘出潜在的关系。此方法存在两个问题：一是难以预定阈值，二是依照频率确定关系特征词时没有考虑到噪声。Chen 等[11]在该研究的基础之上提出一种特征选择的方法来找出表示不同关系的有代表性的特征判别词，并提出一种聚类效果评估公式，该公式可识别出核心的特征子集与上下文的簇数目，此外可以避免因常用词的噪声而影响关系表征。Yan Y 等[12]研究了 Web 的结构特点，基于维基百科中的文本信息，用聚类并以模式组合的方式处理百科文本实体中的语义关系，提出了一种新的无监督的方式。Poon 等[13]提出用递归的方式对依赖树片段执行聚类操作，通过这种方式可以将表达含义相同的各种信息都分在一起。Yao 等[14]使用概率生成模型对所有关系实例进行聚类。针对中文领域的实体关系抽取，黄晨等[15]提出了卷积树核的方法，该方法用最短包含树对关系类进行表示，同时用核函数的计算方法计算不同结构化信息之间的相似度，最后使用三种分层聚类算法来进行关系抽取。马超[16]在传统的无监督学习方法基础之上，提出加入样例概念权重，并提出一种改进的聚类算法迭代更新所有样例概念对的权重，使得可以从单样例多概念对中正确的抽取概念关系。

基于无监督的方法最明显的优势是不依赖标注的语料，解放了人力也解决了机器学习领域中常存在的数据不足问题，并且在不同领域适性强，具有良好的可移植性。但是该方法的缺陷也是很显然的，需要预先定义好聚类的阈值，阈值往往很难确定，而且特征的选取也过度依赖个人经验，聚类结果中关系边界不清晰。与此同时，现有的基于无监督的方法有待进一步研究跟完善，且抽取效果相比于有监督抽取方法有较大的差距。

1.2.2 半监督的方法

半监督的方式可以利用相对较少的标注数据进行关系抽取任务。一般而言，半监督的方法要去先设计好关系类型，然后人工加入一些实体对将其作为训练语料的种子，进一步通过不断迭代的方式最终产生关系数据集。该过程需要一定的人工干预调整，但相比于有监督方法，半监督方法大大减轻了对标注数据的依赖性。

在基于半监督的实体关系抽取方法中，Brin[17]提出 Bootstrapping 的方法，将少量的关于书与作者名的实体对关系作为种子，从语料中抽到实例，并将其视为标注样本，以此种方式来建立关系抽取模板；然后基于该模板发现新的关系，并通过不断调整、迭代将结果添加入建立的模板中。Agichtein 等[18]对新抽的实例进行置信度评估，该方

法用向量去表达实体及其关系，通过迭代计算向量间的相似度来扩充样本，在迭代中模型会自动评估这些实体关系元组的置信度，选取高置信度样例作为下一次迭代的种子。陈锦秀等[19]从另一个角度采用基于图的策略，构建了基于网络图的关系抽取模型，这为该研究提供了新思路。Bootstrapping 方法高度依赖初始种子选取的质量，在迭代中噪声会累积并不断放大，从而最终会发生语义漂移现象。为了解决这个存在的问题，研究人员将目光投入到协同训练方法中。一般而言，在协同训练方法中，两个特征集表示一个实例，这样可以补充实例的信息，采用相互预测的方式将分类器各自对数据的最佳预测添加到另外一个分类器的种子集中。Cvitas[20]设计了基于协同训练的方法，该方法在一定程度上有效缓解了 Bootstrapping 算法中的语义漂移问题，但在预测的可信度方面仍存在不足。

半监督实体关系抽取大大降低了关系抽取的学习过程中对于人工干预的依赖，只需人工构造最初的种子集，利用少量的训练语料即可构建模型，弥补了有监督方法的不足。但是，其对初始种子的标注质量要求较高，对建立复杂的模板要求较高，在迭代过程中语义漂移现象依然会存在，在如何设置初始种子，以及在迭代过程中关系模式如何去评估等方面还有很广阔的研究空间。

1.2.3 远程监督的方法

远程监督方法的提出基于两个目的，一是为了降低对标注语料的过分依赖，二是为了提升抽取模型的迁移能力。远程监督基于这样的基本假设：假如某对实体在某知识库中存在着某种特定关系，那么包括这两个实体的全部句子都表达这种关系。远程监督的基本流程是先将待标注的语料与远程知识库中的信息自动对齐，以此种方式对样本进行自动标注从而得到标记样本。

Mintz 等[21]首次提出远程监督方法，他们将新闻领域文本与 FreeBase 知识库中实体关系进行对齐，并利用远程监督的方式自动对齐文本匹配关系，从而训练关系分类模型。Zeng 等[22]在传统的卷积神经网络中融合了多示例学习，这在一定程度上缓解了远程监督中存在的噪声问题，但是结合多示例学习方法的同时也会造成很多有用信息的丢失，此外，对于关系重叠问题也无法处理。Lin 等[23]用注意力机制来主动学习并筛选样本中的有用信息，并且有效地避免了多示例学习过程中造成的信息丢失问题。结果表示该模型可以学到的注意力权重信息，以此缓解远程噪声问题，同时还找出了有用的特征。Feng 等[24]提出一种新的思路，将样本对齐转化为强化学习问题，该模型分为样本选择器以及关系分类器，关系分类器可以很好地过滤掉远程监督中的一些噪声数据。Qin 等[25]利用强化学习方法训练模型正负样本感知器。但以上两种采用流水线方式的模型都无法处理关系重叠的问题，流水线方式也无法将实体抽取与关系抽取两个子任务间的联系结合起来。针对流水线方式存在的问题，Ren 等[26]提出基于

远程监督的联合抽取模型进行实体关系抽取任务，实验证明该方法能有效地减轻噪声的传播，同时还具备较好的可迁移性。黄杨琛等[27]针对远程监督过程中存在的噪声问题提出了多示例学习的方法，利用 TF-IDF 公式计算指导词，并将词法与句法特征结合起来作为关系特征向量用于模型的学习。

远程监督的抽取方法虽然能快速得到大量的标记样本，大大减少人工标记带来的庞大的工作量，但这种假设过于强烈，本身就存在不足，因此不可避免地使得到的数据集中包含大量的错误标签，从而引起错误传播进而影响关系抽取的结果。

1.2.4 有监督的方法

有监督方法可以被定义为分类问题，即通过句子中两个实体间的相关特征来预测它们之间的关系是否属于预先定义好的某一种类别，主要包含传统的特征向量的方法、核函数的方法以及本文重点研究的深度学习的方法。

基于特征向量的主流的做法是从语料文本中提取词法、句法、语法等关键信息，然后根据关键信息构造出特征向量，最后通过计算特征向量之间相似度的方式来训练出关系分类模型，该方法往往基于相似度高的三元组属于相同的语义关系的原则。该方法根据任务可以分为三个主要流程，分别是特征选择、向量表示、分类器构建。具体而言，就是首先得到训练语料的语法、词法等特征，然后进一步将特征向量化，最后用分类器训练机器学习模型，并将其应用在文本中实现关系抽取任务。Kambhatla[28]提出最大熵模型结合词法、句法和语义特征的模型实现关系抽取，该模型降低了对特征提取树的依赖。大多数基于特征向量的方法都是依赖于传统机器学习的算法来实现，并在关系抽取任务上有着优异的表现。Giuliano 等[29]提出了一种利用上下文和距离等特征结合浅层和深层信息，使用核方法提取名词之间语义关系，并用支持向量机作为分类模型，在实验中取得良好的效果。Tratz 等[30]提出的最大熵分类模型，通过训练得到最大熵分类器，将其用于关系抽取。Culotta 等[31]提出了基于条件随机场同时加入了正则化参数的关系分类方法，实验表现较好。以上所有采用特征向量的方法都能比较好地实现关系抽取的任务，但在特征构建过程中过于依靠设计人员的主观经验，主观性较强且对于特征需要大量地验证证明，并且对文本句子本身的信息利用不够，于是为了更好地挖掘出并利用到文本中的结构化特征，后来提出了核函数的方法。

核函数方法一般是把文本结构树作为研究对象，用结构树间的相似程度来完成关系抽取任务。Zelenko 等[32]提出并使用核函数的方法进行关系抽取任务，使用了浅层解析树核并设计了计算核函数的方式，然后将设计出来的核与支持向量机和投票感知机结合使用完成实体关系抽取任务。Zhao 等[33]将多个核融合在一起，采取多核组合的方式处理语法特征。Culotta 等[34]面向领域数据集，通过依存树核对语料库中实体间的关系进行了分类，并分别研究了词性特征、实体类型特征等不同特性的影响效果。

Bunescu 等[35]基于依赖图中实体间的最短路径提取核，图中两个实体间的最短路径可得到实体关系，并利用最短依存树核进行改进，实验表明该方法对关系分类任务有明显改善。Zhang 等[36]提出将卷积核应用到解析树上，并在英文公开数据集上有不错的表现。庄成龙等[37]提出了一种基于树核的改进方法，该模型是在原结构化的关系实例中融合实体的语义信息，同时去除无关信息进而在整体上提高关系抽取的效果。总体而言，基于核函数的思想往往以树作为基本研究对象，通过计算子树之间的相似性来完成关系抽取任务。基于核函数的方法，其核心也是难点在于如何选择恰当的核函数，设计出的核函数将会直接影响到最终效果。此外，复合的核函数构造也导致了训练速度过于缓慢，因此该方法并不适合大规模数据。

以上所述方法往往都依靠 NLP 相关的工具，而工具本身就会带来错误噪声，在迭代过程中会不断放大，严重影响关系抽取的效果，降低算法的性能。神经网络模型仅通过输入数据便可主动学习到数据本身蕴含的隐藏特征，同时无须人工进行特征选择，且在关系抽取的任务中取得了相当不错的性能。

Socher 等[38]采用递归神经网络，在解析树中，对每个节点分配一个向量和矩阵，分别捕捉组成部分的内在含义和邻近的单词或短语的含义，通过 RNN 学习多种句法类型信息和句子的向量表示，实验显示该方法在公开数据集上表现出了不错的效果。Zeng 等[39]首次将 CNN 用到该任务中，基于卷积神经网络提取句子特征，并结合词汇特征等进行关系抽取，不需要复杂的 NLP 工具进行处理，达到了当时最好的效果。Nguyen 等[40]在此研究的基础上在卷积层中加入了多尺寸大小的卷积核来提取更丰富的 N-Gram 特征，该实验证实了多尺寸卷积核对关系抽取任务有提升。Lin 等[41]基于传统的 CNN 提出了 PCNN，利用两个实体将池化特征分三段分别进行池化，这种分段池化操作使得模型可以更好地利用实体间的上下文信息，同时通过引入注意力机制来处理错误标签问题。Santos 等[42]在 Zeng[39]的基础上提出一种新的模型，该模型的主要创新点在于其计算损失函数 Ranking loss 上，相较于传统的 Softmax 函数，Ranking loss 函数可以使模型分别对正负类别进行采样。Zhou 等[43]用双向长短期记忆网络对句子建模，并使用基于词的注意力机制学习句子的重要信息来提升结果。闫雄[44]等采用卷积神经网络和自注意力融合的方法，通过在输入层加入自注意力机制来学习输入序列词与词之间的相互关系，从而丰富输入层的表示，提升了模型的效果。随着图卷积神经网络[45]（Graph Neural Network，GNN）的提出和广泛应用，有研究者将图卷积的思想引入到该任务，基于依存句法分析对句子进行建模进而建立树状网络图，然后将图卷积神经网络应用在该网络图上进行关系抽取。Guo[46]等人提出了 AGGCN 模型，该模型使用了图卷积网络对句子的依存树进行编码，主要创新点体现在通过注意力机制对树进行类似剪枝的操作，使模型能有选择地学习到依存树中对关系抽取任务有用的结构信息。

综上所述，采用深度学习的方法优势在于可以利用神经网络自动抽取特征便可以取得很好的表现效果。但现阶段该方法的研究依旧存在不足：一方面是大部分的实体关系抽取方法都在探索如何结合外部的知识库特征，外部知识库特征的有效性需要耗费很多的时间精力去验证，此外，依赖领域知识库特征的通用性不强，而且某些知识库的获取方式也会受到一定的制约；另一方面，模型对于数据本身的学习能力不足，而且对特征的优化跟组合相对缺乏更进一步的研究。本文将围绕着以上两点展开进行研究，使得本文提出的新的实体关系抽取模型可以自动挖掘出数据蕴含的丰富特征，在不必依赖外部知识库资源的同时可以取得比传统及主流模型更好的实验效果。

1.3 论文研究内容

本文的研究关注于句子级二元关系抽取，即抽取同一句子中的两实体间的语义关联关系，主要关注有监督的句子级的关系抽取。针对目前关系抽取方法中存在依赖外部知识库特征以及模型对于数据本身的学习能力不足，对特征的优化跟组合缺乏更进一步的研究等问题，本文的研究内容主要可以分为以下三个方面：

（1）本文提出一种融合词特征与相邻词间特征的关系抽取模型，该模型可以有效利用卷积神经网络以及双向长短期记忆网络的特点，并结合注意力机制来提取出基于词的句子特征，基于相邻词间关系的句子特征，进一步挖掘出自然语言文本中潜在蕴含的语义信息。实验结果显示，本模型在 SemEval-2010 Task8 数据集以及 Wiki80 数据集上进行训练和测试，测试结果相比于对比模型均有所提高。

（2）提出结合实体相关信息的多特征组合的关系抽取模型：实体及实体左右的上下文背景信息有利于实体在句中语义关系的确定，这对关系抽取任务起到非常重要的作用。为了使模型能够更好地利用到实体信息，学习到句中与给定实体有关的重要信息，本文对句子使用实体注意力进行加权处理，生成实体注意力向量特征，强化模型对句子重要信息的理解，同时减少句子中无关信息的噪音。实验结果表明，本文提出的结合实体相关信息的多特征组合的关系抽取模型有效提升了实体关系抽取效果。

（3）本文对各类特征的有效性进行验证、分析，并将模型中的各类特征进行不同的优化组合实验。

1.4 论文组织结构

本论文共分为五章，各章内容如下：

第一章：绪论。主要阐明了本文研究的方向，介绍关系抽取的研究背景及意义，并对其研究的国内外现状进行了详细叙述，最后介绍本文的核心研究内容以及各章节内容安排。

第二章：相关背景和技术。本章重点介绍与关系抽取任务有关的基础知识，词向量化的技术及模型，关系抽取任务常用数据集、评价指标，以及关系抽取常用的神经网络跟相关技术。

第三章：融合词特征与相邻词间特征的关系抽取模型。本章将主要介绍该模型的研究思路和设计方案。首先介绍模型的整体架构，然后重点对网络分模块讲解其工作机制和作用。最后，给出所需实验环境并将本模型与传统及主流方法进行对比实验，分析实验结果并解释本模型的优势。

第四章：结合实体相关信息的多特征组合的关系抽取模型。本章首先给出整体模型结构按模块详细展开，并且说明了实验环境、实验设置以及实验参数；最后对实验结果进行思考与分析。

第五章：总结和展望。对本文研究的工作做一个系统性的总结，介绍本文所做的贡献以及不足之处，并对未来的关系抽取的研究方向与方法进行了展望。

第二章　相关背景和技术

本章节主要描述与关系抽取有关的基本概念、核心技术等。同时介绍该任务中常用的两个评测数据集和评测指标，并对核心关键技术如神经网络模型和注意力机制进行阐述。

2.1 关系抽取基础

2.1.1 实体和关系

一般来讲，实体关系抽取包含两部分，分别是实体抽取以及关系抽取。实体一般指客观存在的具体的或抽象存在的事物，在实际应用中是根据具体任务而定义的。比如人名、地名、时间、组织机构等。实体识别经多年研究目前相对成熟，其在 NLP 领域中是一项基础性的任务，现在实体关系抽取研究任务更多关注于实体间关系的抽取。目前通用实体关系抽取任务一般是在给定语料及实体的前提下，抽取出句中实体对之间的关系。比如在 Wiki80、SemEval-2010 Task8 等公开的关系抽取数据集中，句子中的实体对是预先标注好的，任务目标是抽取句子中已标注的实体对间的关系。

实体一般指客观存在的具体的物化的或抽象存在的事物。实体一般包含通用实体及领域实体。通用实体如人名、地名、组织等等，领域实体如生物医学领域中的基因名称、蛋白质名称等。

关系一般表示成<实体 1，关系 1，实体 2>的形式，指不同实体或不同语言组件单元或概念间的关系。如在句子“知识图谱是典型的多学科交叉领域，是人工智能的重要分支”中，“人工智能”和“知识图谱”是“包含”关系。关系抽取模型的作用是可以自动在句子中识别出实体“人工智能”和实体“知识图谱”之间的关系是“包含”关系。

2.1.2 关系抽取

关系抽取任务是信息抽取任务中的一项重要挑战，其根本目标是抽取实体间语义关联关系。依据语料涉及范围不同分为面向开放域和面向限定域。在开放域一般不会预先设定好关系类别，而是将具有相似表达的实体归纳为同种类关系。在限定域需要事先将关系集合定义好，此时关系抽取问题转化成关系分类问题，在限定域的关系抽取任务中，其目的是判定出实体之间属于哪种给定的关系，其本质属于一个关系分类任务。本文研究的核心是限定域的句子级关系抽取任务，在句子级的关系抽取任务中，

待抽关系的实体都存在于同一个句子中。该任务是关系抽取中最基础、最经典的一项任务，深受广泛关注。例如，在句子“The fire inside WTC was caused by exploding fuel.”中，关系抽取模型能够提取出实体 fire 与实体 fuel 之间为 Cause-Effect 关系。

2.1.3 关系抽取的一般步骤

本文重点研究的是面向限定域的句子级关系抽取任务，其本质是一个关系分类问题。用深度学习的方法解决关系抽取任务，一般需要对数据进行预处理，并对句子向量化之后输入到神经网络中，然后即可通过神经网络自动地提取特征，最后将特征用于分类输出，将最终训练得到的向量信息通过归一化处理，最终映射成在预定义类别上的概率分布。

以本文研究为例，本文实验所用数据集为关系抽取领域公开数据集，分别是 Wiki80 数据集和 SemEval 数据集，这类数据集都由人工标注的属于关系抽取任务的标准数据集。原始数据集中的数据一般需要经过分词，去除特殊符号，转换数据格式，划分训练集、测试集、验证集等一系列预处理后才能输入到神经网络中提取特征。

词的向量化表示是深度学习模型进行关系抽取任务的基础。在深度学习领域，文本、图像、视频等等数据都需要将其嵌入成词向量后才能进行输入到神经网络中进行后续特征提取。本文将会在下面 2.2 节重点介绍 GloVe 这个在 NLP 领域中常用的词向量模型。

词向量首先会被输入到神经网络中，之后会在特征提取器中经过复杂的传播计算提取特征。在关系抽取任务中，常常会使用卷积神经网络（Convolutional Neural Networks,CNN）、长短期记忆网络（Long Short-Term Memory,LSTM）、注意力机制（Attention Mechanism）、图卷积神经网络(Graph Convolutional Network,GCN)以及各自相关的变种等作为特征提取器，具体可根据内容选择不同的提取器或相关组合。

将提取后的抽象句子语义特征进一步进行拼接组合成关系向量，最后经过归一化处理将提取到的关系特征归一化成概率分布，在概率分布中选择最大预测值所对应的分类结果。

2.2 词向量

在 NLP 任务中，第一步需要考虑的就是词如何在计算机中表示出来，因为计算机是不能理解文本中自然语言所蕴含的信息。在深度学习中，词通常采用独热表示(one-hot representation)或分布式表示(distributed representation)的方式。而词向量化技术就是一种将文本词转换成计算机可处理的数值向量的技术。

在初期，词向量表示方式一般采用比较简单的独热表示的方式，具体而言就是用一个一维向量表示一个单词，向量的长度等于词表的大小，向量内部的所有分量中只有一个 1，其余元素全为 0。标记为 1 的位置表示该单词在词表中的索引。其缺点也很明显，当词表中单词量增多时，词向量的维度也随之增大，编码向量就会变得很稀疏，维度灾难问题就会发生，如果词表一共有三万个单词，那么向量的维度就是三万维，单词所在索引位置为 1，其余全部为 0。除此之外，因为每个词只有索引位置的编码为 1，所以词向量间的关系是完全相互独立的，无法体现词的语义信息也无法反映出词间关系。

由于独热表示本身存在的问题，Hinton 等人[47]随后提出了词的分布式表示，其主要思想是将语料中每个词转换成固定维度的实值向量，所有单词在同一个向量空间里，语义相似的词在向量空间中也会相距更近，这种方式的好处便是能更好地捕捉到词的语义特征。深度学习的崛起带动了词嵌入技术的发展，各种词向量模型不断被提出，下面将会介绍本文所涉及到的 GloVe 模型。

GloVe[48]是一种词向量表示工具，其基本思想是基于全局词频统计，根据词的上下文背景关系来建模从而表示词的语义信息，最终将单词的语义信息表示为一个固定维度的实值向量。一般含义接近的词会在相似的语境中出现，从数学的角度用向量的余弦相似度可以很好的体现词与词之间的语义相似关系。使用 GloVe 模型从大量文本中获得词向量一般遵循以下两个步骤。

步骤 1　构建基于词的共现矩阵

首先需根据语料库中的文本内容构建一个词间共现矩阵 X，矩阵中每一个元素$X_{i,j}$代表单词 i 和单词 j 在一个滑动窗口中同时出现的次数，当滑动窗口完成对整个语料库的遍历时，就可得出共现矩阵 X。同时，为了减少距离对单词所占权重的影响，使用权重衰减因子对矩阵中的元素进行调整，衰减系数可以设为 1/L（L 表示窗口中两个单词之间距离）。

步骤 2　训练词向量

最后采用梯度下降的方式学习语料中每个词的向量化表示，其损失函数如下：

$$J = \sum_{i,j}^{N} f(X_{i,j})(v_i^T v_j + b_i + b_j - \log(X_{i,j}))^2 \tag{2.1}$$

其中，v_i和v_j是训练求得的词向量，b_i和b_j代表偏移差，N 为词库的大小。$f(X_{i,j})$表示权重函数，一般用其削弱在训练过程在某些高频词带来的干扰。

Word2Vec 词向量模型提取词特征是基于局部语料库并采用窗口滑窗的方法，通过中心词所在的上下文窗口预测词向量表示；而 GloVe 采用基于所有单词共现概率的方式，充分考虑了全局语料信息，因此其词向量在表达单词的潜在特征方面更具优势。

2.3 常用数据集和评价指标

本文选取关系抽取任务中公开的标准数据集，分别使用 SemEval-2010 Task8 数据集[49]和清华大学公开发布的有监督关系抽取数据集 Wiki80[50]。

2.3.1 SemEval-2010 Task8 数据集

SemEval-2010 Task8 数据集在关系抽取任务中经常被用来使用，该数据集是 2010 年由国际语义评估会议（International Workshop on Semantic Evaluation）发布的。该数据集存储在 txt 格式的文件中，如表 2.1 所示，一共包含 10717 个样本，其中训练数据 8000 条，测试数据 2717 条，每个句子的实体及实体关系都已人工标出，准确性很高。每一条数据都包含句子、标注实体、关系以及注释，其中头实体用标签“<e1></e1>”标出，尾实体用标签“<e2></e2>”标出，并给出有实体方向的关系类型，最后是每条数据的相关说明内容。该数据集有带实体方向的关系 9 种和不带方向的关系 other（表示实体间不存在关系）类型共 19 种关系。采用 NLP 工具 NLTK 对语料进行预处理。关系分布如下所示，如在例子“cancers were caused by radiation exposures”中，下划线的词是标注好的实体，在句子中两实体之间的 Cause-Effect（因果）关系类别已经给出。有监督的方法目的是基于标注好的数据，训练神经网络模型，进而自动地在句子中抽取到实体 cancers 与实体 exposures 之间的 Cause-Effect 关系。

表 2.1　关系类别及其分布

关系类别	例子	个数（比例）
Other	people filled with joy	1410(17.63%)
Cause-Effect	cancers were caused by radiation exposures	1003(12.54%)
Component-Whole	my apartment has a large kitchen	941(11.76%)
Entity-Destination	the boy went to bed	845(10.56%)
Product-Producer	a factory manufactures suits	717(8.96%)
Entity-Origin	letter from the city	716(8.95%)
Member-Collection	there are many trees in the forest	690(8.63%)
Message-Topic	the lectures was about semantics	634(7.92%)
Content-Container	a bottle full of honey was weighed	540(6.75%)
Instrument-Agency	phone operator	504(6.30%)

2.3.2 Wiki80 数据集

Wiki80 评测数据集是清华大学于 2019 年发布的，一般将其用在有监督关系抽取任务上，数据格式为 JSON 格式，每个样本由句子、实体、实体 ID、实体位置及其关系标签构成。关系集合中共计 80 种关系标签，均来自于维基数据中定义的关系，如：“participant of”、“director”、“instance of”、“place served by transport hub”等等。Wiki80 数据集中实体关系没有方向性，且不含“Other”关系。样本语料是来自维基百科并由人工精确标注得到的不含噪声的数据。每种关系都有 700 条样本，共计样本量 56000 条，其中 50400 条训练数据集，5600 条测试数据集。样本格式如下：{"token": ["The", "nearest", "airport", "is", "Raja", "Bhoj", "Airport", "Bhopal", "."], "h": {"name": "raja bhoj airport", "id": "Q7285421", "pos": [4, 7]}, "t": {"name": "bhopal", "id": "Q80989", "pos": [7, 8]}, "relation": "place served by transport hub"}。其中，token 对应的是句子，h 对应的是头实体以及头实体的 ID 编号信息，pos 表示实体位置信息，t 对应的是尾实体以及尾实体的 ID 编号信息，relation 表示两实体之间的关系类别。本文研究的关系抽取网络模型的目的是：当该条样本输入给模型时，模型能够根据输入样本中的句子及标记实体自动地判断出标注实体"raja bhoj airport"与"bhopal"之间存在的关系是"place served by transport hub"。

2.3.3 评价指标

在二分类任务中，会根据模型对样本的预测类别与其真实类别将抽取结果分成：真正例（True Positive，TP），真反例（True Negative，TN），假正例（False Positive，FP），假反例（False Negative，FN）。

TP：实际为正例，预测为正例。

FP：实际为负例，预测为正例。

FN：实际为正例，预测为负例。

TN：实际为负例，预测也为负例。

在关系抽取任务中常用准确率（Precision）、召回率（Recall）、F1 值（F1-Score）来评测抽取性能好坏。准确率即查准率，指的是在所有预测为正例的样本中实际的正例所占比例；召回率即查全率，指的是所有正例样本中被预测为正例的比例。其计算方式如下：

$$Precision = \frac{TP}{TP+FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.3}$$

一般而言，当精确率或召回率其中一项指标相对较高时另一个会相对较低。为了综合衡量精确率和召回率整体的效果，关系抽取任务常会选取 F1 值作为最终的评价指标。

$$F1-Score=\frac{2Precision*Recall}{Precision+Recall} \tag{2.4}$$

为了评测模型在整个语料数据集上的效果，本研究采用每个类别性能指标的宏平均（macro-average），其定义如公式（2.5）所示，公式中 n 表示关系类别数。本文的所有 Baseline 方法和本文提出的改进模型都采用宏平均作为评价指标，用 macro-F1 值来评估模型性能。

$$macro-F1=\frac{1}{n}\sum_{i=1}^{n}F1-Score_i \tag{2.5}$$

2.4 关系抽取常用神经网络

在深度学习的发展过程中，卷积神经网络始终扮演着了非常重要的角色，它是将研究人脑如何获取并理解深刻信息成功应用在机器学习上的典型例子[51]。卷积神经网络是近些年神经网络在计算机视觉领域取得突飞猛进的前提。在其它如 NLP 领域，卷积神经网络也同样被广泛使用。在深度学习领域中常被用于处理序列数据的循环神经网络也是一类非常重要且功能强大的神经网络模型，其变体模型已成功应用在许多任务上，语言翻译模型、语音及文本识别、视频分类、文本分类等等[52]，都需要一个模型能够学习这类具有序列性的数据。接下来将介绍几种关系抽取任务中常用的深度学习模型，包括卷积神经网络、循环神经网络以及针对其进行改进的长短期记忆网络和双向长短期记忆网络。

2.4.1 卷积神经网络

卷积神经网络(Convolutional Neural Network，CNN)最早是针对图像处理任务而提出的，后来也用在文本处理领域。对于序列数据 CNN 也能处理并表现出不错的效果，Zeng 首次应用 CNN 来解决关系分类问题[39]。在关系抽取任务中，相较于传统的机器学习方法，卷积神经网络不需手工手动选择特征且不依赖 NLP 工具可以取得很好的效果。

CNN 通过卷积核与原始网格数据之间的卷积运算有效地获取数据的局部信息，通过卷积核在原始网格数据上的平移与多层卷积的方式进而获取全局信息。CNN 含有卷积和池化操作，且在结构上有局部连接、权重共享的特点。局部连接指 CNN 每次会从局部的小部分数据中抽取特征，再由局部到全局；权值共享指卷积核在输入数据上滑动的时候，输入数据在变化，但卷积核权值不变，权值共享使每个卷积核在输

入的不同位置检测同一种特征提取一种特征，不同的卷积核提取不同特征，这样降低了网络的复杂性，使需要学习的参数更少，学习速度更高。

在关系抽取任务上，给定文本的二维词向量矩阵输入，卷积神经网络在输入矩阵上，通过卷积、池化、激活等操作，通过设置不同的卷积核来提取文本语义层面上不同的特征，最后用得到的高级语义特征进行关系分类任务。卷积神经网络应用在关系抽取任务中的基本流程如图 2.1 所示[39]。

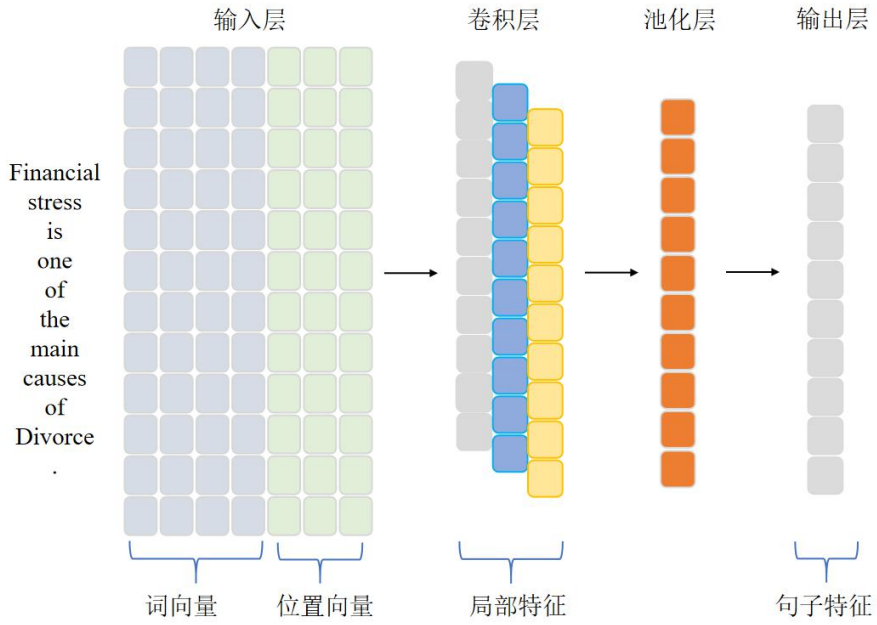

图 2.1　基于卷积神经网络关系抽取框架

（1）输入层。对输入数据进行分词、去除特殊符号等预处理后得到输入序列，加载词向量对句子进行初始化，得到输入句子的向量矩阵；一般在关系抽取任务中会同时会加入位置向量，将词向量与位置信息结合最终输入到神经网络中。

（2）卷积层。在文本处理时，卷积操作被定义为每个权重矩阵$W^k \in R^{c\times d}$(W^k表示第 k 个卷积核，c 表示卷积核大小，d 表示词向量维度)和每个句子向量$S = (e_1, e_2, ..., e_n)$之间的操作，其中$e^i \in R^d$，n 是句子长度。卷积过程用滑动窗口大小为 C 的卷积核与每个输入句子向量进行卷积计算，每个卷积核对句子向量进行卷积计算都会得到一个特征向量，最终得到一个矩阵c_k(由 K 个不同卷积核提取到的特征向量)，计算公式如下：

$$c_k = f(W^k e^{i:i+c-1} + b^k) \tag{2.6}$$

其中，$e^{i:i+j}$表示$e^i ... e^{i+j}$特征向量的连接，$i = 1,2, ..., n-c+1$,f表示 ReLu 激活函数，W^k，b^k是要学习的参数，卷积核的步长通常设为 1，在输入的向量矩阵上进行滑动，最终获得的局部特征表示集合c_k。

（3）池化层。池化层是对上一层得到的特征进行降采样，一定程度上可以减小过拟合并可以保留重要信息。本文采用最大池化操作，最大池化是选择池化区域中的最大值代表整个区域。将上述得到的局部特征按最大池化层挑选c_k中最大的值，计算如下。

$$c_{max}=\max(c_k) \tag{2.7}$$

（4）输出层。在关系抽取中，输出层一般是将池化操作得到的向量作为关系分类的最终向量，将最终的特征向量输入 softmax 进行分类，输出每一种关系的预测概率，对于每个句子，与最大概率值对应的类别就是分类结果。

2.4.2 循环神经网络

CNN 虽然可以从句子中抽取词间的深层次的抽象语义信息，但每一层的神经元，前一个输入和后一个输入之间不存在关联关系，所有输出都是独立的，因而无法获取句子中上下文间的依赖信息。现实生活中很多数据间有很强的前后逻辑关系，比如在自然语言表达中，句子与句子之间，词与词之间都具有很紧密的前后逻辑联系。例如，“我的祖国是中国，所以我的（）说的很好。”在这个例子中，容易根据句子背景信息理解得到括号内应表达“汉语”的意思，但是机器却不知道。因此，在这种具有逻辑性的序列任务上就要求模型具备信息记忆的功能，模型每一时刻的输出都能考虑到之前的输入信息。所以，为了让模型具有一定记忆的能力，循环神经网络(Recurrent Neural Network, RNN)被提出并不断发展起来。

循环神经网络在时间序列上是不断循环重复的，其循环结构能考虑到每一时刻的输入信息，对应文本中词的从前往后的每个位置顺序，能够学习句子前文依赖信息。循环神经网络中待学习的参数在不同时刻是共享的，其网络结构如图 2.2 所示，左侧展示的是 RNN 的基本结构单元，右侧是将其在时间维度上进行展开的网络形式。

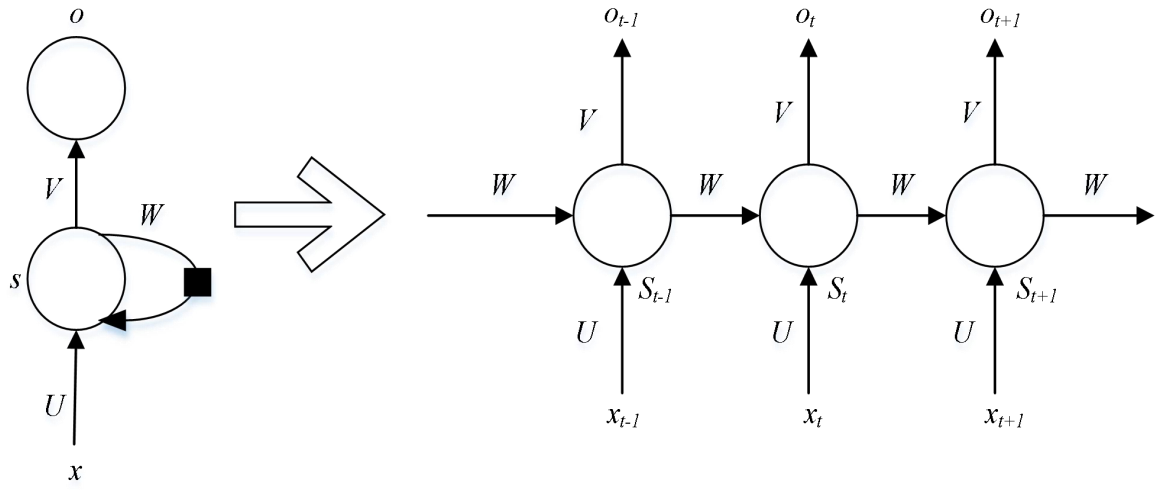

图 2.2　RNN 模型结构图

在图 2.2 中，x_t表示 t 时刻序列的输入，S_t表示 t 时刻的隐藏状态的值，表示记忆单元信息，记忆单元信息取决于当前时刻的输入x_t与前一时刻的单元状态S_{t-1}，计算公式(2.8)所示。

$$S_t = f(Ux_t + WS_{t-1}) \tag{2.8}$$

上式中，U、V 分别表示的是输入层到隐藏层以及隐藏层到输出层之间的权重系数矩阵；f 指的是神经网络中的激活函数，常常使用 tanh 函数。

t 时刻，输出向量o_t取决于 t 时刻的单元状态S_t，其计算过程如公式(2.9)所示，V 代表隐层到输出层的权重系数矩阵。

$$o_t = VS_t \tag{2.9}$$

当前 t 时刻的输出表示前 t-1 时刻跟当前时刻输入的所有信息，最后时刻的输出代表学习的整个句子的信息。最后，将o_t输入到 Softmax 函数中得到各种关系对应的概率分布。在整个训练过程中，参数矩阵 U、V、W 在时间维度上是共享的，这大大降低了参数量同时提高了训练速度，并采用反向传播算法对训练参数进行更新。

虽然 RNN 相比于前馈神经网络在处理序列任务时更有优势，但其自身也存在一些问题，当序列文本较长的时候，在反向传播的训练过程中需要计算损失函数对权重的梯度，但随着向后传播的加深，梯度会逐渐变小，这意味着在网络中靠前的一些神经元的训练速度要比后面的慢很多，甚至不会发生变化，致使结果不准确，训练时间非常长，长距离的依赖信息无法获取，易出现梯度消失或爆炸的问题。

2.4.3 长短期记忆网络

长短期记忆（Long short-term memory, LSTM）是在 RNN 的基础之上改进的一种结构更复杂的循环神经网络，其主要目的为了解决对长序列进行训练的过程中发生的梯度消失或梯度爆炸问题。简单地说，相比原始的 RNN，当序列比较长的时候，LSTM 会表现出更好的效果。长短期记忆网络使用一种特殊的记忆单元，通过精心设计的“门”结构选择性控制输入序列在网络记忆单元中信息的流入流出，这种“门”结构使得关键信息得以保留，无关紧要的信息被丢弃。门指的是将不同的信息进行按位点乘之后再经过一个 sigmoid 层的操作。sigmoid 函数可以将任意输入控制在 0 到 1 之间，这种特性代表着让信息通过的概率。

目前，基于长短期记忆网络的模型在机器翻译、文本生成、语音识别、生成图像描述、关系抽取等众多方面表现出了非常好的效果。长短期记忆网络的神经元因具有“门”结构，故而相比于基本循环神经网络显得更复杂，其模型结构如下图 2.3 所示。

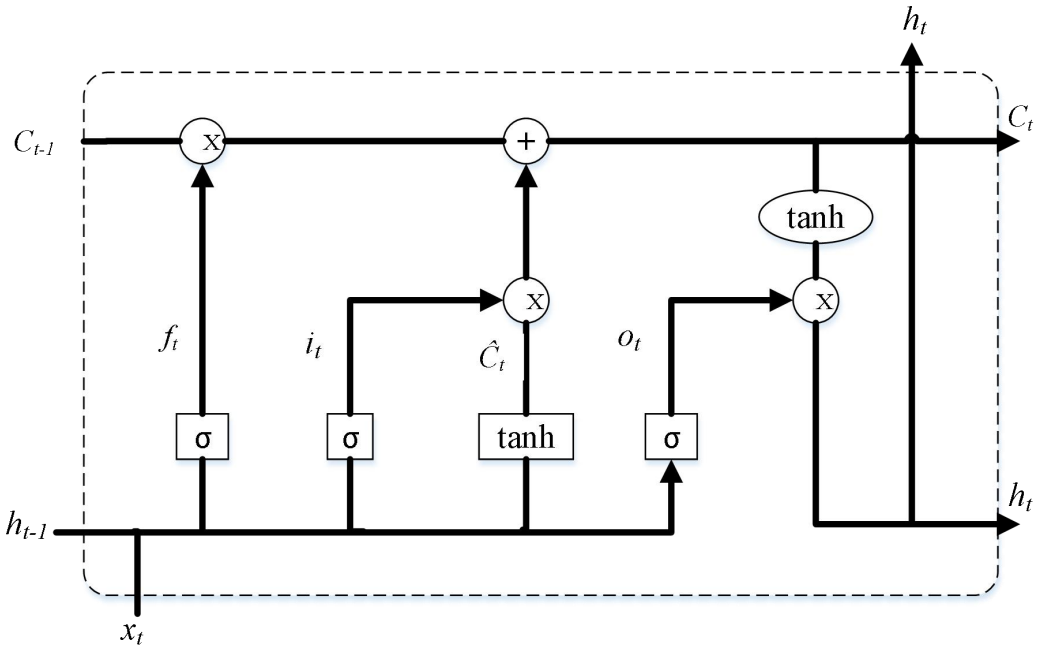

图 2.3　LSTM 单元结构图

LSTM 单元由一个记忆单元三个基本门组成：输入门i_t通过相应的权重系数矩阵W_i，偏执项b_i控制输入到单元中的信息；遗忘门f_t通过相应的权重系数矩阵W_f，偏执项b_f控制单元中信息的去除；输出门o_t通过权重系数矩阵W_o，偏执项b_o控制信息从单元的输出。所有门都需要当前时刻的输入x_t与上一时刻的隐含层状态向量h_{t-1},它们的计算如下：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.10}$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.11}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.12}$$

其中，$[h_{t-1}, x_t]$表示将前一时刻单元的隐层状态信息与当前时刻的输入进行拼接形成一个向量表示，σ表示 sigmoid 函数。

当前时刻特征向量的计算方式如下，当前时刻的特征也需要上一时刻的隐层状态向量h_{t-1}与当前的输入x_t。

$$\widetilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_g) \tag{2.13}$$

然后，当前记忆单元C_t会整合上一刻的记忆单元信息c_{t-1}与当前时刻的特征$\tilde{C}_t$，并结合输入门以及遗忘门来选择性地保留跟丢弃某些信息。初始记忆单元C_0初始化为全零向量，计算方式如下。

$$C_t = f_i \cdot C_{t-1} + i_t \cdot \widetilde{C}_t \tag{2.14}$$

通过输出门之后最终会得到最后时刻的隐层状态向量h_t。

$$h_t = o_i \cdot \tanh \quad (C_t) \tag{2.15}$$

h_t表示最终输出，取决于输出门的结果o_t和当前时刻的记忆单元状态C_t。将代表句子序列的二维向量输入到长短期记忆网络中，经过上述一系列的计算过程最终得出输出向量h_t。

2.4.4 注意力机制

注意力机制（Attention Mechanism）在各类机器学习任务上普遍表现出良好的性能，因此成为当下深度学习领域中的研究热点，越来越多的研究人员也将注意力机制引入到 NLP 相关任务中来[53]。注意力机制提出的灵感来源于我们人类，当我们用眼观察图像时，一般会不自觉地将注意力放到一些相对特殊的或引起我们兴趣的区域，而不会关注图像上的所有细节。

注意力机制最初用在 NLP 领域中机器翻译任务上，并较好地提升了翻译模型的性能。在翻译模型中，注意力机制的目的是学习所需要翻译的词与每个原始待翻译词之间的关键特征信息，具体而言就是通过注意力机制对输入句子中的不同词赋予不同的权重，从而使得网络模型可以学到重要部分的信息同时削弱不重要信息的影响，最终让模型做出更精准化的判断。另一方面，注意力机制亦可缓解神经网络中存在的梯度爆炸或梯度消失问题[54]。注意力机制的本质是一个函数，通过计算获得对数据的注意力概率分布。

“编码-解码”结构常用在机器翻译、关系抽取等序列到序列的任务中，如图 2.4 所示为传统的编码-解码过程图。

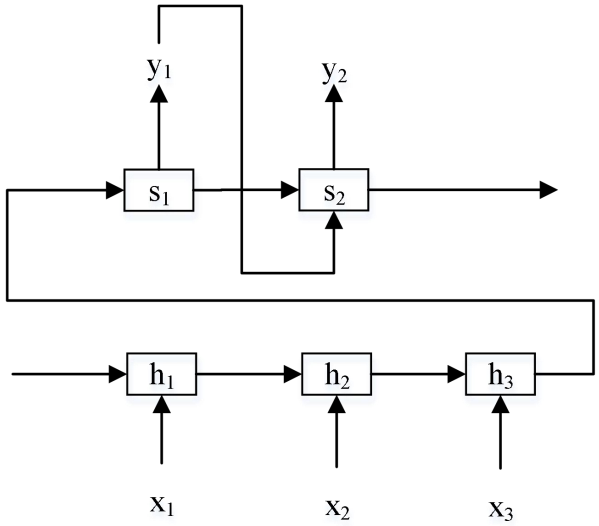

图 2.4　编解码结构图

首先将文本序列（$x_1,x_2,\cdots,x_t$）输入到编码器内，其中 t 表示序列的长度，通过编码器（如 LSTM）将输入序列编码成长度大小固定的隐藏向量（$h_1,h_2,\cdots,h_t$），然后将固定长度的向量h_t作为输入，最后经过解码器（如 LSTM）的解码操作最终生成与输入序列相对应的输出序列（$y_1,y_2,\cdots,y_t$）。传统的编码-解码结构存在着两个问题。一是传统的编码结构会将输入序列编码为一个固定不变的中间向量h_t，解码器基于这个固定的中间向量进行解码会导致信息丢失。其次，对输入和输出序列无法实现自动对齐。

注意力模型通过让解码器遍历整个编码后的输入序列（$h_1,h_2,\cdots,h_t$）来解决以上问题。具体而言是就是通过引入注意力权重α对输入序列进行加权处理，以此来关注某些具有相关关系的特定位置的信息，然后生成下一个输出。

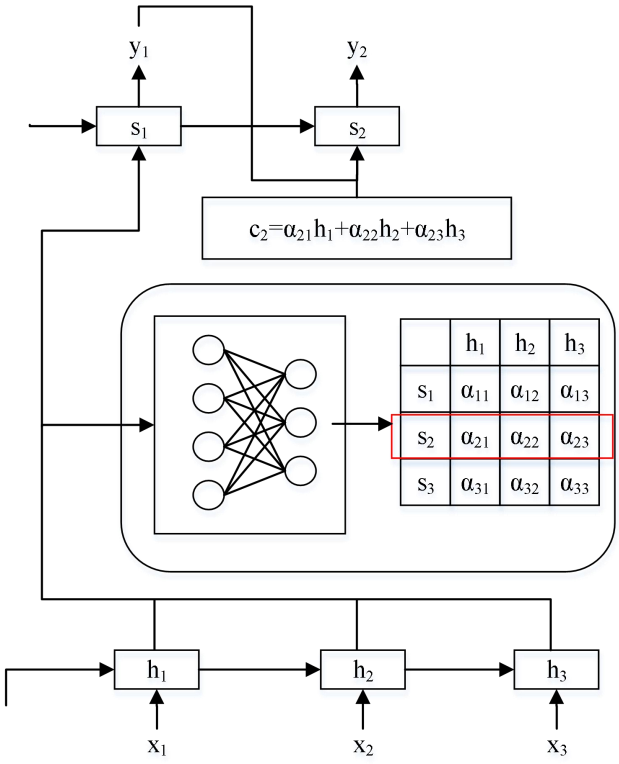

图 2.5　加入注意力机制的编解码结构图

加入注意力机制的编-解码器结构如图 2.5 所示。网络中的注意力模块通过自动捕捉编码器隐藏状态h_i（也称候选状态）和解码器隐藏状态S_j（也称查询状态）之间的相关关系来学习注意力权重α_{ij}。最后，将所有注意力权重与编码器的隐藏状态结合构建出当前时刻的向量 C，并将其传给解码器。针对每个待解码的位置 j，向量C_j是将编码器中的所有隐藏状态与其对应的注意权重进行加权求和，该过程计算方式如下 2.16 所示。

$$C_j = \sum_{i=1}^{T} \alpha_{ij} h_i \tag{2.16}$$

注意力权重由网络中引入的前馈神经网络学习得到，注意权重α_{ij}是关于每个编码器隐藏状态h_i和解码器隐藏状态S_{j-1}的函数。第 j 个位置的注意力得分α_{ij}的计算公式如 2.17 所示。

$$\alpha_{ij} = \frac{\exp\ (sim(s_{j-1}, h_i))}{\sum_{k=1}^{T} \exp\ (sim(s_{j-1}, h_k))} \tag{2.17}$$

2.5 本章小结

本章为关系抽取相关背景和技术部分，介绍了本文研究的理论基础、常用数据集、性能评判标准及相关技术。首先介绍了相关理论基础，如什么是实体、关系是什么、词的向量表示等；然后具体介绍了常用的开源数据集并详细介绍了评价模型性能的评价标准；最后对本文所涉及到的技术做了详细的介绍，如 CNN、BiLSTM 以及注意力机制等。

第三章　融合词特征和相邻词间特征的关系抽取模型

3.1 研究动机

针对大多数现有的实体关系抽取方法过于依赖外部特征资源，且未能充分考虑并挖掘出句子内部所蕴含的信息从而阻碍了关系分类性能进一步提升的问题，本章提出了融合词特征和相邻词间特征的关系抽取模型。本文提出将词特征、相邻词间关系特征融合到深度神经网络模型中。具体来说，在卷积神经网络中，利用卷积核与句子之间的卷积操作可以有效地提取到基于局部相邻词的句子特征，通过卷积核在输入数据上的平移与多层卷积的方式进而获取全局信息。卷积神经网络虽然可以从句子中抽取词间的深层次的抽象语义信息，但每一层的神经元前一个输入跟后一个输入之间没有任何联系，所有输出都是相互独立的，因而卷积神经网络无法获取到句子中上下文前后的依赖信息。双向长短期记忆网络对信息具有记忆的功能，能够学到序列中每个词的上下文信息从而能对每个当前词做出基于上下文背景信息的精准判断。本章在结合了卷积神经网络、双向长短期记忆网络各自的特征优势并弥补了各自在特征抽取上不足的同时引入注意力机制，提出了一种新的关系抽取模型对句子进行建模。

本文在充分研究了句子内部的潜在结构信息之后，提出了相邻词特征的概念，为了更好地挖掘并利用到句子中的信息，达到提升关系抽取效果的目的，本文提出融合词特征以及相邻词间的特征的关系抽取模型，该模型能够更有效地建模句子的局部跟全局语义信息，提升了模型抽取的性能。

3.2 模型设计

3.2.1 模型整体架构

本文提出一种新的融合词特征与相邻词间特征的关系抽取模型，该模型通过 Att-BiLSTM 网络模块与 CNN-AttBiLSTM 网络模块分别提取出句子的词特征、相邻词间特征，进一步挖掘出自然语言文本中潜在的语义信息，最终将两类特征进行融合完成关系分类任务。模型结构如图 3.1 所示。

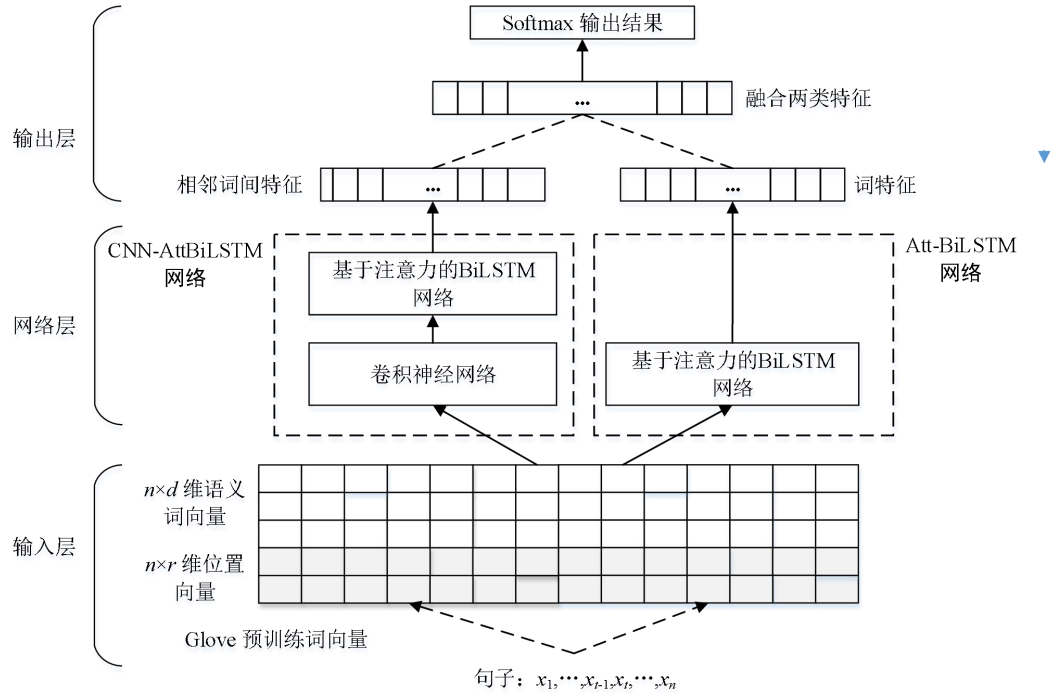

图 3.1　融合词特征和相邻词间特征的关系抽取模型

3.2.2 输入层

神经网络不能对输入文本直接进行编码，所以在用神经网络处理自然语言文本时首先要对语料进行预处理。对原始数据进行分词、去除特殊符号、大写转小写、分离标点符号等操作，并将预处理后的原始语料转化成特定格式（如 JSON 格式），进而将其转化为向量形式输入到神经网络中。本文在输入层使用了词嵌入以及位置嵌入两部分，将词与位置向量结合作为模型的输入。

（1）词向量

本文研究所用的 GloVe 向量是由斯坦福大学提供的，通过预训练的词向量来初始化句子的表示，预训练的词向量比随机初始化表示更能准确体现其语义信息，将每个词映射成 N 维实值向量进而输入到模型中。如给定句子 $S = (w_1, w_2, ..., w_m)$，通过 GloVe 词向量矩阵，每个词都将被映射为向量表示。

（2）位置向量

词向量虽然能表示词在语义层面上的含义，却无法蕴含词所在句中的位置信息。一般而言，距离实体越近的词往往越能体现出关键的信息。本文参考了 Zeng[48]的计算方式，通过计算句中每个词分别到两个标记实体的相对距离来表达位置特征。例如:在句子“My cat has a problem with his paw.”中，词 problem 到实体 cat 的相对距离是 2，到实体 paw 的相对距离是-2。两个相对位置信息会被映射成两个随机初始化的 p 维实值的向量。

最终，将句子中各个词的词向量表示与位置向量表示拼接起来，作为模型的输入，用矩阵 $D \in R^{s \times d}$表示（s 指句子长度，d 表示向量拼接后的维度，即 $d = d_w + 2 * d_p$）。

3.2.3 Att-BiLSTM 网络模块

长短期记忆网络无法获取到输入序列每个位置词的上下文信息，为了捕获原始句子每个时间步 t 时刻之前和之后的上下文信息，本章采用基于双向长短期记忆网络模型来捕获基于词的句子特征。BiLSTM 是对 LSTM 的一种改进，从正反两个方向学习句子序列的语义信息。

本文采用基于注意力机制的 BiLSTM 网络学习句子的词级特征，具体模型结构如图 3.2 所示。模型的输入为序列$[X_1, X_2, ..., X_n]$，在输入层得到每个词的向量表示形式 V_i（词向量与位置向量的组合），并将其输入到 BiLSTM 中。BiLSTM 层有前向后向两个不同方向的子网络。Attention 层会对每个在 BiLSTM 层输出的向量做加权计算[43]。该模型引入注意力机制的目的是找到对关系分类影响更大的词并赋予更大的权重，从而在句子内部捕捉更重要的语义信息。

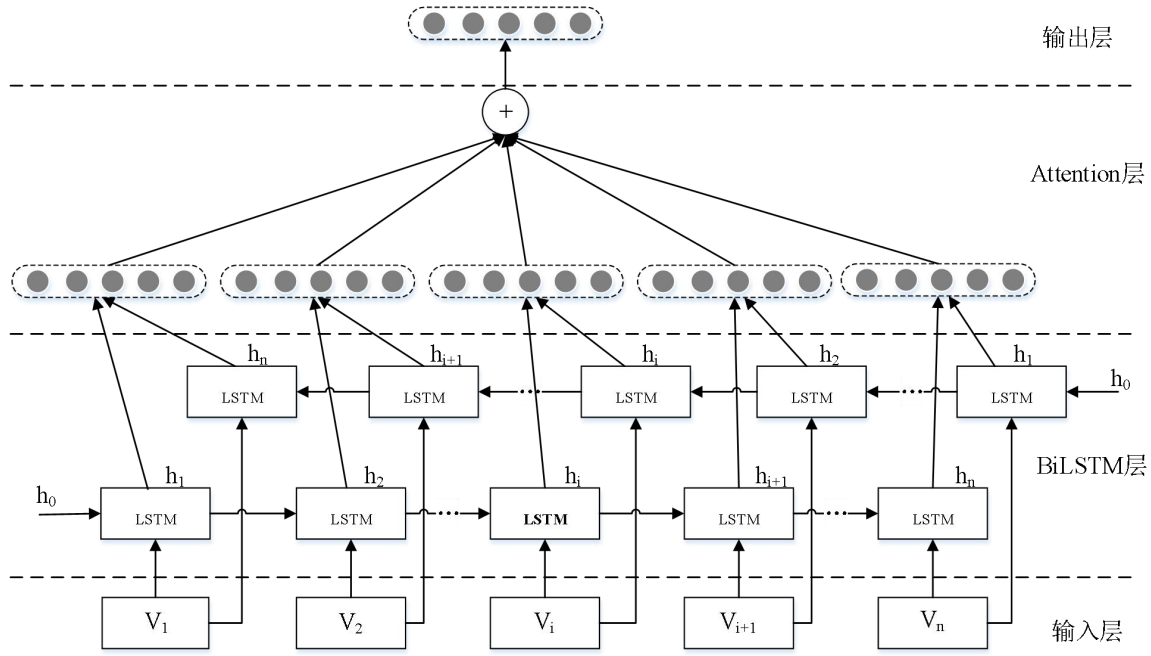

图 3.2　Att-BiLSTM 网络结构图

为了捕获原始句子每个时间步 t 时刻之前和之后的上下文信息，构建了双向长短期记忆网络结构。BiLSTM 最终的隐层状态是由前向与后向 LSTM 得到的隐层状态拼接而成，正向 LSTM 能够学到当前输入的前文信息，反向的 LSTM 能够学到当前输入的后文信息。BiLSTM 模型从两个方向更好地学习到每个词的上下文背景信息，得到每个词两个方向上的隐藏状态向量。LSTM 输出门最终输出隐层状态向量h_t。

$$h_t = o_i \cdot \tanh \quad (C_t) \tag{3.1}$$

最后将两个序列的输出相加得到 H，即为通过双向 LSTM 网络的最后结果。双向 LSTM 隐藏状态向量h_t的计算公式为：

$$h_t = \left[\overrightarrow{h_t} \oplus \overleftarrow{h_t}\right] \tag{3.2}$$

将原始句子的词向量矩阵输入到 BiLSTM 编码器，输出是一个新的矩阵，表示为：

$$H = [h_1, h_2, ..., h_n](H \in R^{n*d^h}) \tag{3.3}$$

其中，d^h是隐藏状态向量h_i的维数。

句子中的每个词与实体关系的相关性是不同的，有些词可能在句子的语义中起着重要的作用，而有些字符可能不起作用。因此，利用注意机制来学习每个词的权重，从而衡量其对实体之间关系的贡献。所有权重的计算如下所示：

$$M = \tanh \quad (H) \tag{3.4}$$

$$\alpha = \text{softmax} \quad (w^T M) \tag{3.5}$$

$$r = H\alpha^T \tag{3.6}$$

其中 w 是维数为d_h的随机初始化向量，α是需要训练的 n 维向量，即为注意力权重参数，r 是注意力加权后的维数为d_h的句子表示向量。最后H_1=tanh(r)为基于词特征向量的句子表示，Att_BiLSTM 网络模块最终得到基于词的句子表示H_1。

3.2.4 CNN-AttBiLSTM 网络模块

上文使用 Att-BiLSTM 编码器已经学习了原始句子基于词的特征，但是模型仅基于词特征其表达能力有限，无法获得句子中相邻词间特征，这使得捕获句子的信息不够充分。在这种情况下，本文首先使用卷积层用于学习基于词嵌入的相邻词局部特征。卷积层的输出包含窗口内上一个词到下一个词相邻词间的局部信息，然后再将卷积层的输出输入到 Att-BiLSTM 编码器以学习相邻词间特征，具体网络模型如图 3.3 所示。

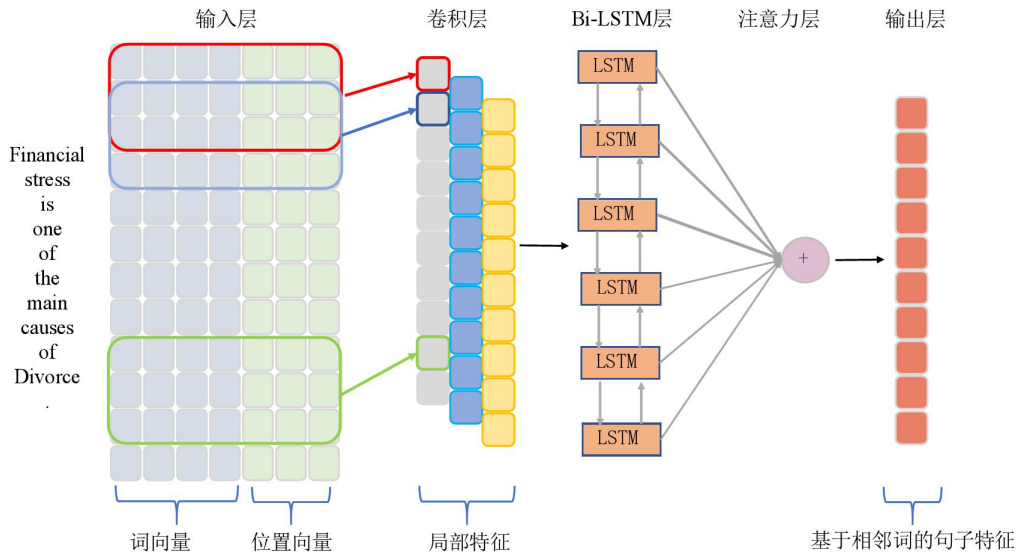

图 3.3　CNN-AttBiLSTM 网络结构图

卷积被定义为每个权重矩阵$W^k \in R^{c\times d}$(W^k代表第 K 个滤波器，C 表示滤波器大小，d 是词向量的维度)和每个句子向量 $S = (e_1, e_2, ..., e_n)$之间的操作，其中$e^i \in R^d$，n 是句子长度。卷积操作指滑动窗口在每个句子向量上的卷积计算，每个滤波器的卷积都会得出一个特征向量，最终得到一个矩阵c_k，计算公式如下：

$$c_k = f(W^k e^{i:i+c-1} + b^k) \tag{3.7}$$

其中，$e^{i:i+j}$表示$e^i ... e^{i+j}$特征向量的连接，$i = 1,2, ..., n-c+1$,f表示 ReLu 激活函数，W^k，b^k是需要学习的参数。

从卷积层输出的局部特征向量被视为相邻词特征向量。通过一系列不同滤波器，每个输出向量表示不同的语义信息。CNN 的输出向量可以表示为$H = [h_1, h_2, \ldots, h_k]$的相邻词间向量矩阵，其中 k 是滤波器的数目。最后，将卷积操作得到的相邻词特征向量矩阵输入到 Att-BiLSTM 网络中，从而最终获得基于相邻词间关系特征的句子向量表示H_2。

3.2.5 关系分类层

本文提出的融合词特征和相邻词间特征的关系抽取模型关系抽取模型，以句子的词向量及位置向量作为输入，通过 Att-BiLSTM 网络模块与 CNN-AttBiLSTM 网络模块分别提取出句子的词特征H_1、相邻词间特征H_2。然后，将词特征向量H_1，相邻词间特征向量H_2拼接作为关系分类的最终特征向量O，并将其输入 softmax 分类器中以计算概率：

$$O = [H_1, H_2] \tag{3.8}$$

$$p(y) = softmax(O) \tag{3.9}$$

对于每个句子，与最大概率值对应的类别就是分类结果。

$$\hat{y} = arg\max_{y} p(y) \tag{3.10}$$

目标函数是带L_2正则化的交叉熵损失函数，如式(12)所示。

$$J(\theta) = -\sum_{i=1}^{m} t_i \log(y_i) + \lambda \|\theta\|_F^2 \tag{3.11}$$

其中，λ表示正则化系数，t_i是真实标签的 One-hot 表示，y_i表示每个类别的预测概率，m 代表预定义的关系分类的数目，为了防止过拟合，在模型的输入层、隐藏层以及输出层都加入了 dropout 操作。

3.3 实验与分析

本节将主要介绍实验数据、实验过程中的实验环境及配置、参数设置具体细节以及实验的最终结果，并对结果进行总结分析。

3.3.1 数据集

本研究任务使用 2.3 小节介绍的 SemEval 2010 Task8 数据集和 Wiki80 数据集。其中，SemEval 2010 Task8 数据集包含 10717 条标注句子，共 19 类关系，训练数据集有 8000 条样本，测试集 2717 条。Wiki80 数据集含 80 种关系，每种关系对应 700 条样本，训练集规模有 50400 条数据，测试集 5600 条数据。对两种数据分别进行预处理操作，将大写转小写、去除特殊标识符并转换成特定的 JSON 格式。如：{"id": "8001", "relation": "Message-Topic(e1,e2)", "head": "audits", "tail": "waste", "subj_start": 4, "subj_end": 4, "obj_start": 9, "obj_end": 9, "sentence": ["The", "most", "common", "<e1>", "audits", "</e1>", "were", "about", "<e2>", "waste", "</e2>", "and", "recycling", "."], "comment": " Assuming an audit = an audit document."}，其中 id 表示样本编号，relation 表示人工标注的实体在句中的关系，head 表示头实体，tail 表示尾实体，subj_start 表示头实体的起始位置，subj_end 表示头实体的末尾位置，obj_start 与 obj_end 分别表示尾实体的起始与末尾位置，sentence 表示实体所在的句子，comment 是对样本的描述。

3.3.2 实验环境

本文研究模型是基于 PyTorch 实现的，并在 SemEval-2010 Task8 数据集及清华大学发布的 Wiki80 数据集上进行实验。实验环境与配置如表 3.1 所示。

表 3.1 实验环境与配置

实验环境	配置
操作系统	Ubuntu16.04.6
CPU	Intel(R)Xeon(R)Gold5218CPU
GPU	GeForceRTX2080Ti1
内存	256G
编程语言	Python3.8
深度学习框架	PyTorch1.8.0

3.3.3 模型参数设置

实验调整超参数采取的是网络搜索方法，超参数范围是一定可选的，进行大量实验来优化模型参数，同时为了不同实验之间进行对比，对本文所有 baseline 模型以及改进方案中的相同的模块选择同样的参数。

表 3.2　超参数设置

参数	值
词向量维度	100
位置向量维度	10
句子长度	100
批处理大小	10
学习率	1.0
L2 正则化系数	1.00×10^{-5}
词嵌入层 dropout	0.3
LSTM 层 dropout	0.3
其他 dropout	0.5
CNN 卷积核窗口大小	3
CNN 卷积核个数	110
LSTM 单元大小	100
随机种子	5782
epoch	100

实验中的词向量使用预先训练好的 100 维的 GloVe 词向量，位置向量的维度取 10 维。使用梯度下降法来更新网络中的参数。学习率设置为 1.0。使用 dropout 方法和L_2正则化来处理过拟合问题，嵌入层以及 BiLSTM 层的 dropout 值设置为 0.3，其他部分设置为 0.5。L_2正则化的系数 λ 取值为 1.0×10^{-5}。模型中的其他参数矩阵采用正

态分布进行随机初始化。通过比较不同参数下的结果，本文实验最优结果部分参数设置如上表 3.2 所示。

3.3.4 实验结果分析

为了验证本章模型对于实体关系抽取的稳定性和有效性，采用对比实验方法，同时在 SemEval-2010 Task8 和 Wiki80 数据集上与如下 3 种经典主流模型对比。本文选取经典的 CNN、CRCNN 模型以及主流的 BiLSTM 模型与本文提出模型进行对比，在输入层，都采用统一的词向量和位置向量作为模型输入。

CNN：文献[39]的方法。该方法由 Zeng 等于 2014 年首次提出用卷积神经网络做关系抽取任务，文中提出了位置特征，将句子特征与语法特征进行融合进行关系抽取去任务，语法特征包括实体词及实体词左右词，WordNet。

CRCNN：文献[42]提出的方法。该方法由 Santos 等于 2015 年提出，CRCNN 通过对 softmax 损失函数进行改进，用加权的 Softmax 损失函数解决类别不平衡问题，在采样方式上提高类别数比较少的权重，同时去除了 WordNet 等复杂的特征。

BiLSTM：参照文献[43]提出的方法。该方法由 Zhou 等于 2016 年提出，该方法用经典的 BiLSTM 作为模型的主要模块，利用了注意力机制对 BiLSTM 的输出进行注意力加权求和，最后用加权后的向量进行关系分类。

Our_model1：本文提出的融合词特征与相邻词间特征的方法，该方法利用 Att-BiLSTM 提取基于词的句子特征，在此基础之上，利用 CNN-AttBiLSTM 提取出句子内部的相邻词间关系特征，并通过注意力机制对特征向量加权处理，最后进行关系分类。

下面分别在两个公开数据集上进行两组实验，实验结果如表 3.3、表 3.4 所示。

表 3.3　SemEval 实验结果对比

Models	Featureset	macro-F1/%
CNN	word,position,word pair,words around word pair	79.47
CR-CNN	word,position	80.24
BiLSTM	word,position	82.23
Ourmodel_1	word,position,Word and Adjacent-word	83.52

表 3.4　Wiki80 实验结果对比

Models	Featureset	macro-F1/%
CNN	word,position,word pair,words around word pair	78.00
CR-CNN	word,position	78.64
BiLSTM	word,position	79.78
Ourmodel_1	word，position,Word and Adjacent-word	81.72

由表中的实验结果可以发现，本文提出的模型效果比 CNN、CR-CNN 以及 BiLSTM 模型更好。本文提出的改进的模型结合 CNN 与 BiLSTM 各自的优势与特性，并弥补了因各自网络本身的局限性所带来的在特征抽取问题上的不足，抽取出句子内部蕴含的相邻词间特征以及词特征，然后通过注意力机制对特征进行进一步的筛选优化。最后，将两种不同的特种组合作为最终的关系分类向量。在使用相同输入的前提下，本模型考虑到句子内部的结构特征，直接从原始句子中获取特征，在不依赖人工经验设计特征且不依赖外部知识库的同时从内部挖掘出更多的特征信息，从而提升了模型的效果。

为了直观地显现在不同数据上各模型的表现性能，进一步将实验结果用折线图形式表示，下图 3.4、图 3.5 对应上表 3.3 及 3.4 中 CNN、CRCNN、LSTM、Our_model1 四种模型分别在 SemEval-2010 Task8 及 Wiki80 数据集上 macro-F1 值随迭代次数的变化曲线。由图可知，两种数据集上 F1 值随着各模型迭代次数的增加趋于稳定，本文提出的模型整体效果优于所有对比模型。

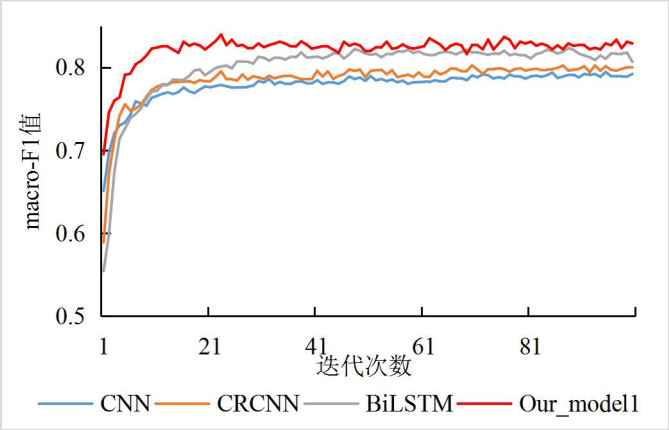

图 3.4　SemEval 数据集上 macro-F1 值随迭代次数变化曲线

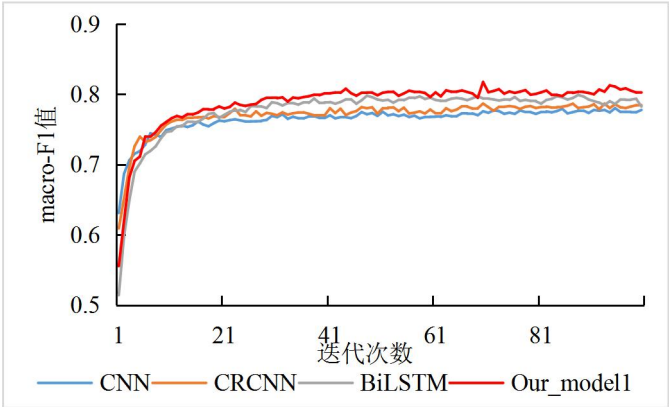

图 3.5　Wiki80 数据集上 macro-F1 值随迭代次数变化曲线

另外，为了发现实验结果与数据集各类别分布之间的关系，本文对测试数据进行统计分析，探索不同类别的数据在关系分类表现上的差异。本实验统计了在两种数据集下各个关系分类下各模型准确率、召回率以及 F1 值的性能。图 3.6、图 3.7、图 3.8 为各模型在 SemEval 数据集下关系的准确率、召回率以及 F1 值曲线，可以发现，本文提出的融合词特征和相邻词间特征的关系抽取模型在各类关系上均优于对比模型，整体上的表现效果更好。在某些特定关系类别下明显优于其他类别。

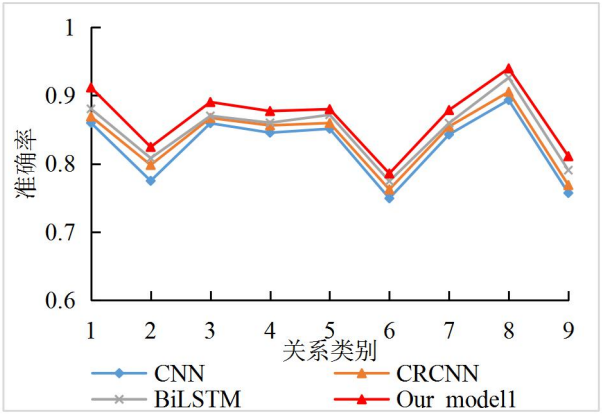

图 3.6　Sem 数据集上各关系准确率比较

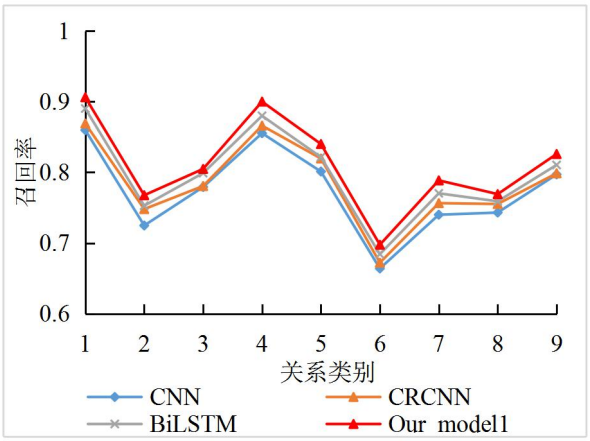

图 3.7　Sem 数据集上各关系召回率比较

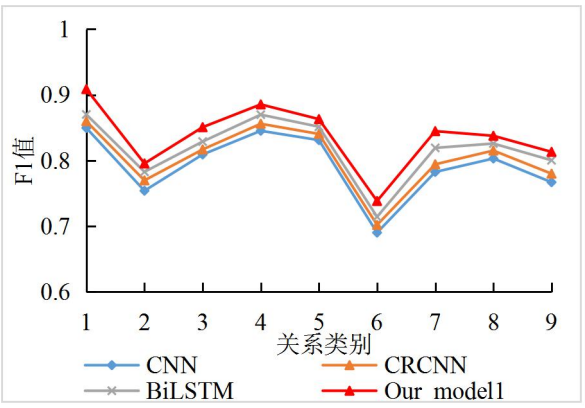

图 3.8　Sem 数据集上各关系 F1 值比较

图 3.6、图 3.7 以及图 3.8 中数据集数字编号对应的关系类别如下表 3.5 所示：

表 3.5　关系类别对应编号

关系类别	关系类别编码
Cause-Effect	1
Component-Whole	2
Content-Container	3
Entity-Destination	4
Entity-Origin	5
Instrument-Agency	6
Member-Collection	7
Message-Topic	8
Product-Producer	9

图 3.9、图 3.10、图 3.11 中，展示了 Wiki80 数据集前 20 种关系类别各自对应的准确率、召回率以及 F1 值曲线，进一步验证了本文提出模型的有效性。

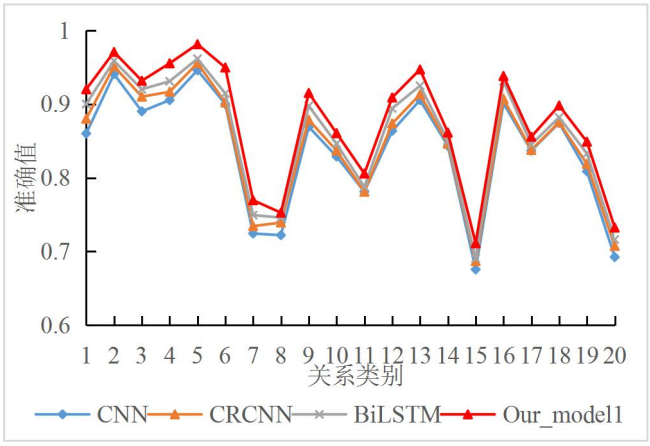

图 3.9　Wiki80 数据集上前 20 种关系类别准确率比较

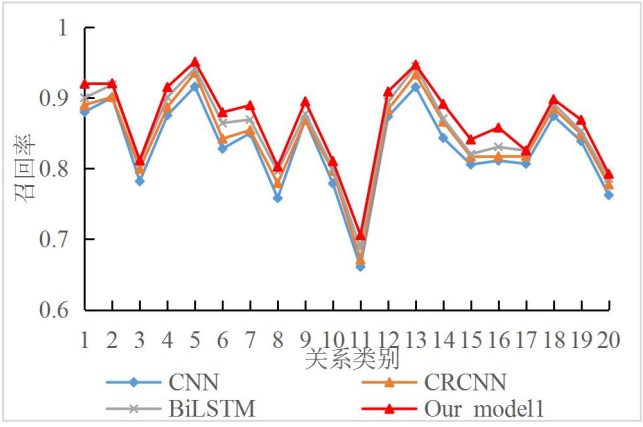

图 3.10　Wiki80 数据集上前 20 种关系类别召回率比较

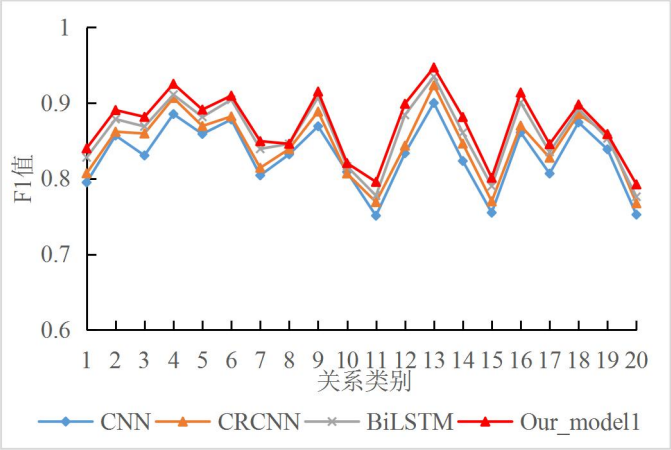

图 3.11　Wiki80 数据集上前 20 种关系类别 F1 值比较

上图 Wiki80 数据集中数字编号对应 Wiki80 的部分关系类别如下表 3.6 所示：

表 3.6 关系类别对应编号

关系类别	关系类别编码
place served by transport hub	1
mountain range	2
religion	3
participating team	4
contains administrative territorial entity	5
head of government	6
country of citizenship	7
original network	8
heritage designation	9
performer	10
participant of	11
position held	12
has part	13
location of formation	14
located on terrain feature	15
architect	16
country of origin	17
publisher	18
director	19
father	20

3.4 本章小结

本文提出了一种新的结合卷积神经网络、双向长短期记忆网络和注意力机制的网络模型，该模型可以有效抽取出基于词的句子特征以及基于相邻词间特征的句子特征，实验结果表明，在上述两个公开数据集上都取得了不错的效果。

第四章　结合实体相关信息的多特征组合的关系抽取模型

4.1 研究动机

通过构造 Att-BiLSTM 编码器和 CNN-AttBiLSTM 编码器，捕获句子中基于词级和相邻词间的句子特征，有效地表示了原始句子的语义信息。此外，实体以及实体左右的上下文背景信息有利于确定实体在句中的语义关系。因此，在不依赖外部知识的前提下，构建了一个能学习句子连续特征，又能捕获到实体及实体在句中复杂背景关系的神经网络模型对关系抽取研究具有重要意义。为了让模型学到句中与实体词相关的重要信息，本文通过在模型中加入了实体注意力机制，来帮助模型学习到原始句子中有关实体的相关语义信息。

实体注意力网络模块使得模型能够主动去确定句子的各个部分对两个给定的实体词的影响力强弱，因此需要对每个实体词与句子进行注意力加权计算，举例如下。

句子："The<e1>burst</e1>has been caused by water hammer<e2>pressure</e2>."

关系的类型是：”Cause-Effect(e2,e1)”

在该句中，“caused”是背景词，其在确定关系“Cause-Effect”方面具有特殊意义。因此，可以利用两个实体词“burst”与“pressure”在句中存在的背景词“caused”来帮助确定其关系类型。本文引入了一种实体注意机制来定量衡量句中词与目标实体词之间的上下文相关性。

4.2 模型架构设计

本文提出一种新的网络结构并结合注意力机制的模型，该模型不仅能提取出基于词级别以及相邻词之间关系的句子特征；还能在不依赖外部知识的前提下，结合句中实体相关特征。分别利用 CNN-AttBiLSTM 网络模块、Att-BiLSTM 网络模块与实体注意力网络对句子进行编码，最终将三类特征进行线性加权组合完成关系分类任务。

如图 4.1 所示，本文提出的模型包含三个部分。

（1）输入层：首先进行数据预处理，然后将句子的词向量信息，位置向量信息拼接起来作为句子的输入，分别输入到网络层的 Att-BiLSTM 网络模块、CNN-AttBiLSTM 网络模块以及实体注意力网络模块中。

（2）网络层：分别通过 Att-BiLSTM 网络模块、CNN-AttBiLSTM 网络模块、实体注意力网络模块学习到基于词，相邻词间关系的句子特征以及句中实体相关特征。

（3）输出层：将三类特征进行线性加权组合，输入 Softmax 中进行关系分类。

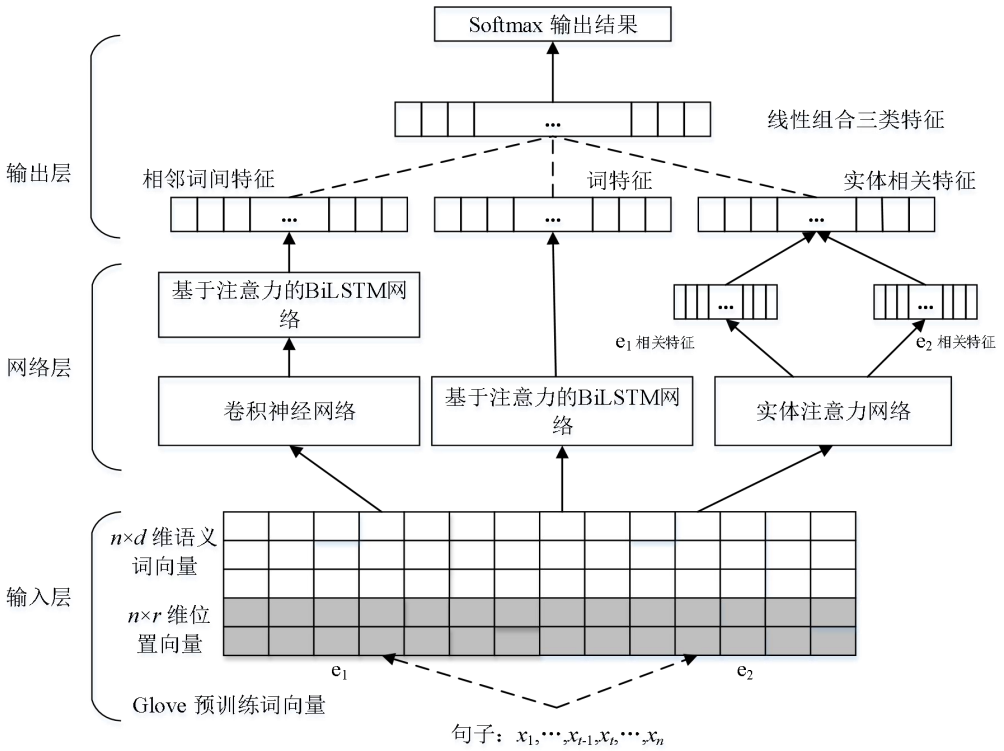

图 4.1　基于注意力机制的多特征融合的关系抽取模型

4.2.1 实体注意力网络

为了计算句子中每个单词的权重，本文将句子中的每个单词和每个实体输入到前馈神经网络中。假设每个句子都包含 T 个词。w_{it}表示第 i 个句子中的第 t 个单词($t \in [1,T]$）。$e_{ij}(j \in [1,2]$）表示第 i 个句中的第 j 个实体。将实体e_{ij}的表示与单词w_{it}的表示连接起来，得到单词 t 的新表示，即h_{it}^{j}=$[w_{it}, e_{ij}]$。h_{it}^{j}量化了第 i 个句子中第 t 个单词与第 j 个实体的相关程度。相关性由多层感知机根据单词w_{it}和实体e_{ij}各自词嵌入计算出来。本文将相关程度命名为词注意权重，即u_{it}^{j}。u_{it}^{j}的计算过程如下：

$$h_{it}^{j}=[w_{it}, e_{ij}] \tag{4.1}$$

$$u_{it}^{j} = W_a\left[\tanh\left(W_{we}h_{it}^{j} + b_{we}\right)\right] + b_a \tag{4.2}$$

注意力多层感知机网络的输出是u_{it}^{j}，然后通过 softmax 函数归一化重要性权重。

$$\alpha_{it}^{j} = \frac{\exp\ (u_{it}^{j})}{\sum_t \exp\ (u_{it}^{j})} \tag{4.3}$$

然后，我们通过以下公式计算句中实体相关性注意力权重，再进行加权求和作为句子的实体注意力向量s_{ij}：

$$s_{ij} = \sum_t \alpha_{it}^{j} w_{it} \tag{4.4}$$

在训练过程中，注意力权重在多层感知机网络被随机初始化。最后，可以分别得到两个实体的句子注意力向量。然后，我们将两个输出向量进行拼接输入到前馈神经网络中形成一个固定长度的特征向量作为最终的实体相关特征向量。

4.2.2 关系分类

本文提出一种关系抽取模型，以句子的词向量及位置向量作为输入，输入到网络层，分别学到词特征、相邻词间特征以及实体相关特征。最终，可以获得三种不同特征向量，即词向量、相邻词间向量、实体相关特征向量。

实体相关特征向量被拼接为$h_e = [h_{e1}, h_{e2}]$（h_{e1}, h_{e2}分别表示句中两个实体词对应的实体相关特征向量），通过全连接层将实体相关特征向量h_e转换成向量H_e。然后，将词特征向量H_1，相邻词间特征向量H_2和实体相关特征H_e线性加权求和作为关系分类的最终向量O，并将其输入 softmax 分类器以计算概率：

$$O = a * H_1 + b * H_2 + (1 - a - b) * H_e \tag{4.5}$$

$$p(y) = softmax(O) \tag{4.6}$$

对于每个句子，与最大概率值对应的类别就是分类结果。

$$\hat{y} = arg\max_y p(y) \tag{4.7}$$

目标函数是带L_2正则化的交叉熵损失函数，如式(12)所示。

$$J(\theta) = -\sum_{i=1}^{m} t_i \log(y_i) + \lambda \|\theta\|_F^2 \tag{4.8}$$

4.3 实验与分析

4.3.1 实验设置

本文实验部分参数设置如下表 4.1 所示。其中，词向量、位置向量以及学习率等参数跟上文实验保持一致。通过研究不同权重组合对实验结果的影响，最终得到权重 a 与 b 的最优值。

表 4.1 超参数设置

参数	值
词向量维度	100
位置向量维度	10
句子长度	100
批处理大小	10
学习率	1.0
L2 正则化系数	1.00×10^{-5}
词嵌入层 dropout	0.3
LSTM 层 dropout	0.3
其他 dropout	0.5
CNN 卷积核窗口大小	3
CNN 卷积核个数	110
LSTM 单元大小	100
a	0.8
b	0.1
随机种子	5782
epoch	100

4.3.2 实验结果及分析

在前一章研究的基础上，为了进一步验证实体有关信息对关系抽取任务的鲁棒性和有效性，依旧采用对比实验方法，同时在 SemEval-2010 Task8 数据集和 Wiki80 数据集上与 3 种经典主流模型 CNN、CRCNN、BiLSTM 以及上文提出的模型进行对比，且统一采用词向量和位置向量作为模型的输入，表 4.2 与表 4.3 分别表示 4 种模型在两种数据集上的 macro-F1 值结果。其中，Our_model2 是本文提出的结合实体相关信息的多特征组合的方法。

表 4.2　SemEval 实验结果对比

Models	Featureset	macro-F1/%
CNN	word,position,word pair,words around word pair	79.47
CR-CNN	word,position	80.24
BiLSTM	word,position	82.23
Ourmodel_1	word,position,Word and Adjacent-word	83.52
Ourmodel_2	word,position,Word and Adjacent-word,entity attention	84.17

表 4.3 Wiki80 实验结果对比

Models	Featureset	macro-F1/%
CNN	word,position,word pair,words around word pair	78.00
CR-CNN	word,position	78.64
BiLSTM	word,position	79.78
Ourmodel_1	word,position,Word and Adjacent-word	81.72
Ourmodel_2	word,position,Word and Adjacent-word,entity attention	82.08

由表中的实验结果可以发现，本文模型比 CNN、CR-CNN 以及 BiLSTM 模型的效果更好，在使用同样输入的情况下，首先结合 CNN 与 BiLSTM 的特性，抽取出句子中相邻词间特征，然后通过 Att-BiLSTM 学习到重要的基于词级别的句子特征。此外，对于句子级的关系抽取任务，实体信息与句子信息都至关重要，实体以及实体左右的上下文背景信息有利于确定实体在句中的语义关系。为了进一步挖掘实体相关信息，本模型采用了实体注意力机制挖掘句子中实体相关信息，在句子中挖掘出与两实体词相关的背景信息；最后，将基于词的句子特征、基于相邻词的句子特征与实体词相关的背景相邻信息共三种特种进行线性组合，最终达到了最优的效果，实验表明，在 SemEval 数据集上模型能到 84.17%的结果，在 Wiki80 数据集上达到 82.08%的结果。

下图 4.2、图 4.3 对应上表 4.2 及表 4.3 中 CNN、CRCNN、LSTM、Our_model1、Our_model2 五种模型分别在 SemEval-2010 Task8 数据集以及 Wiki80 数据集上 macro-F1 值随迭代次数的变化曲线。

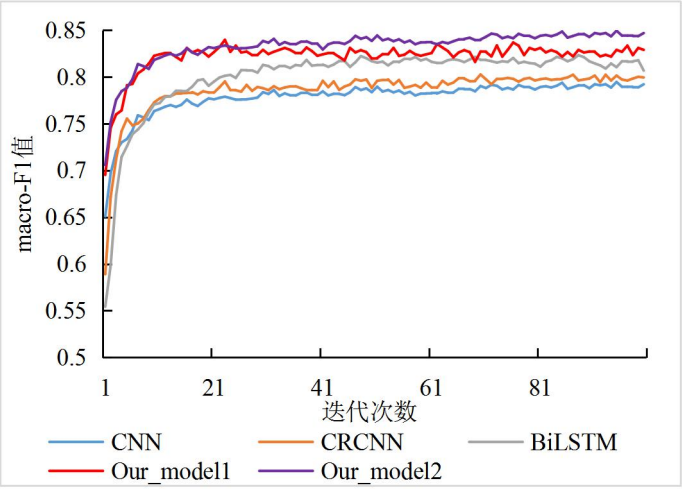

图 4.2　SemEval 数据集上 macro-F1 值随迭代次数变化曲线

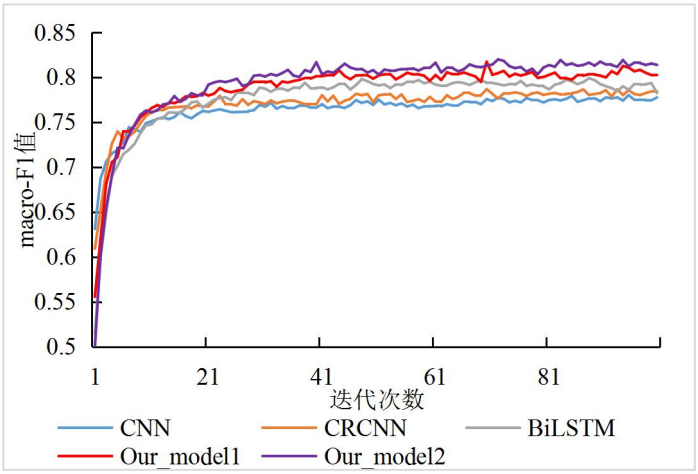

图 4.3　Wiki80 数据集上 macro-F1 值随迭代次数变化曲线

上图实验结果显示，本章提出的结合实体相关信息的多特征组合的关系抽取模型在两种数据集上的表现效果整体都优于对比模型，证明了模型的有效性与稳定性。与此同时，本章实验也统计了在两种数据集下各个关系分类下各模型准确率、召回率以及 F1 值的性能。图 4.4、图 4.5、图 4.6 为各模型在两数据集下各关系的准确率、召回率以及 F1 值曲线，可以发现，本文提出的结合实体相关信息的多特征组合的关系抽取模型在各类关系表现上均优于对比模型，整体上的表现效果更好。

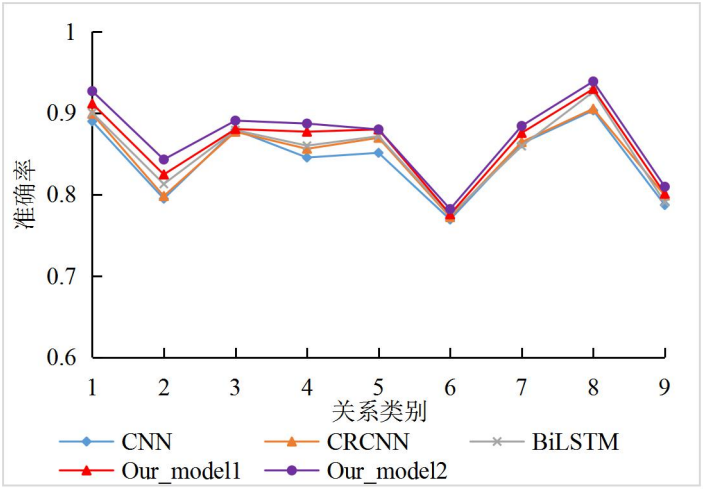

图 4.4　Sem 数据集上各关系准确率比较

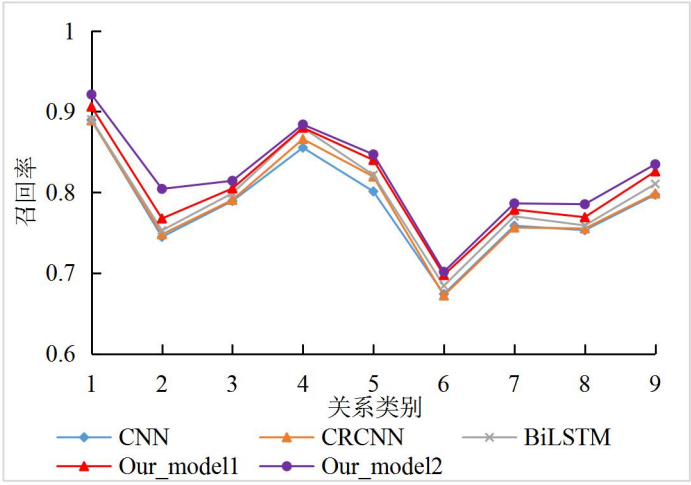

图 4.5　Sem 数据集上各关系召回率比较

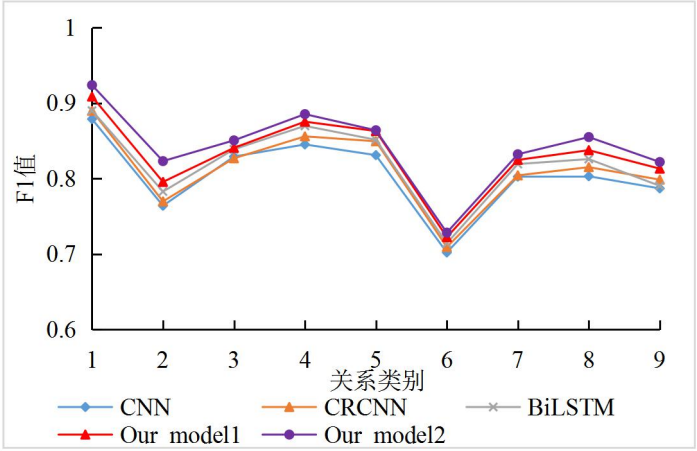

图 4.6　Sem 数据集上各关系 F1 值比较

在图 4.7、图 4.8、图 4.9 中，展示了 Wiki80 数据集前 20 种关系类别分别对应的准确率、召回率以及 F1 值曲线，进一步验证了本文提出模型的有效性。

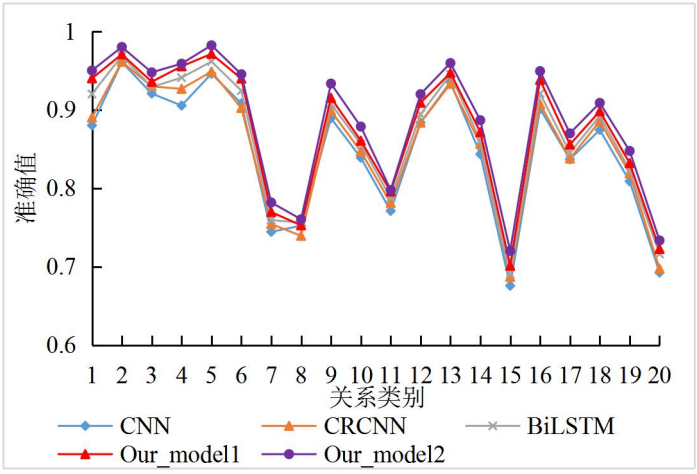

图 4.7　Wiki80 数据集上前 20 种关系类别准确率比较

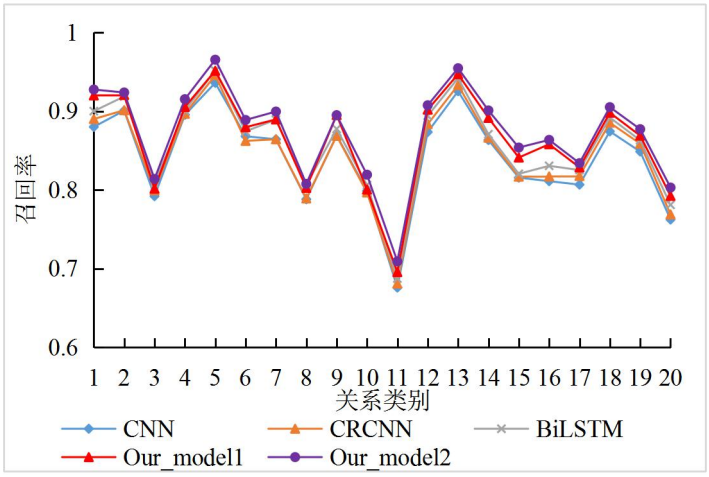

图 4.8　Wiki80 数据集上前 20 种关系类别的召回率比较

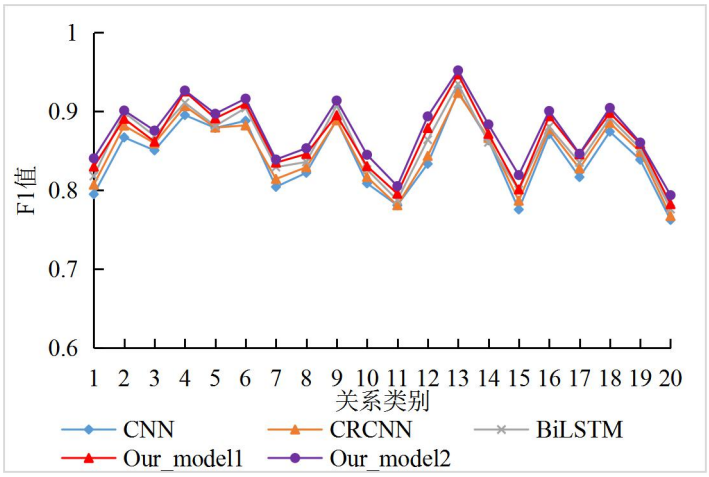

图 4.9　Wiki80 数据集上前 20 种关系类别的 F1 值比较

4.4 特征有效性验证

特征会直接影响模型的性能，为了验证本文涉及的各个特征的有效性，分别在两个数据集上做特征验证实验，表 4.4 以及表 4.5 展示了不同特征对 macro-F1 值的影响。

表 4.4　SemEval-2010 Task8 数据集上特征对结果的影响

Model	Feature set	macro-F1/%
Our_model	Entity attention	80.05
	Word Feature	82.23
	Adjacent-word Feature	81.00
	Word Feature and Adjacent-word Feature	83.52
	All above	84.17

表 4.5 Wiki80 数据集上特征对结果的影响

Model	Feature set	macro-F1/%
Our_model	Entity attention	78.23
	Word Feature	79.78
	Adjacent-word Feature	78.86
	Word Feature and Adjacent-word Feature	81.72
	All above	82.08

以 SemEval-2010 Task8 数据集上特征对结果的影响为例进行说明。实验结果表明，直接用本文模型中的实体注意力网络部分抽取到的实体相关特征进行关系抽取任务，macro-F1 为 80.05%，同样的，单独用词特征进行关系分类得到 82.23%的效果，只采用相邻词间关系特征能达到 81%的效果。将词特征与相邻词间特征两种特征拼接组合进行关系分类任务，macro-F1 能达到 83.52%，优于每个单特征的抽取效果。在此基础之上，将实体相关特征加入到模型中，对三种特征信息进行线性加权求和能达到 84.17%的结果。实验结果验证了本文提出的相邻词间特征以及实体相关特征对关系抽取任务的有效性。

4.5 多特征组合对结果的影响

为了进一步研究多特征组合的方式对实验效果的影响，本小节基于上一节对特征有效性研究的基础上，对三种重要特征按其效果赋予不同权重进行组合实验，实验结果如下表 4.6 及 4.7 所示：

表 4.6　SemEval-2010 Task8 数据集上特征组合对结果的影响

词特征(H_1)权重系数	相邻词特(H_2)权重系数	实体相关特征(H_e)权重系数	最终特征组合形式	macro-F1/%
0.5	0.4	0.1	$0.5*H_1+0.4*H_2+0.1*H_e$	82.23
0.6	0.3	0.1	$0.6*H_1+0.3*H_2+0.1*H_e$	82.49
0.7	0.2	0.1	$0.7*H_1+0.2*H_2+0.1*H_e$	82.56
0.8	0.1	0.1	$0.8*H_1+0.1*H_2+0.1*H_e$	84.17
0.9	0.05	0.05	$0.9*H_1+0.05*H_2+0.05*H_e$	81.58
1	1	1	$H_1+H_2+H_e$	81.65
向量拼接$[H_1, H_2, H_e]$				81.62

表 4.7 Wiki80 数据集上特征组合对结果的影响

词特征(H_1)权重系数	相邻词特征(H_2)权重系数	实体相关特征(H_e)权重系数	最终特征组合形式	macro-F1/%
0.5	0.4	0.1	$0.5*H_1+0.4*H_2+0.1*H_e$	80.50
0.6	0.3	0.1	$0.6*H_1+0.3*H_2+0.1*H_e$	80.61
0.7	0.2	0.1	$0.7*H_1+0.2*H_2+0.1*H_e$	81.56
0.8	0.1	0.1	$0.8*H_1+0.1*H_2+0.1*H_e$	82.08
0.9	0.05	0.05	$0.9*H_1+0.05*H_2+0.05*H_e$	80.58
1	1	1	$H_1+H_2+H_e$	79.65
向量拼接$[H_1, H_2, H_e]$				80.60

表 4.6 跟表 4.7 通过不同的组合方式验证了特征有效性的同时也说明了不同的特征组合方式会对实验结果产生不同的影响。不同的特征对关系抽取效果产生不同程度的影响，根据特征在关系抽取任务上的影响效果不同，对三类特征赋予不同的权重进行线性组合，从上表 4.4 及表 4.5 也可以看出，词特征的对关系分类的影响程度最大，故而本文以词特征为主对不同特征进行线性加权，对词特征赋予最高的权重系数，同时融入其他特征丰富并加强对特征的表达。实验结果表明，直接将三类特征向量进行拼接组合，并将其作为最终用于关系分类的特征，实验在两种数据集上的结果分别为 81.62%、80.60%的结果；对不同的特征进行线性组合，当词特征的权重系数设置为 0.8，相邻词特征权重系数为 0.1，实体相关特征权重系数为 0.1 时，其线性组合的最终特征在关系抽取任务上取得最好的效果。

4.6 本章小结

本章首先介绍了结合实体相关信息的多特征组合的关系抽取模型的研究动机和模型架构；然后对实体注意力网络模块进行了描述，并对关系分类层特征的组合进行了详细地说明；通过在公用数据集上进行验证，以宏平均作为性能评测指标与其它经典、主流的模型以及第三章的模型进行对比。实验显示，结合实体相关信息的多特征组合的关系抽取模型可以挖掘并利用句中实体词上下文背景信息、词特征、相邻词间关系特征，因此相比其他模型取得更好的结果。同时本章验证了实体词所在的上下文背景信息能进一步提升模型的效果，也证明了研究实体及其背景信息对于关系抽取任务而言是有意义的，表明了句中实体词的上下文背景信息对该任务研究的重要性。

第五章　总结与展望

5.1 论文总结

关系抽取属于 NLP 领域中信息抽取里面的一个重要且基础的任务，它能为很多任务提供支持，如知识库构建、智能问答系统、智能决策系统、智能推荐系统等等。怎样在自由文本里自动抽取出实体间的存在的关系这一问题已经得到了广泛的研究。万维网的快速发展导致每天都会产生数以亿计的海量文本数据，以人工方式处理这些海量的数据变得不现实，在这种背景下，关系自动抽取自然便成了具有挑战性的研究领域。

本文针对目前基于深度学习的二元关系抽取研究中存在的问题：依赖外部知识库特征、模型对于句子信息挖掘不充分、对特征的优化跟组合相对缺乏更进一步的研究，针对这些问题，本文提出了相应的解决方法，主要研究内容和贡献如下：

（1）本文提出融合词特征与相邻词间特征的关系抽取模型，该模型能够有效利用卷积神经网络、双向长短期记忆网络以及注意力机制的各自的优势提取词特征、相邻词间特征，进一步挖掘出自然语言文本中潜在的语义信息。实验结果显示，本模型在 SemEval-2010 Task8 数据集以及 Wiki80 数据集上进行训练和测试，测试结果相比于其它 3 个模型均有所提高。

（2）提出结合实体相关信息的多特征组合的关系抽取模型：研究表明，实体信息对关系抽取任务起到非常重要的作用，为了使模型能够更好地利用到实体信息，学习到句中与实体相关的重要上下文向量信息，本文使用实体注意力对句子进行加权处理，生成实体注意力向量特征，强化模型对句子重要信息的理解，同时减小句子中无关信息的噪音。

（3）本文对各类特征的有效性单独进行验证跟分析，并在模型中将学到的各类特征进行优化组合实验，对特征进行深入研究。

5.2 展望

未来的工作可以从以下两个方面展开：

(1) 在句子输入层，除了词特征以及位置特征之外，还有词性、依存句法、语义角色特征、实体类别和语法关系等特征，将来可以尝试寻找更多未使用过的关键特征，也可以研究引入外部资源特征去丰富输入层的信息表示。

(2) 对模型的探索还有广阔的空间。机器学习的大部分工作都在做数据处理跟特征处理，特征能决定模型的上限，如果模型能学到更多更有用的特征无疑会提升关系抽取的效果。比如图卷积神经网络近些年来在很多 NLP 领域上都有着不错的效果，在关系抽取任务上未来可以探索新的网络模型。

参考文献

[1] 王昊奋,漆桂林,陈华钧.《知识图谱：方法、实践与应用》[J].自动化博览,2020,37(01):7.

[2] Pawar S, Palshikar G K, Bhattacharyya P. Relation extraction: A survey[J]. arXiv preprint arXiv:1712.05191, 2017.

[3] 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]. Journal of Software, 2019, 30(6).

[4] 庄传志, 靳小龙, 朱伟建, 等. 基于深度学习的关系抽取研究综述[J]. 中文信息学报, 2019, 33(12): 1-18.

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

[6] 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.

[7] Kumar S. A survey of deep learning methods for relation extraction[J]. arXiv preprint arXiv:1705.03645, 2017.

[8] 李冬梅, 张扬, 李东远, 等. 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.

[9] 谢德鹏,常青.关系抽取综述[J].计算机应用研究,2020,37(07):1921-1924+1930.

[10] 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 (ACL-04). 2004: 415-422.

[11] Chen J, Ji D, Tan C L, et al. Unsupervised feature selection for relation extraction[C]//Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts. 2005.

[12] Yan Y, Okazaki N, Matsuo Y, et al. Unsupervised relation extraction by mining wikipedia texts using information from the web[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009: 1021-1029.

[13] Poon H, Domingos P. Unsupervised semantic parsing[C]//Proceedings of the 2009 conference on empirical methods in natural language processing. 2009: 1-10.

[14] Yao L, Haghighi A, Riedel S, et al. Structured relation discovery using generative models[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011: 1456-1466.

[15] 黄晨,钱龙华,周国栋, 等.基于卷积树核的无指导中文实体关系抽取研究[J].中文信息学报,2010,24(04):11-17.

[16] 马超.基于 Web 信息使用改进的无监督关系抽取方法构建交通本体[J].计算机系统应用,2015,24(12):273-276.

[17] Brin S. Extracting patterns and relations from the world wide web[C]//International workshop on the world wide web and databases. Springer, Berlin, Heidelberg, 1998: 172-183.

[18] Agichtein E, Gravano L. Extracting relations from large plain-text collections[J]. 1999.

[19] 陈锦秀,姬东鸿.基于图的半监督关系抽取[J].软件学报,2008(11):2843-2852.

[20] Cvitaš A. Relation extraction from text documents[C]//2011 Proceedings of the 34th International Convention MIPRO. IEEE, 2011: 1565-1570.

[21] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 2009: 1003-1011.

[22] Zeng D, Liu K, Chen Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 1753-1762.

[23] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 2124-2133.

[24] Feng J, Huang M, Zhao L, et al. Reinforcement learning for relation classification from noisy data[C]//Proceedings of the aaai conference on artificial intelligence. 2018, 32(1).

[25] Qin P, Xu W, Wang W Y. Dsgan: Generative adversarial training for distant supervision relation extraction[J]. arXiv preprint arXiv:1805.09929, 2018.

[26] Ren X, Wu Z, He W, et al. Cotype: Joint extraction of typed entities and relations with knowledge bases[C]//Proceedings of the 26th International Conference on World Wide Web. 2017: 1015-1024.

[27] 黄杨琛,贾焰,甘亮, 等.基于远程监督的多因子人物关系抽取模型[J].通信学报,2018,39(07):103-112.

[28] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction[C]//Proceedings of the ACL Interactive Poster and Demonstration Sessions. 2004: 178-181.

[29] Giuliano C, Lavelli A, Pighin D, et al. FBK-IRST: Kernel methods for semantic relation extraction[C]//Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). 2007: 141-144.

[30] Tratz S, Hovy E. Isi: automatic classification of relations between nominals using a maximum entropy classifier[C]//Proceedings of the 5th International Workshop on Semantic Evaluation. 2010: 222-225.

[31] Culotta A, McCallum A, Betz J. Integrating probabilistic extraction models and data mining to discover relations and patterns in text[C]//Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. 2006: 296-303.

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

[33] Zhao S, Grishman R. Extracting relations with integrated information using kernel methods[C]//Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05). 2005: 419-426.

[34] Culotta A, Sorensen J. Dependency tree kernels for relation extraction[C]//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). 2004: 423-429.

[35] Bunescu R, Mooney R. A shortest path dependency kernel for relation extraction[C]//Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. 2005: 724-731.

[36] Zhang M, Zhang J, Su J. Exploring syntactic features for relation extraction using a convolution tree kernel[C]//Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. 2006: 288-295.

[37] 庄成龙, 钱龙华, 周国栋. 基于树核函数的实体语义关系抽取方法研究[J]. 中文信息学报, 2009, 23(1): 3.

[38] 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.

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

[40] Nguyen T H, Grishman R. Relation extraction: Perspective from convolutional neural networks[C]//Proceedings of the 1st workshop on vector space modeling for natural language processing. 2015: 39-48.

[41] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 2124-2133.

[42] Santos C N, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks[J]. arXiv preprint arXiv:1504.06580, 2015.

[43] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th annual meeting of the association for

computational linguistics (volume 2: Short papers). 2016: 207-212.

[44] 闫雄,段跃兴,张泽华.采用自注意力机制和 CNN 融合的实体关系抽取[J].计算机工程与科学,2020,42(11):2059-2066.

[45] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[46] Guo Z, Zhang Y, Lu W. Attention guided graph convolutional networks for relation extraction[J]. arXiv preprint arXiv:1906.07510, 2019.

[47] Paccanaro A, Hinton G E. Learning distributed representations of concepts using linear relational embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(2): 232-244.

[48] 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 (EMNLP). 2014: 1532-1543.

[49] Hendrickx I, Kim S N, Kozareva Z, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals[J]. arXiv preprint arXiv:1911.10422, 2019.

[50] Han X, Gao T, Yao Y, et al. OpenNRE: An open and extensible toolkit for neural relation extraction[J]. arXiv preprint arXiv:1909.13078, 2019.

[51] 周飞燕,金林鹏,董军.卷积神经网络研究综述[J].计算机学报,2017,40(06):1229-1251.

[52] 杨丽,吴雨茜,王俊丽, 等.循环神经网络研究综述[J].计算机应用,2018,38(S2):1-6+26.

[53] 石磊,王毅,成颖, 等.自然语言处理中的注意力机制研究综述[J].数据分析与知识发现,2020,4(05):1-14.

[54] 朱张莉,饶元,吴渊, 等.注意力机制在深度学习中的研究进展[J].中文信息学报,2019,33(06):1-11.

攻读硕士学位期间发表的论文和科研成果

1. 学术论文

[1] 赵占芳,刘鹏鹏,李雪山.基于改进 TextRank 的铁路文献关键词抽取算法[J].北京交通大学学报,2021,45(02):80-86.

[2] 刘鹏鹏,赵占芳,王楠. 基于标记属性图的 Wikidata 人物关系可视化数据分析[J]. 新一代信息技术,2021,4(12):13-18. DOI:10.3969/j.issn.2096-6091.2021.12.003.

2. 获奖成果

[1] 2019 年 10 月　河北地质大学学业奖学金三等奖

[2] 2020 年 10 月　河北地质大学学业奖学金三等奖

3. 参与项目

[1] 河北省社会科学基金项目“网络叙词表的语义表示及知识服务机制研究”(项目编号：HB20TQ003)

[2] 中国铁道科学研究院基金项目（2018YJ134）：中国工程院项目（CKCEST2019-2-11）

作者简介

刘鹏鹏，男，汉族，河南省信阳市人。

1995 年 10 月 15 日出生于河南省信阳市。

2015 年 09 月考入南阳理工学院软件学院软件设计专业。

2019 年 06 月本科毕业于南阳理工学院，获得工学学士学位。

2019 年 09 月考入河北地质大学攻读计算机应用技术专业硕士研究生，师从亢俊健教授，赵占芳教授。

研究生期间完成课程数 16 门，共修学分 34 分，达到了本专业研究生培养方案。

致　　谢

流光容易把人抛，红了樱桃，绿了芭蕉，两年半的研究生生涯就要宣告结束，又到了毕业季，又到了告别分开的时候，值此之际，心里万分留恋不舍。回想这段时光，心里尽是不舍跟感谢。在此，我要衷心的感谢在这段学习生活中给予我指导和帮助的老师、同学、家人以及朋友们。

首先要特别地深深感谢我的导师亢俊健教授和赵占芳教授，两位导师不仅是我的学术导师更是我的人生导师。亢老师知识渊博、为人谦和、对学生的体贴、关心，每次跟老师沟通学习，都给人一种如沐春风、醍醐灌顶的感觉，让我明白任何时候都不要荒废自己、都不能浪费时间，不畏困难挑战，从亢老师身上我学会了一些做人做事的道理，这些都深深刻在我的脑子里指引着我，是我的精神原动力。赵老师是我另外一个要特别感谢的人，刚认识赵老师的时候，她严谨的学术态度、精益求精的工作作风、认真务实的实干精神深深触动了我。这两年多来，我跟着赵老师学到了很多东西，有很多很多让我难忘的记忆，引领我做科研，教会我道理，给我启迪、提高我的认知。两位老师的言行教育使我刻骨铭心、终生难忘。此外，我要感谢实验室的其他老师，柴变芳老师、尹立杰老师、马立肖老师、陈嶷瑛等老师，在这两年多的时间里，我们一起学习交流，感谢你们无私的帮助、关心与照顾。同时要感谢实验室的兄弟姐妹们，我们经常探讨学术，感谢学长学姐们的指教，也感谢学弟学妹们的热情。

最后，我特别要感谢我的家人，你们给我支撑，感谢你们无私奉献跟默默支持，让我顺利完成学业，你们是我人生前中的最坚强最坚定的后盾，永远感谢家人。我还要感谢一个人，杨蕾，我们一起成长，一起度过了难忘美好的时光，感谢你对我的帮助，祝我们未来可期，明天更好。

感谢审稿专家、答辩委员会以及各位老师对我提出的宝贵意见和建议。

最后，再次衷心感谢所有给予我帮助和关心的老师、同学以及朋友们！