MASTER THESIS XIDIAN UNIVERSITY

西安电子科技大学

硕士学位论文

深度关系抽取算法的研究与应用

作者姓名 杨文成

学校导师姓名、职称 李雁妮 教授

企业导师姓名、职称 刘志鹏 高工

申请学位类别 工程硕士

学校代码 10701 学 号 18031211510
分 类 号 TP39 密 级 公开

西安电子科技大学

硕士学位论文

深度关系抽取算法的研究与应用

作者姓名：杨文成

领　　域：计算机技术

学位类别：工程硕士

学校导师姓名、职称：李雁妮教授

企业导师姓名、职称：刘志鹏 高工

学　　院：计算机科学与技术学院

提交日期：2021 年 6 月

Research on and Application of Relation Extraction Algorithms based on Deep Learning

A thesis submitted to

XIDIAN UNIVERSITY

in partial fulfillment of the requirements

for the degree of Master

in Engineering

By

Wencheng Yang

Supervisor: Yanni Li Title: Professor

Supervisor: Zhipeng Liu Title: Senior Engineer

June 2021

西安电子科技大学
学位论文独创性（或创新性）声明

秉承学校严谨的学风和优良的科学道德，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中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同事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。

学位论文若有不实之处，本人承担一切法律责任。

本人签名：杨文成　　日　　期：2021.6.1

西安电子科技大学
关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权属于西安电子科技大学。学校有权保留送交论文的复印件，允许查阅、借阅论文；学校可以公布论文的全部或部分内容，允许采用影印、缩印或其它复制手段保存论文。同时本人保证，结合学位论文研究成果完成的论文、发明专利等成果，署名单位为西安电子科技大学。

保密的学位论文在___年解密后适用本授权书。

本人签名：杨文成　　导师签名：李雅娇

日　　期：2021.6.1　　日　　期：2021.6.1

摘要

关系抽取是一种从文本中提取一对实体间的语义关系的任务。作为信息抽取的重要组成部分，关系抽取为构建机器可以识别的结构化数据提供了便利。**深度关系抽取算法**是指使用深度学习方法解决关系抽取问题的算法，可以对关系抽取问题进行端对端处理。**有监督关系抽取算法**充分利用人工标注的标签信息来训练模型，是目前关系抽取的主流方法，但是标注数据所耗费的成本过高。**远程监督关系抽取算法**通过将知识库与非结构化文本对齐来自动构建大量训练数据，减少模型对人工标注数据的依赖。

近年来，国内外专家学者使用深度学习方法在有监督关系抽取与远程监督关系抽取领域进行了深入研究，并提出了一些较为有效的算法，但这些算法基本面临着如下问题：1) 有监督关系抽取算法大都选择从单个句子中学习实体对的特征，该方式学习到的片面特征会影响到关系抽取的精度；2) 远程监督关系抽取算法采用远程对齐的方式构建训练数据会引入大量噪声，这些噪声会极大地干扰关系抽取的结果；3) 无论是有监督关系抽取还是远程监督关系抽取，其提取实体和句子特征的精度都有待提高。

为克服现有的关系抽取算法的上述缺陷，作者对国内外关系抽取算法进行了深入研究，并分别提出了一种高效的有监督关系抽取算法，以及一种高效的远程监督关系抽取算法，本文的主要工作与创新如下：

1) 本文针对现有有监督关系抽取算法忽略外部知识的问题，建立了一种可以存储语料库知识的实体关系图，并为此设计了一种基于语义相似度的图神经网络来有效筛选知识以丰富对实体的认识，在此基础上提出了一种新的基于实体关系图的有监督关系抽取算法 ERGSRE(A Novel **E**ntity **R**elation **G**raph for **S**upervised **R**elation **E**xtraction)；
2) 针对远程监督关系抽取特征提取不完全，噪声识别精度不足，难以充分利用噪声数据等缺陷，引入图神经网络丰富句子结构化信息，并设计了一种基于边距度量的噪声识别方式识别数据中的噪声，此外将噪声数据作为负样本与正确标注的正样本一起训练算法以提升关系抽取的精度，据此提出了一种基于边距度量的远程监督关系抽取算法 MMDSRE(A **M**argin **M**etric for **D**istantly **S**upervised **R**elation **E**xtraction)；
3) 在真实数据集上对本文所提出的 ERGSRE 算法和 MMDSRE 算法进行了仿真实验，实验结果表明，ERGSRE 算法和 MMDSRE 算法的性能均高于当前最新最好的关系抽取算法。此外将本文提出的两个算法集成至开放式关系抽取工具包 OpenNRE，可以方便地一键式抽取实体对的关系，同时也提升了

OpenNRE 工具的性能。

尽管本文提出的两种算法可以分别提升有监督关系抽取和远程监督关系抽取算法的性能，但是这两个算法对于训练数据的依赖过高，当算法加入新的训练任务后，很容易忘记之前的训练任务，而终身学习机制可以有效解决这一现象，因此如何引入终身学习机制以提升模型的抗遗忘能力是作者未来的研究方向。

关 键 词：关系抽取，深度学习，有监督学习，远程监督学习

ABSTRACT

Relation extraction is a way of extracting semantic relations between entities from text. As an important part of information extraction, relation extraction provides convenience for constructing structured data that can be recognized by machines. Relation extraction algorithms based on deep learning refer to the algorithms that use deep learning method to solve the problem of relation extraction. The supervised relation extraction algorithms make full use of manually labeled information to train the model. It is the mainstream method of relation extraction at present, but the cost of labeling data is too high. The distantly supervised relation extraction algorithm automatically builds a large amount of training data by aligning the knowledge base with unstructured text, reducing the model's dependence on artificially labeled data.

In recent years, countless experts and scholars have used deep learning methods to conduct deep research in the fields of supervised relation extraction and distantly supervised relation extraction, and have proposed some effective relation extraction algorithms, but these algorithms basically face the following problems: 1)The supervised relation extraction algorithms mostly choose to the features of entities from a single sentence, the one-sided features learned in this way will affect the accuracy of relation extraction. 2)Because of the data from distantly supervised relation extraction is constructed by aligning the knowledge base with the corpus, there is a lot of noise in the data, which will greatly interfere with the results of relation extraction. 3)Whether it is supervised relation extraction or distantly supervised relation extraction, the accuracy of extracting entity and sentence features needs to be improved.

The topic of the thesis comes from the general project of the National Natural Science Foundation of China. In order to overcome the shortcomings of the existing relation extraction algorithms, the author has conducted deep research on the relation extraction algorithms at home and abroad, and respectively proposed an efficient supervised relation extraction algorithm and an efficient distantly supervised relation extraction algorithm, the main work of this paper is as follows.

1) Aiming at the problem that existing supervised relation extraction algorithms ignore external knowledge, an entity semantic relationship graph that can represent the

connection relationship between different entities is established, which can store the knowledge of the corpus, and for this purpose, a graph neural network based on semantic similarity is designed to effectively filter knowledge to enrich the understanding of entities. On this basis, A Novel **E**ntity **R**elation **G**raph for **S**upervised **R**elation **E**xtraction (ERGSRE) algorithm is proposed.

2) In view of the defects such as incomplete feature extraction, low accuracy of noise identification and difficult to make full use of noise data, etc., a graph neural network is used to enrich sentence structure information, and a noise recognition method based on margin metric is designed to identify the noise in the data. In addition, we use the incorrectly labeled samples with correctly labeled samples to train the algorithm to improve the accuracy of relation extraction. Based on this, A **M**argin **M**etrics Denoising Method for **D**istantly **S**upervised **R**elation **E**xtraction algorithm (MMDSRE) is proposed.
3) The ERGSRE algorithm and the MMDSRE algorithm proposed in this article are simulated on real datasets. The test results show that the performance of the ERGSRE algorithm and the MMDSRE algorithm is higher than the current state-of-the-art relation extraction algorithms. The two algorithms proposed in this paper are integrated into OpenNRE, an open relation extraction toolkit, which can conveniently extract the relationship of entity pairs with one click, and meanwhile improve the performance of OpenNRE.

Although the two algorithms proposed in this article can respectively improve the performance of supervised relation extraction and distantly supervised relation extraction algorithms, this deep learning-based approach relies too much on training data, and it is easy to forget the old task after learning a new task, so it is difficult to meet the actual demand. While the lifelong learning mechanism can effectively solve this phenomenon. Therefore, how to introduce the lifelong learning mechanism to improve the anti-forgetting ability of the model is the future research direction of the author.

Keywords: Relation extraction, deep learning, supervised learning, distantly supervised learning

插图索引

表格索引

符号对照表

符号	符号名称
e	实体
s	句子
w_i	句子中的第 i 个单词
r	实体间的关系
R	所有的实体关系集合
G	图
V	图中的顶点集合
E	图中的边集合
A	图中顶点的邻接矩阵
D	图中顶点的度矩阵
W	参数矩阵
d^w	单词的特征向量维度
h_s	句子的特征向量
h_e	实体的特征向量
L	损失函数
D_{train}	训练数据集
D_{val}	验证数据集
D_{test}	测试数据集
$\oplus$	拼接操作
[*CLS*]	ALBERT 输入中的句子开始标记
[*SEP*]	ALBERT 输入中的句子结束标记
α	ERG 中句子间的影响权重
σ	激活函数
tanh()	tanh 激活函数
*F*1-*score*	模型的 F1 衡量指标
P@N	模型的 P@N 衡量指标
recall	模型预测结果的召回率
precision	模型预测结果的准确率

缩略语对照表

缩略语	英文全称	中文对照
CNN	Convolutional Neural Network	卷积神经网络
ERG	Entity Relation Graph	实体关系图
GCN	Graph Convolutional Neural Network	图卷积神经网络
GNN	Graph Neural Network	图神经网络
LSTM	Long Short-Term Memory	长短期记忆网络
MIL	Multiple Instance Learning	多实例学习
NA	Not Applicable	无可用的
NER	Named Entity Recognition	命名实体识别
NLP	Natural Language Processing	自然语言处理
NSP	Next Sentence Prediction	下一个句子预测
NYT	New York Times	纽约时报
PCNN	Piece Wise Convolutional Neural Network	分段卷积神经网络
P-R	Precision-Recall	准确率-召回率
RE	Relation Extraction	关系抽取
RNN	Recurrent Neural Network	循环神经网络

目录

第一章 绪论

本章首先对论文选题背景及意义进行了介绍，之后，对国内外研究现状进行了分析，最后，对本文的主要工作、创新点及论文组织结构进行了介绍。

1.1 选题背景与研究意义

信息抽取是自然语言处理的重要任务之一，其目的是从非结构化文本中提取结构化信息，这些结构化信息可以很容易地被人类或计算机程序使用。随着互联网的发展，人们在互联网上创造和分享了大量内容，这些大量的文本形式数据几乎不可能通过人工方式进行分析。因此，建立一个高效的信息抽取系统是很必要的，它可以从文本中自动提取有意义的数据，建立结构化知识库，用于自动问答[1]、机器翻译[2]、知识图谱[3]等应用。结构化知识库主要由实体以及实体之间的关系组成，如何精确地识别文本数据中的实体以及实体间的关系成为了构建高效的信息抽取系统的关键。所谓**关系抽取**是指一种从非结构化文本中提取实体间的语义关系的任务。它是信息抽取的重要组成部分，因此，一个精确的关系抽取算法对构建高效的信息抽取系统至关重要。

在过去几十年中，大量的关系抽取算法被提出，根据训练过程中对样本标签的使用情况，关系抽取算法被分为四大类：无监督方式[4]、半监督方式[5]、有监督方式[6]以及远程监督方式[7]。其中，所谓**有监督关系抽取算法**是指充分利用人工标注的标签信息来训练模型，并利用训练好的模型预测实体关系的关系抽取算法，该类算法可以取得较好的关系抽取效果，是关系抽取的主流方法。但是由于有监督的关系抽取方法需要大量的人工标记数据，要花费大量的时间与精力来标记数据集，代价较高。**无监督关系抽取**是根据命名实体之间的上下文词的相似性对命名实体进行聚类，虽然无监督关系抽取算法不需要标注数据，可以使用大量的数据，也可以提取大量的关系。但是，由于没有标准形式的关系，输出结果可能不容易映射到特定知识库所需的关系。**半监督关系抽取**利用模式和关系集之间的对偶性，从少部分有标签数据开始迭代地从语料库匹配更多相关句子来抽取关系，但是由于该类方法通常会遇到语义漂移和精度差的问题，因此半监督学习方法难以在关系抽取领域被广泛应用。为此，Mintz[8]提出了远程监督关系抽取算法，他提出了一个假设：如果一个句子包含一个关系中涉及的实体对，则该句子将描述该实体对的关系。所谓**远程监督关系抽取算法**是指通过将知识库与非结构化文本对齐来自动构建大量训练数据，并用这些自动构建的数据进行关系抽取的算法。这类算法减少了模型对人工标注数据的依赖，近年来，基于远程监督方式的关系抽取算法受到越来越多的人的关注。

传统的非深度学习的关系抽取方法通常通过有监督方式进行。它们可以分为两类:基于特征的方法[9]和基于内核的方法[10]。在这两种方法中，使用已有的自然语言处理系统进行特征提取和内核设计，导致下游各模块的误差积累。**深度关系抽取算法**是指使用深度学习方法解决关系抽取问题的算法，它可以对关系抽取问题进行端对端处理。随着深度学习技术的快速发展，基于深度学习的方法[6, 7, 11-15, 17-22, 35-45]可以有效消除手工构造特征的需求，因此深度关系抽取算法逐渐成为关系抽取的主流方法。经典的深度关系抽取算法可概括为：基于卷积神经网络(CNN)的方法[11, 30]，基于循环神经网络(RNN)的方法[12, 13]以及基于预训练模型的方法[14, 15, 35]。这些方法虽然取得了一定效果，但是受深度学习模型的限制，一些模型无法理解上下文相关性，从而导致重要的隐含信息丢失，严重影响关系抽取的准确性。

图是一种为一组对象(节点)及其关系(边)建模的数据结构。由于图具有较强的表达能力，基于图的深度学习方法已经受到越来越多的关注。图神经网络(GNN)[16]是一种在图领域上进行图计算的深度学习的方法，近年来，由于其高效的性能和较高的可解释性，GNN已成为广泛应用的图分析方法。基于图神经网络的关系抽取方法[17-20, 37, 38]可以有效弥补基于CNN、RNN和预训练模型的方法的缺陷。该方法可以从图结构中获取语句的拓扑信息或引入句子以外的知识来丰富语义信息，从而提升关系抽取的准确性。

本文拟对有监督关系抽取算法与远程监督关系抽取算法进行深入研究，具有重要的理论意义与应用价值。

1.2 国内外研究现状

近几十年来，国内外无数学者和科技工作者致力于对关系抽取算法的研究，同时也随着深度学习技术的飞速发展，越来越多的关系抽取算法被提出。回顾已有的关系抽取算法的发展和演变过程，跟据训练过程中是对样本中标签信息的使用方式，可将现有的关系抽取算法分为四大类：无监督方式[4, 33]、半监督方式[5, 34]、有监督方式[6, 9-15, 17-20]以及远程监督方式[7, 30, 31, 38-45]。但是由于基于无监督和半监督方式的关系抽取算法所抽取的关系精度较低，且难以与现实知识库中的关系对齐，实用意义较低，因此本文将重点研究基于有监督和远程监督的关系抽取算法。

1.2.1 有监督关系抽取算法

有监督关系抽取算法通过大量的标注了实体对及其关系的样本数据训练模型，然后利用训练好的模型预测实体对的关系，因此在有监督方法中，关系抽取被视为分类任务。基于有监督的关系抽取算法可以简单地分为三种类型：基于特征向量的方法[9,

[21, 22]，基于内核的方法[10]以及基于神经网络的方法[6, 11-15, 17-20]。

基于特征向量的方法需要从包含实体对的语句中提取语义特征，并转换为特征向量，然后使用条件随机场[21]，支持向量机[22]等模型进行关系抽取。该方法的关键是在将结构化表示转换为特征向量时如何选择合适的特征集。但是，特征的选择通常需要凭借直觉以及实验得出，不稳定因素过多，在分类效果上很难进一步提升。基于核函数的关系抽取算法[10]避免了特征向量方法中的显式的特征工程，在基于核函数的方法中，利用核函数计算两个关系实例之间的相似度，并使用分类器进行分类。这一方法的重点是如何设置能够准确识别不同关系实例之间相似度的核函数，这一工作需要大量的人工操作，在大规模语料的关系抽取任务上有较大局限性。

以上所有方法的性能在很大程度上取决于从现有自然语言处理（NLP）工具派生的抽取功能的质量。这些工具在处理过程中不可避免地会产生错误。因此，如何通过减少对现有 NLP 工具的使用来提取特征成为研究的重点。神经网络（NN）的兴起为此类问题提供了新的方式。Socher 等人[23]首先将神经网络应用于关系抽取。从那时起，许多神经网络（例如卷积神经网络(CNN)，循环神经网络(RNN)，预训练模型以及图神经网络(GNN)等）被广泛用于关系抽取任务。Zeng 等人[11]使用卷积神经网络（CNN）捕获每个单词周围的文本信息来学习句子及实体的特征, Xu[12]和 Zhou[13]等人采用长短期记忆（LSTM）网络捕获句子中的语义信息, 这些方法受模型限制，特征提取能力有限。近年来，由 Devlin 等人提出的预训练模型 BERT[24]可以有效地改善许多 NLP 任务。Wu[14]和 Soares[15]等人将 BERT 引入他们的模型中以学习上下文信息。这类方法凭借 BERT 高效的特征提取能力在关系抽取领域取得了巨大突破。由于图神经网络(GNN)[16]具有较强的信息表达能力，因此被广泛用于关系抽取领域[17-20]。Guo[17]等人以及 Sun 等人[20]利用图神经网络学习句子的语法结构，以丰富句子语义信息。Zhang[18]等人，Yang[25]等人引人了额外的知识库，并提出了基于知识的关系抽取模型，但是由于从知识库和从目标语句中学习的信息属于异构信息，难以有效融合，关系抽取的精度有待提高。为了避免异构信息融合，Zhao 等人[19]设计了一种新的实体对图（EPGNN），它可以表达句子之间的相关性，但他没有考虑依据句子语义挑选有效信息，结果有待进一步提高。

现有的有监督关系抽取算法大都存在以下缺陷：1）现有的方式大都选择从单个句子中学习句子以及实体的特征，但是由于语言的多样性导致从单个句子学习到的实体特征往往比较片面，难以对实体有全面的认识，而关系抽取最终是根据两个实体的特征进行关系分类，因此片面的信息往往会影响到关系抽取的精度；2）一些方法尝试从额外的知识库或利用其它 NLP 工具引入知识以增加对实体的认识，但是这种方式会引入异构信息，或因 NLP 工具性能问题导致级联错误，因此这类方法的关系抽取精度有待提高；3）由于知识的多样性以及复杂性，简单的将实体所涉及的知识全

部引入模型势必会导致不相干信息的干扰，因此，现阶段缺少一种可以筛选与实体对关联密切的知识的一种方式。

1.2.2 远程监督关系抽取算法

由于基于有监督的关系抽取算法需要大量的训练数据来训练模型，使用手工标注的数据集进行关系抽取需要花费大量的时间和精力来构建数据集。而在已经建立的许多知识库（例如 DBpedia[26]，Freebase[27]，Wikidata[28]等）中存在大量实体关系三元组，只需要将这些关系三元组标记与原始文本中相应的句子对齐，就可以利用这些知识库中的语义信息来提高关系抽取的性能。因此，Mintz[8]首次提出了远程监督关系抽取，远程监督关系抽取利用知识库作为训练数据源，将所有包含目标实体对的句子的标签都标记为知识库中对应的关系，这样会导致训练数据中存在很多错误标签，这些具有错误标签的句子带来了噪声。因此，如何去除数据集中的噪声或增强模型的鲁棒性成为了远程监督关系抽取的关键问题。为解决错误标注的问题，已有的研究[29]将远程监督关系抽取任务视为多实例学习（MIL）。MIL 将包含相同实体对的句子划分成一个句包，句包中以知识库为单位标注实体对所提及的关系。在这些研究中，针对噪声的处理方式不同，远程监督关系抽取算法主要分为以下两种[7]：基于选择性的方法和基于降噪的方法。

基于选择性的远程监督关系抽取算法的基本思想是有选择地从句包中筛选有用的句子进行关系抽取。Zeng 等人[30]提出了 PCNN 模型，他们只选择句包中一个最有可能标注正确的语句来预测实体对的关系。这意味着，该模型忽略了句包中几乎所有的其他的句子。针对 PCNN 模型仅使用句包中最相关的一句话作为正样本的缺点，Lin 等人[31]在句包中的所有实例上使用了注意机制[32]来选择有意义样本，该机制赋予句包中每个句子一个权重来表示不同句子的重要性。此后，基于选择性注意力机制的方法[35-38]被广泛应用于远程监督关系抽取中。然而，这种基于注意力的方法容易受到一个句包所有句子都是噪声的影响。为缓解这一问题，Li 等人[39]提出了选择性门控机制将句子的表示聚合为句包的表示来进行远程监督关系抽取。虽然基于选择性的远程监督关系抽取算法一定程度上削弱了噪声对关系抽取结果的影响，但仍不能克服句包里所有句子都贴错标签的问题。

与基于选择性的远程监督关系抽取算法不同的是，基于降噪的远程监督关系抽取侧重于关注并处理句包中的噪声数据。为了降低句包中的噪声，Qin 等人[40]提出了一种基于生成对抗网络的方法来去除数据中的噪声，Jia 等人[41]采用基于匹配的方式为每个标签类别设置一种模式，通过样本与模式的匹配程度来识别噪声数据，以上两种方式都需要人工干预，局限性较大。此外，基于强化学习的去噪方法[42,43]利用强化学习模型来识别并移除训练数据中的错误样本，从而降低训练数据的噪声，然而由于强

化学习的奖励延迟机制导致难以寻找最优决策，降噪精度有待提高。Shang 等人[44]指出，以往的去噪方法忽略了产生噪声标注问题的根本原因——缺少正确的关系标注，他们通过对噪声数据重新标注来避免有用信息的丢失，但是由于难以衡量重新标注的标签的准确性，该算法的精度有待提升。He 等人[45]将噪声数据作为负样本，与正样本一起训练模型有效利用噪声数据提升模型精度，但其采用强化学习的降噪方式局限性较大，因此关系抽取的性能有待提升。基于降噪的方法可以从根本上解决远程监督关系抽取面临的噪声问题，但是该方法严重依赖噪声识别器的识别精度，此外如何有效利用噪声数据也是当前面临的一大难题，因此基于降噪的远程监督关系抽取算法的性能还有很大提升空间。

通过对基于远程监督的关系抽取算法的研究可以发现，现阶段的远程监督关系抽取算法的缺陷可概括为：1）基于选择性的远程监督关系抽取算法由于关注的是标注正确的数据，因此很难解决句包中所有句子都标注错误的问题；2）当前存在的识别噪声的方式都有较大局限性，现阶段缺乏一种简单有效的降噪方式来准确识别出远程监督句包中的噪声数据；3）噪声数据中往往包含着可以利用的信息，而现有方法在识别出噪声数据后大都选择将噪声数据丢弃，但是如果能够充分利用噪声数据，将会对关系抽取的精度有很大提升；4）一个优秀的特征提取器可以大大降低噪声识别的难度，因此如何设计一种高性能的特征提取器提取句子以及实体的特征也是当前面临的难点。

总之，有监督/远程监督关系抽取算法所面临的问题是信息抽取领域的瓶颈，国内的研究相对较为落后，而国外的研究也面临着各种挑战，因此本文将致力于这些关键性问题的研究，旨在设计出一种高效的有监督关系抽取算法和一种高效的远程监督关系抽取算法。

1.3 论文的主要工作及创新点

随着网络数据的日益暴增，传统关系抽取算法已难以满足实际应用需求，本文通过查阅相关文献，对现有的基于深度学习的有监督关系抽取算法以及远程监督关系抽取算法进行深入研究，并总结其缺陷与不足，针对现有算法的不足，分别提出一种高效的有监督关系抽取算法和一种高效的远程监督关系抽取算法。本文主要工作与创新点如下：

1）本文针对现有有监督关系抽取算法仅关注于单个句子，而忽略了具有相同实体的句子间的联系的问题，建立了一种表示不同句子中实体之间关联的实体关系图，并为此设计了一种基于语义相似度的图神经网络，用于筛选最相关的信息，并且尽可能排除不相干信息对于关系抽取的影响。在此基础上提出

了一种新的基于实体关系图的有监督关系抽取算法 ERGSRE(A Novel **E**ntity **R**elation **G**raph for **S**upervised **R**elation **E**xtraction)；

2）本文通过对现有远程监督关系抽取算法的研究，针对其特征提取不完全，噪声识别精度不足，难以充分利用噪声数据等缺陷，利用图神经网络学习句子的结构信息以完善句子特征，并提出了一种基于数据边距度量的降噪方式，此外将噪声数据作为负样本通过负样本学习的方式提升远程监督关系抽取算法的精度。据此提出了一种基于边距度量的远程监督关系抽取算法 MMDSRE(A **M**argin **M**etric for **D**istantly **S**upervised **R**elation **E**xtraction)；

3）本文在真实数据集上对所提出的有监督以及远程监督关系抽取算法进行了仿真实验，通过与经典算法以及现有最优算法的对比，证明了作者所提出的算法的有效性.此外，本文将所提出的有监督以及远程监督关系抽取算法集成至开源的关系抽取工具包 OpenNRE 中，可以一键式地抽取给定语句中的实体关系，并提升了 OpenNRE 的性能。

1.4 论文的组织结构

文章的组织结构如下：

第一章 绪论。介绍本文的选题背景和意义，并对国内外的关系抽取算法的研究现状进行了相关总结，同时介绍了本文的主要工作与创新点和论文的组织结构。

第二章 基础理论与相关工作综述。首先给出了有监督关系抽取和远程监督关系抽取的相关形式化定义，并对有监督和远程监督关系抽取工作进行了综述与分析。

第三章 ERGSRE：基于实体关系图的有监督关系抽取算法。首先对 ERGSRE 算法的设计动机、算法思想与算法框架进行了简单介绍，然后详细阐述了本算法的设计策略，最后通过在真实数据集上的仿真实验，分析论证了本算法的有效性。

第四章 MMDSRE：基于边距度量降噪的远程监督关系抽取算法。首先介绍了 MMDSRE 算法的设计动机和算法框架，然后对本算法的各个设计细节进行详细介绍，之后通过 MDSRE 与现有算法在真实数据集上进行实验对比，证明 MMDSRE 的高效性。

第五章 与 OpenNRE 的集成。将本文所提出的有监督以及远程监督关系抽取算法集成至开放的关系抽取工具包 OpenNRE 中，方便地抽取给定语句中的实体关系。

第六章 总结与展望。对本文工作进行概括总结，并对本文今后的研究方向进行展望。

第二章 问题定义与相关工作综述

本章首先对本文涉及的相关概念进行形式化定义，之后，从基于深度学习的有监督关系抽取和远程监督关系抽取算法中选取具有代表性的算法进行了研究与综述，最后，对这些相关算法所采取的机制及性能进行了对比。

2.1 问题定义

为方便本文的论述，本节首先给出相关的基本概念与形式化定义。

定义 2.1 **实体** 实体 e 指客观存在并可相互区别的事物，实体可以是具体的人、事、物，也可以是抽象的概念或联系。

定义 2.2 **实体关系** 实体关系是指两个实体之间的关联。形式化地，将实体关系定义为三元组(e_i,r,e_j)，其中 e_i 和 e_j 是两个实体，r 是来自所有可能的关系集合 R ($R = \{r_1, r_2, \dots, r_k\}$，$k$ 是关系数量)的关系。

定义 2.3 **关系抽取** 给定一个包含两个实体 e_i 和 e_j 的句子 s，关系抽取是预测出实体 e_i, e_j 间的实体关系(e_i,r,e_j)，$r \in R = \{r_1, r_2, \dots, r_k\}$。因此，关系抽取被定义为一个分类任务。

定义 2.4 **远程监督关系抽取** 给定一个包含 b 个句子的句包 S_b, $S_b = \{s_1, s_2, \dots, s_b\}$，这 b 个句子都包含了相同的两个实体 e_i, e_j，远程监督关系抽取的目的是预测出 S_b 中包含的实体对间的关系(e_i,r,e_j)，$r \in R = \{r_1, r_2, \dots, r_k\}$。

2.2 有监督关系抽取算法的研究与综述

根据 1.2.1 节所述，传统的有监督关系抽取算法[9,10]以及基于 CNN[11]、RNN[12,13] 模型的有监督关系抽取算法由于模型局限性较大导致关系抽取结果较差，而基于预训练模型[14,15,18]和基于图神经网络[17,19,20]的有监督关系抽取算法近年来取得了较大突破，因此本节将对现有的具有代表性的基于预训练模型和基于图神经网络的有监督关系抽取算法进行研究与综述。

2.2.1 基于预训练模型的关系抽取算法

最近，一些研究工作将关系抽取作为预训练语言模型的下游任务，并取得了较好的效果。特别是自从 2018 年 Google 发布预训练模型 BERT[24]以来，BERT 或其改进模型 RoBERTa[46]，ALBERT[47]等被广泛应用于关系抽取任务[14,15,18]。

BERT 的模型结构如图 2.1 所示，除了输出层之外，模型的预训练和微调都使用

相同的结构。在使用 BERT 进行不同下游任务时，首先使用相同的模型参数进行初始化。在微调过程中，会对所有参数进行调整。BERT 输入的设计是为了使其能够在一个输入中同时表示一个或一对文本句子，每个输入表示由其对应单词的特征，位置信息等组成。两个句子之间以特殊分隔符“[SEP]”标记，每个输入前以“[CLS]”标记。“[CLS]”标记位置的输出常用于表示分类任务中整个句子的特征。BERT 通过一个预训练任务（masked language model，MLM）对模型参数进行预训练，该任务随机屏蔽输入中的一些单词，并根据上下文信息来预测被屏蔽词。与从左到右的语言模型的预训练不同，MLM 任务可以充分学习一个单词的上下文信息。除了 MLM 外，BERT 还训练了一个预测下一个句子(Next Sentence Prediction，NSP)的任务，通过随机生成句子对，来让模型判断句子对是否连续。BERT 通过 MLM 和 NSP 任务对模型进行联合预训练，之后，在进行下游任务（如关系抽取、智能问答等）时，只需要加载训练好的预训练模型，并对其参数进行微调即可。改进版的 BERT 如 RoBERTa[46]，ALBERT[47]等模型的基本原理、输入输出与 BERT 相同，只是将 BERT 模型进行压缩或改变预训练任务。

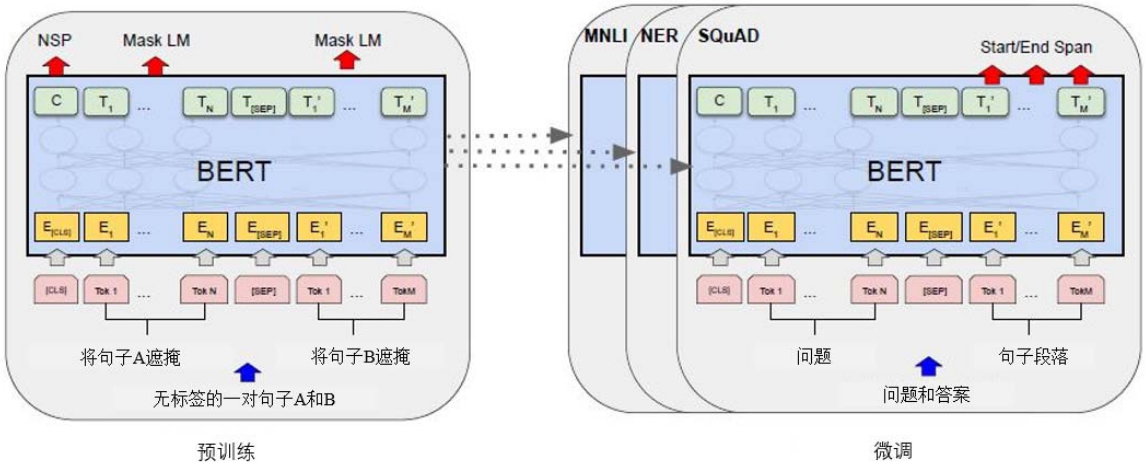

图2.1 BERT 预训练和微调模型结构示意图

基于 BERT 的关系抽取算法是在预训练模型 BERT 上进行参数微调，并适当修改模型输入格式。R-BERT[14]通过将两个实体前后分别加入特殊符号“$”和“#”来标记两个实体，将带有实体标记的句子输入 BERT 后，将“[CLS]”位置输出作为句子语义特征表示，将“$”标记位之间和“#”标记位之间的输出分别进行平均作为两个实体的特征表示。最后将句子的特征表示与实体的特征表示拼接作为分类层的输入来抽取关系。与 R-BERT 类似，$BERT_{EM}$[15]分别在第一个实体前后加入“[E1]”、“[/E1]”，在第二个实体后加入“[E2]”、“[/E2]”来标记实体。不同的是作者将实体开始标记位“[E1]”、“[E2]”的输出作为实体的特征表示，且仅将实体表示作为分类层的输入。

以上两个算法的方法类似，取得的结果也较为接近，其共同缺点是仅考虑了单个句子内部的语义信息，这种方法所获取的信息有限，因此关系抽取的精度有待提高。

为了解决上述缺陷，ERNIE[18]将知识图谱引入关系抽取模型中，作者使用 TransE 知识嵌入算法[49]将知识图谱中的结构化知识编码，然后将编码后的知识信息与 BERT 抽取的语义信息融合进行关系抽取。为了更好地将语义和知识信息融合起来，模型改进了 BERT 模型的架构，并设计了新的预训练任务。作者设计了文本编码器 T-Encoder 和知识编码器 K-Encoder 两个模块。对于 T-Encoder，和 BERT 模型一致，它负责获取输入句子的词法和句法等语义信息。对于 K-Encoder，它将额外知识信息与来自底层的文本信息有效结合，以缓解来自 BERT 以及知识图谱的异构信息融合问题。该方法虽然能引入额外的知识信息以提高关系抽取精度，但是其需要依赖模型以外的 NLP 工具如 TransE，知识图谱等，容易发生级联错误。

综上所述，基于预训练模型的有监督关系抽取算法可以利用预训练模型从句子中学习到较为准确的特征，取得较好的关系抽取效果，但是，这种仅利用单个句子学习实体信息的方法由于学习到的实体信息比较片面，结果仍可以进一步提升。虽然已有工作尝试从外部引入知识来丰富实体对的信息，但是现有的引入知识的方法难以解决异构信息融合影响关系抽取精度的问题，因此如何从句子以外学习到有效知识仍是有监督有关系抽取面临的一大难题。

2.2.2 基于图神经网络的关系抽取算法

图是由节点 V 和边 E 组成的数据结构，给定一个含有 n 个顶点的图$G = (V, E)$，可以用邻接矩阵$A = [a_{ij}] \in R^{n\times n}$表示各个顶点间的邻接关系，当顶点 v_i 与顶点 v_j 间有边连接时，a_{ij}=a_{ji}=1，否则 a_{ij}=a_{ji}=0。为考虑顶点自身的关系，将自连接加入邻接矩阵，因此可得新的邻接矩阵$\tilde{A} = A + I_N$，其中$I_N \in R^{n\times n}$是单位矩阵。度矩阵$\widetilde{D}$可表示为：$\widetilde{D} = diag(\tilde{d}_1, \tilde{d}_2, \cdots, \tilde{d}_n)$，其中，$\tilde{d}_i = \sum_j \tilde{a}_{ij}$是顶点 v_i 的度。

图神经网络(GNN)是基于深度学习的在图上进行相关操作的神经网络。基于频谱卷积的图神经网络 GCN[48]在图信号处理中具有坚实的基础，它通过在频谱域上分解图像信号，然后应用频谱滤波器来定义卷积操作，对于一个图信号$X \in R^{n\times d}$，其中 d 是顶点特征向量的维度，具有频谱滤波器 W 的频谱卷积运算可以用公式(2-1)形式化地表示：

$$H = GCN(X, A) = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}XW) \tag{2-1}$$

$$H^{t+1} = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^tW^t) \tag{2-2}$$

其中，H 是图卷积得到的状态表示，W 是可训练的参数矩阵，σ是激活函数（例如$ReLU(\cdot)$等）。公式(2-2)是公式(2-1)的迭代形式，H^t 是 H 的第 t 次迭代，即第 t 层图卷积得到的状态表示，且 $H^0=X$，W^t 是第 t 层的可训练的参数矩阵。

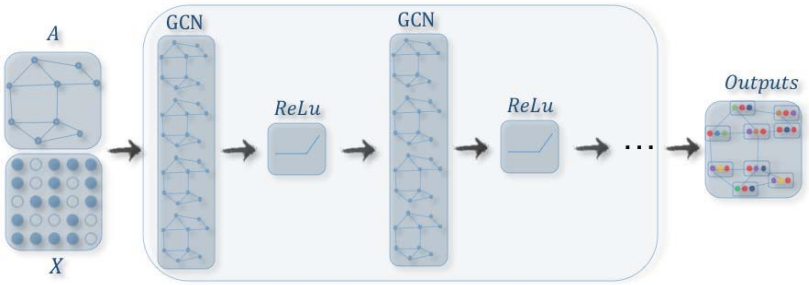

图2.2 多层图卷积神经网络示意图

如图 2.2 所示，GCN 通过聚合邻居节点的特征信息来封装每个节点的隐藏表示。在特征聚合之后，对结果输出应用非线性变换。通过堆叠多层 GCN，每个节点的最终隐藏表示可以从更远的邻居信息中获取。

AGGCN[17]充分利用了图神经网络来挖掘语句中的语法信息，作者首先依据句子的语法依赖树建图，然后利用两层 GCN 模型学习的语法结构，进而充分理解语义信息。如图 2.3 所示，传统的基于语法依赖树的方法大都采用基于规则的硬剪枝方法，可能会删除树中的一些重要信息，与传统的基于依赖树方法不同，作者将原始依赖树转换为完全连通带权图，这些图的权重被视为节点之间的相关性强度，并使用注意力机制以端到端的方式学习。作者将这种这种基于注意力机制获取图的权重的方法称为软剪枝策略，使用该策略既能避免删除有用信息，又能筛选有效信息。从而能够捕获丰富的局部和非局部依赖信息。

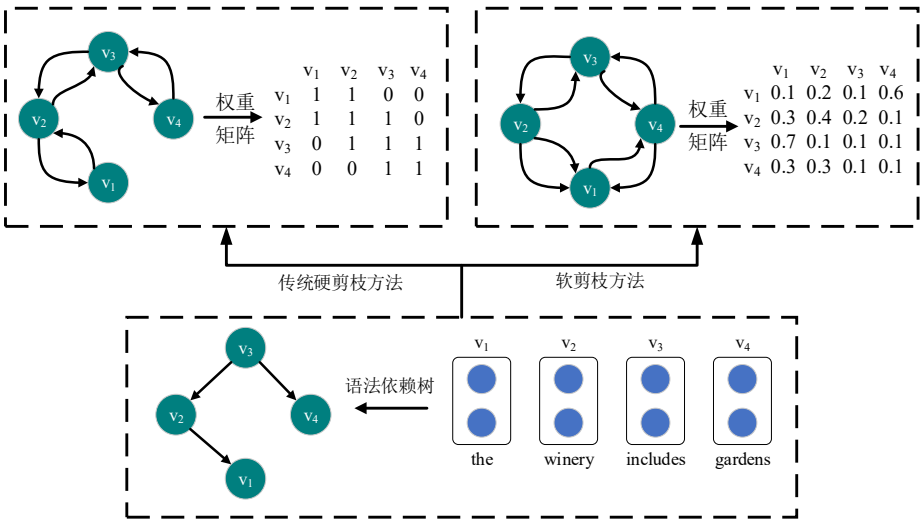

图2.3 AGGCN 算法软剪枝方式与传统硬剪枝方式对比示意图

然而上述这些工作仅对于单个句子建图，试图通过使用图神经网络依据拓扑结构从句子中挖掘更多的信息。近两年，一些研究者试图跨句子建图，将不同句子中有关联的实体相连接，在抽取一个句子中一对实体之间的关系时，使用图神经网络从相关联句子中挖掘有关信息，增强对当前待分类关系实体对的了解，提高分类精度。Zhao等人[19]考虑整个文库中所有实体对之间的联系，将每个句子中的实体对作为图的顶点，用边连接具有相同实体对的两个顶点来建立实体对图，并使用图卷积神经网络来捕获实体对图的拓扑特征。通过这种方式，他们的模型可以充分地利用给定的语料挖掘知识。但是，他们这种方式将所有包含相同实体的句子平等的连接在了一起，没有对跨句子获取的信息进行筛选，容易引入不相关或相关程度很低的信息，一定程度上影响分类精度。Nan 等人[20]针对文档级别关系抽取问题，将文档中出现的所有实体作为顶点建图，在文档的多个句子之内和之间整合信息。通过将建图范围限定在同一个文档的句子，他们的模型可以有效避免不相干信息的干扰，然而这种强硬的限定方式势必会造成一些有效信息的损失。

基于图神经网络的关系抽取算法可以有效利用节点邻居信息来丰富语义信息，特别是当不局限于单个句子建图时，还可以从其他句子中获取有用信息。但是，如何跨句子建图，以及如何设计一种可以有效筛选有用信息的图计算方式是基于图神经网络的关系抽取算法所面临的关键问题。

为更加直观展示有监督关系抽取算法的特点，本文分别从算法所采用的主要机制以及算法性能等方面将上述算法进行对比，如表 2.1 所示：

表2.1 有监督关系抽取算法采用主要机制及性能对比表

算法名称	分类	端对端模型	考虑句子结构	引入额外知识	筛选知识	性能
R-BERT[14](2019)	预训练模型	是	否	否	否	中等
$BERT_{EM}$[15](2019)	预训练模型	是	否	否	否	中等
AGGCN[17](2019)	图神经网络	否	是	否	否	较低
ERNIE[18](2019)	预训练模型	否	否	是	是	较低
EPPGN[19](2019)	图神经网络	是	是	是	否	较高
LST-AGCN[20](2020)	图神经网络	是	是	是	否	中等

表 2.1 展示了一些具有代表性算法采取的主要机制以及算法性能，其中算法性能是根据算法在真实数据集 Sem-Eval 2010 Task 8[52]上的 *F*1-*score* 值进行划分，由于各算法的 *F*1-*score* 值集中在 80%-90%，因此根据其 *F*1-*score* 可划分为低(<80%)、较低(80%~85%)、中等(85%~90%)、较高(90%~95%)、高(>95%)五个等级。从表 2.1 可以

看出，EPGNN[19]算法由于引入了额外知识信息，且在特征提取时充分考虑句子结构，取得了最好的结果，但是由于其未能有效筛选知识，性能仍受到较大限制。对以上所有算法进行分析可以发现，关系抽取算法随着人们的深入研究越来越完善，特别是引入额外的知识这一方法逐渐引起关注，但是，目前通过引入额外知识的方法普遍存在如下缺陷：1)通过从额外知识库引入知识，但是由于外部知识库与训练数据分布不一致导致融入异构信息，影响关系抽取的精度，并且该方法需要依赖外部 NLP 工具；2)从训练数据集中挖掘知识的算法（如 EPGNN，LST-AGCN）由于没有对知识进行有效筛选，可能存在不相关信息的干扰。因此，如何在不依靠外部工具，设计一种可以有效引入额外知识的关系抽取算法是当前有监督关系抽取亟待解决的一个关键问题。

2.3 远程监督关系抽取算法研究与综述

正如前文所介绍的，根据对噪声的处理方式不同，远程监督关系抽取算法可分为两类：基于选择性的远程监督关系抽取算法[30, 31, 38, 39]和基于降噪的远程监督关系抽取算法[42, 44, 45]。本节将选择这两类方法中的具有代表性的算法进行深入研究与综述。

2.3.1 基于选择性的远程监督关系抽取算法

基于选择性的远程监督关系抽取算法对于远程监督句包中的噪声的处理方式是忽略这些噪声，选择最能表达实体对关系的句子训练模型。Zeng[30]等人首次提出了 PCNN 模型进行远程监督关系抽取，他们使用 CNN 模型对相邻的单词进行卷积操作来捕获句子中各单词的语义特征，为了使其适用于关系抽取任务，突出实体的特征，作者根据实体的位置将句子分为三段，并提出了分段池化的策略分别对第一个实体之前、两个实体之间以及第二个实体之后的词进行池化操作后得到句子的语义特征，在进行关系抽取时，作者训练模型使其从句包中选择最能表达实体对关系的句子来预测他们的关系。该算法由于受 CNN 卷积核的限制，难以学习句子的全局信息，并且只从句包中选择一个句子的方式会忽略其他句子中的许多有用信息，因此关系抽取的精度有待提升。

为了解决 PCNN 算法中仅从句包选择一个句子进行关系抽取带来的缺陷，Lin[31]等人在 2016 年提出了 PCNN+ATT 算法，它使用 PCNN[30]模型和 Attention 机制[32]相结合的方式来进行远程监督关系抽取。PCNN+ATT 算法模型图如图 2.4 所示。作者首先使用 PCNN 模型作为句子的特征提取器提取句包中的每个句子的特征$h_{s_i}(i \in \{1,2,\cdots,b\})$，之后，作者使用 Attention 机制为每个句子训练一个权重α_i，对能够表达句包中实体对的关系的句子赋予高的权重，否则，赋予较低权重。之后，将每个句子的特征加权后得到整个句包的特征h_S。最后，将得到的句包特征h_S输入分类层进行关

系分类。该方法充分利用了 Attention 的优势以动态方式为句包中的每个句子赋予不同权重，对于噪声数据有较好的抑制效果。但是由于其特征提取器使用的是 PCNN，不可避免地会存在受限于卷积核的大小，难以获取远距离文本的信息的情况，局限性较大。

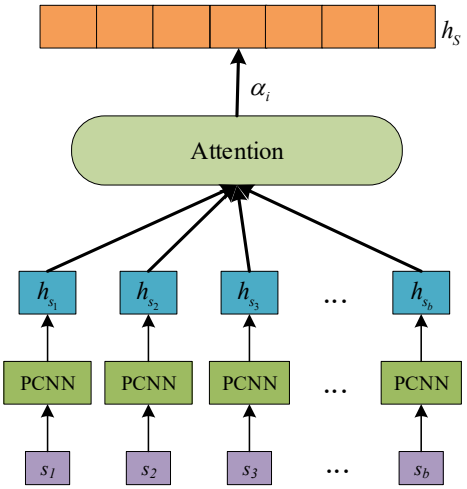

图2.4 PCNN+ATT 模型示意图

与上述工作类似，Duan[38]等人也使用了 Attention 机制为句包中的句子获取权重，不同的是，作者使用图神经网络作为特征提取器得到每个句子的特征。如图 2.5 所示，在作者提出的 HG 模型中，作者分别将实体特征、实体类型、句包特征、关系路径以及实体上下文等作为顶点建立一个混合图（即顶点类型不同的图），并利用图卷积神经网络得到句包中每个句子的特征。之后使用 Attention 机制聚合各个句子得到整个句包的特征，最后将句包表示输入 softmax 层用于关系抽取。作者通过建立混合图并利用图卷积神经网络能够较好的提取句子的特征，最终抽取的关系准确性也较高。

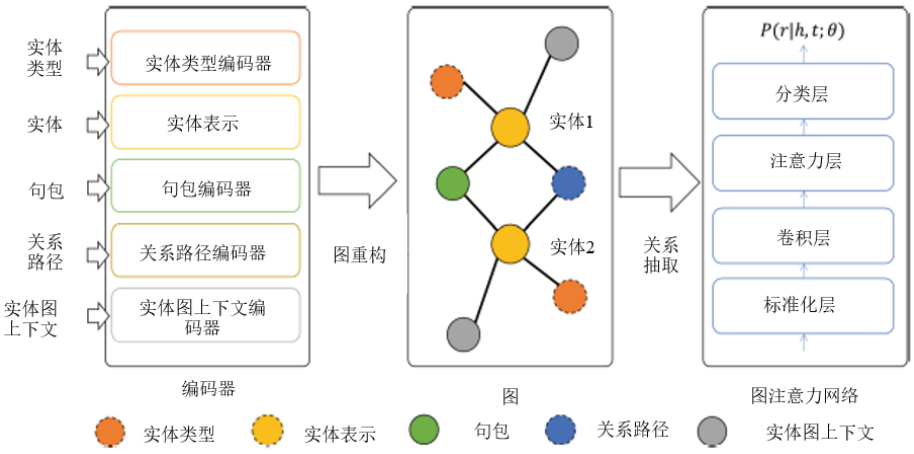

图2.5 HG 模型示意图

上述两种方式都是利用 Attention 的方法为句包中的句子训练权重，并对其进行加权得到整个句包的表示用于关系抽取，但是，这种基于 Attention 的方法由于其权重是进行归一化得到的结果，当句包中只有一个句子时，利用 Attention 计算得到的权重始终为 1，当这个仅存的句子为噪声时，Attention 机制便失去了作用。并且据 Li[37]等人统计，对于一个真实的远程监督数据集，例如 NYT[29]，80%的训练样本是由单个语句组成的句包，并且这些句包中，有 35%左右的句子标签标记错误。

为了解决使用 Attention 机制带来的问题，Li[39]等人提出了 Seg 算法，使用选择性门控网络从句包中选择正样本(即能正确表示句包中实体对关系的样本)。作者使用 PCNN 作为特征提取器提取句子特征h_{s_i}，此外，为解决 PCNN 模型中 CNN 受卷积核限制提取特征不完善的缺陷，作者使用 Attention 机制将每个句子中的单词特征 x_i 聚合为句子特征h_{u_i}。然后作者将由注意力机制得到的句子特征h_{u_i}输入至选择性门控网络为句子训练权重 g_i，如公式(2-3)所示：

$$g_i = \text{sigmoid}(W^{(g1)}\sigma(W^{(g2)}h_{u_i} + b^{(g2)}) + b^{(g1)}), \forall i = 1, \cdots, m \tag{2-3}$$

其中，$W^{(g1)}, W^{(g2)}$是可训练的参数矩阵，$b^{(g1)}, b^{(g2)}$为偏移，$\sigma(\cdot)$为激活函数，$g_i \in (0,1)$是经过选择性门控网络得到的权重。之后通过 g_i 与h_{s_i}的加权和得到最终的句包表示h_s，并将其输入至分类层。

根据以上介绍内容，基于选择性的远程监督关系抽取算法的模型可抽象为图 2.6 所示：

图2.6 基于选择性的远程监督关系抽取算法的模型示意图

如图 2.6 所示，该类方法首先使用句子特征提取器提取每个句子的特征表示h_{s_i}，之后设计合适的选择性机制将句子的特征聚合得到整个句包的特征表示h_S，最后将h_S输入至分类器进行关系抽取。

基于选择性的远程监督关系抽取算法可以有效地从句包中选取正样本，以减少噪声数据的影响，并取得了较好的效果。但是难以解决一个句包所有句子都标注错误的问题，虽然一些工作尝试缓解这一问题，但仍不能从根本上解决。这是基于选择性的远程监督关系抽取算法的固有缺陷。

2.3.2 基于降噪的远程监督关系抽取算法

基于降噪的远程监督关系抽取算法专注于对噪声数据的识别与处理，该类算法通过设计合理的降噪器识别出句包中的噪声数据，然后删除这些噪声数据或为噪声数据重新标记正确标签，最后利用正确的标签数据训练分类器进行关系抽取。

Qin[42]首次使用强化学习来处理句包中的噪声问题，对于强化学习（RL），其拥有的两个必备组件分别是：外部环境（external environment）和 智能体(agent)，而一个具有良好鲁棒性的 agent 正是通过这两个组件的动态交互而训练出来的。作者使用的 agent 的目标是根据关系分类器性能的变化，决定是保留还是移除当前的实例（即一个句子）。然后，框架进一步使基于深度强化学习策略的 agent 学会如何重建一个纯净的远程监督训练数据集。模型采用结果驱动策略，以关系分类器的性能变化为依据，对 agent 的一系列行为决策进行奖励。奖励通过相邻训练步(epochs)的差值来表示，如公式(2-4)所示。

$$R_i = \alpha(F_1^{(i)} - F_1^{(i-1)}) \tag{2-4}$$

其中，$F_1^{(i)}$为第 i 步时，关系分类器的性能得分，如式(2-4)所示，若在第 i 步时，F_1 增加，则 agent 将收到一个正奖励；反之，则 agent 将收到一个负奖励(惩罚)。α 的作用是将 F_1 的差值转换到有理数的范围内。经过上面的强化学习过程，对于每一种关系类型，都得到了一个可以识别出样本中噪声的识别器 agent，识别出数据中的噪声后，选择直接删除这些噪声即可得到纯净的训练数据集，利用净化后的数据进行关系分类，即可抽取出每个句包所描述的关系。

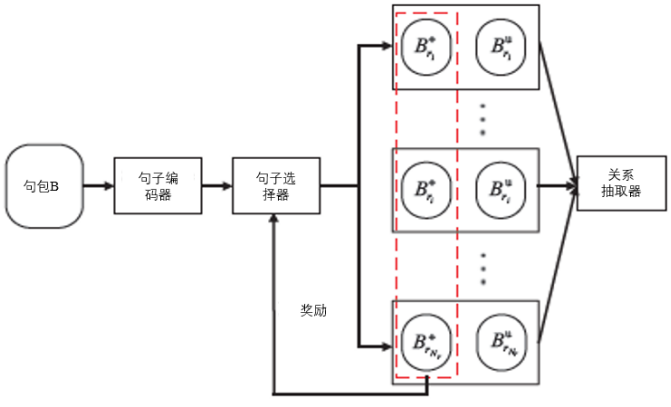

图2.7 强化学习和正负样本训练模型示意图(红色虚线框内为正样本数据组合)

与 Qin 等人[42]的工作类似，Zhang 等人[45]也采用了强化学习来识别训练数据中的噪声，不同的是，在识别出噪声数据后，Zhang 等人[45]将噪声数据作为对应关系类别

的负样本，与正样本(即标注正确的数据)一起训练，以提升分类器的精度。如图 2.7 所示，作者利用强化学习识别出的噪声数据与正样本，将正样本数据组合标记为$B_{r_i}^{+}$（图 2.7 红色矩形框所示），将噪声数据去除标签后作为无标签数据记为$B_{r_i}^{u}$，将$B_{r_i}^{+}$与$B_{r_i}^{u}$的特征线性加权输入至分类器进行关系分类。作者充分利用了噪声数据，提出了两种新的有标签和无标签样本的表示方法。然后以适当的方式将这两种表示结合起来进行关系抽取。

Shang 等人[44]指出以往的去噪方法忽略了产生噪声标注问题的根本原因——缺少正确的关系标注。因此作者在识别出噪声数据后，试图为这些噪声纠正错误标签，从而丰富训练数据。作者设计并提出了 DCRE 模型进行关系抽取，DCRE 模型由特征提取器、去噪器、标签生成器组成。其中特征提取器使用的是 PCNN 模型用于提取每个句子的特征表示，之后，作者使用去噪器识别出噪声数据，并用标签生成器为噪声数据重新标注正确标签。在去噪器中，作者将句包中的每个句子与其对应标签进行相似度匹配，并设置一个阈值，若相似度低于阈值则认为该句子为噪声语句。作者使用聚类方法为噪声数据重新标注标签，具体做法是，首先将所有噪声数据的原标签去除，然后将每类实体关系标签作为一个样本中心点，用 K-means 方法为这些无标签数据聚类。作者提出的为噪声数据重新标注标签的思想很新颖，但是，使用聚类方式标注的标签准确性难以衡量，很容易再次引入噪声，并且作者在去噪器中设置的阈值需要一定的先验知识以及大量实验验证，不能随着样本数据分布而动态变化，因此具有很大的局限性。

基于降噪的远程监督关系抽取算法的模型示意图可抽象为如图 2.8 所示的形式。

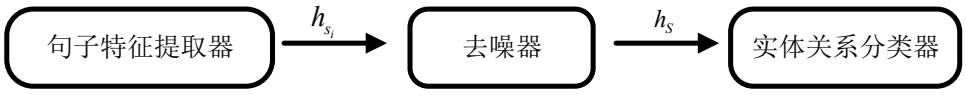

图2.8 基于降噪的远程监督关系抽取算法模型示意图

如图 2.8 所示，基于降噪的远程监督关系抽取算法首先需要根据句子级别的特征提取器准确提取每个句子的特征，然后使用降噪器根据句包中每个句子特征准确识别出与其标签不匹配的噪声句子，最后对句包中的噪声数据采取合适的策略处理后采用实体关系分类器进行关系抽取。

基于去噪方式的远程监督关系抽取方法的关键是噪声的识别以及如何有效利用噪声数据。当前主流的方式如强化学习由于决策的多样性引起环境的变化，会导致模型需要耗费大量资源计算最优决策，且由于延迟的奖励机制，因此寻找最优决策变得更加困难。而另外一些基于匹配的方式往往需要大量先验知识，因此设计一个灵活的去噪器来识别噪声数据是当前研究所面临的一个重大困难。噪声数据中往往包含有用信息，如何合理利用噪声数据提升关系抽取精度仍是一个亟待解决的问题。另外，当

前基于去噪的方式过于关注去噪器的研究，而忽略了最基本的特征提取器，只有准确地提取出句子特征，后续步骤才能有效进行。由于图神经网络在特征学习方面表现的优异性，在特征提取器中加入图神经网络可以提升特征提取器的效果。

为更加直观对比各远程监督关系抽取算法的性能，表 2.2 列举了目前一些主流的算法的特点。

表2.2 远程监督关系抽取算法性能对比表

算法名称	类别	特征提取器	识别噪声方式	有效利用噪声	性能
PCNN+ATT[31](2016)	基于选择	PCNN	无	否	低
HG[38](2019)	基于选择	GNN	无	否	较低
SeG[39](2020)	基于选择	PCNN	无	否	中等
PCNN+RL[42](2018)	基于降噪	PCNN	强化学习	否	较低
DCRE[44](2020)	基于降噪	PCNN	基于匹配	是	中等
PCNN+PU[45](2020)	基于降噪	PCNN	强化学习	是	较高

表 2.2 展示了一些具有代表性算法采取的主要机制以及算法性能，其性能是根据算法在真实数据集 NYT[29]上的精度进行衡量，根据算法精度，其性能可划分为低(<75%)、较低(75%~80%)、中等(80%~85%)、较高(85%~90%)、高(>90%)五个标准。对这些算法进行分析可以得到：1）基于选择的远程监督关系抽取算法由于不能有效识别噪声数据，因此难以挖掘噪声数据中的有用信息，且难以处理句包中所有句子都是噪声的情况；2）现有的基于降噪的方法特征提取器使用的大都是 PCNN 模型，这种方式由于受卷积核大小的限制，难以对长句子的特征进行准确提取；3）现有的识别噪声的方式大都基于匹配或强化学习的方法，正如上文所述，这些方法都有较大的局限性，因此采用合适的特征提取器，设计有效的识别噪声方式并合理挖掘噪声信息是研究远程监督关系抽取问题的关键。

2.4 本章小结

本章首先对本文涉及的一些基本概念进行了形式化定义，然后对基于有监督的关系抽取算法进行深入研究与综述，总结了现有基于预训练模型及基于图神经网络的方法的优势及不足，并且对现有的如何有效引入外部知识的方法进行了综述。最后，深入研究了现有的基于选择性和基于降噪方式的远程监督关系抽取算法，并指出了当前远程监督关系抽取问题的研究难点。

第三章 ERGSRE：基于实体关系图的有监督关系抽取算法

本章首先论述了现有有监督关系抽取算法的不足，从而引入了本章所提出算法的设计动机。之后，详细论述了本章所提出的一种新的基于实体关系图的有监督关系抽取算法 ERGSRE (A Novel **E**ntity **R**elation **G**raph for **S**upervised **R**elation **E**xtraction)。最后，在标准测试数据集上，将本章算法 ERGSRE 与相关基准算法进行了大量的仿真实验，以验证所提出的 ERGSRE 算法的正确性与有效性。

3.1 算法动机

通过分析预训练模型可以发现，它虽然考虑了上下文语义，但是还是缺少相应的知识信息。换句话说，预训练模型通过大量语料的训练可以判别一句话是否通顺，但是它却不知道这句话描述的是什么，它也许能通过训练学到上下文单词之间的一些关系，但是这样的关系还不足以构成知识。举例来说，对于“施耐庵写了《水浒传》”和“贝多芬写了《第五交响曲》”两句话中的实体对“施耐庵”和“水浒传”以及“贝多芬”和“《第五交响曲》”，如果只针对两句话分析，在没有对这些人名以及作品充分了解的条件下，只能推断出两对实体具有相同的关系“写”，而对于更深层次的关系“著作”或者“谱写”，仅从这两句话无法得知具体的关系。而只有引入额外的信息，充分理解了这些实体的含义，例如“《水浒传》”是一本书，而“《第五交响曲》”是一首钢琴曲，才能准确抽取出实体间的关系。这些额外的信息才是知识信息，它们在自然语言中至关重要，如果能够让模型考虑到知识信息，就能让模型不仅在字词、语法层面，还能在知识层面符合人类语言的要求，从而成为一个“有文化”的模型。

通过对前人工作的研究，他们虽然尝试从外部引入知识信息，但是还存在以下缺陷：1）使用 NLP 工具从知识图谱中抽取知识的方式会受到 NLP 工具性能的影响，这种非端对端的方式极易因为前期知识抽取不准确而导致后续关系抽取精度过低；2）从其他语料库或知识图谱中学到的知识很难嵌入到训练数据所在空间，这种异构信息难以有效融合导致关系抽取不准确；3）由于知识的多样性以及复杂性，简单的将实体所涉及的知识全部引入进模型势必会导致不相干信息的干扰，现阶段缺少一种可以筛选有效知识的方式。

本章针对现阶段有监督关系抽取算法的不足，提出了一种基于实体关系图的关系抽取算法 ERGSRE，本章设计的端到端模型不依赖于其他 NLP 工具，且从训练数据所在的语料库建立实体关系图可以有效避免异构信息融合问题，为了筛选有效知识，本章设计了一种基于语义相似度的图计算方式。本章提出的 ERGSRE 算法可以有效

解决上述的三个缺陷。

3.2 算法框架

根据上文所述，有监督关系抽取算法当前面临的最大的难点在于如何在避免使用外部工具的条件下有效引入外部知识以丰富实体对的信息，以提升关系抽取的精度。因此，本章欲设计一种实体关系图(ERG)来存储数据所在语料库的相关知识，并提出一种基于语义相似度的图神经网络从 ERG 中挖掘与实体对相关的知识。此外，为充分学习目标语句的语义，本章使用当前表现较好的预训练模型 ALBERT 作为特征提取器。基于这一思想，本章设计并提出了一种基于实体关系图的有监督关系抽取算法 ERGSRE(A Novel **E**ntity **R**elation **G**raph for **S**upervised **R**elation **E**xtraction)，算法模型框架如图 3.1 所示：

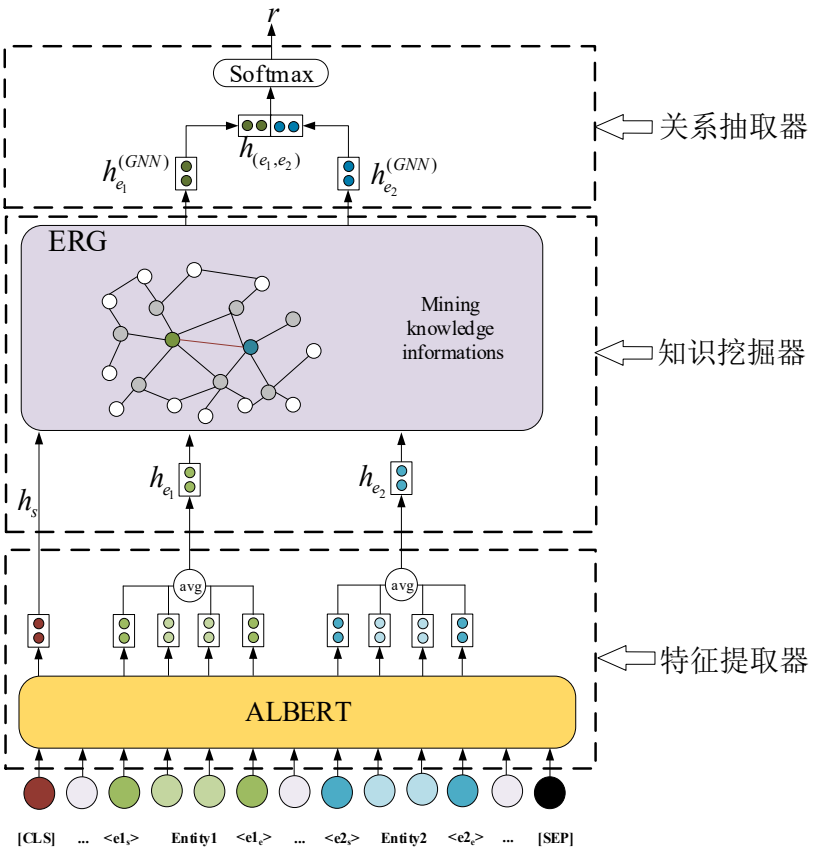

图3.1 ERGSRE 算法模型框架示意图

如图 3.1 所示，本章提出的 ERGSRE 算法主要由特征提取器、知识挖掘器以及关系抽取器三个模块构成，各模块功能如下：

特征提取器：采用 ALBERT 模型对文本句子进行自然语言处理，将各个单词转

化可以包含其上下文、位置等信息的特征向量；

知识挖掘器：本章设计的知识挖掘器从训练数据语料库中构建知识图，并使用基于语义相似度的图卷积神经网络来挖掘知识信息，进而充分理解实体；

关系抽取器：利用分类器对给定的实体对进行关系分类，抽取出其中蕴含的关系。

3.3 算法主要设计策略

3.3.1 特征提取器

特征提取器用来将输入的文本句子编码为计算机可以识别的向量，这些向量需要准确表示出对应单词或句子的特征，因此，一个优秀的特征提取器是后续任务能够顺利进行的基础。

ALBERT 模型[47]是一个由多层的双向 Transformer 编码器[32]组成的预训练语言模型。通过在海量语料的基础上进行自监督学习方法，并利用上下文双向学习为单词学习一个好的特征表示。因此，预训练的 ALBERT 表示比传统的基于任务语料库的单向语言模型生成的词向量包含更多的词汇和语义信息。故本章使用 ALBERT 作为 ERGSRE 算法的特征提取器。

给定一个含有两个目标实体 e_1 和 e_2 的句子$s = \{w_1, w_2, \cdots, w_n\}$，本章使用预先训练好的 ALBERT 模型[47]生成句子的特征表示，并将每个单词转换成一个特征向量来表示其语义特征。对于具有两个目标实体 e_1 和 e_2 的句子 s，为了使 ALBERT 模型捕获这两个实体的位置信息，在第一个实体的开头和结尾插入一个特殊的标记“<e_{1_s}>”，“<e_{1_e}>”，在第二个实体的开头和结尾插入一个特殊的标记“<e_{2_s}>”，“<e_{2_e}>”。同时在句子的开头加入“[CLS]”以表示句子的开始，在句子的末尾加入“[SEP]”以表示句子的结束。例如，对于一个包含实体“kitchen”和实体“house”的句子“The kitchen is the last renovated part of the house .”，加入特殊标记后将转化为“[CLS] The <e_{1_s}> kitchen <e_{1_e}> is the last renovated part of the <e_{2_s}> house <e_{2_e}> . [SEP]”。

将添加特殊标记后的句子 s 输入 ALBERT 模型后，将从 ALBERT 的输出中得到 s 的最终状态表示$H \in R^{n \times d^w}$，其中，n 是输入句子中的单词以及标记个数，d^w是 ALBERT 模型中的单词向量维度。由于实体可能由多个单词组成，假设实体 e_1 由第 i 个单词至第 j 个单词组成，实体 e_2 由第 k 个单词至第 m 个单词组成，那么 e_1 的特征向量可以由其对应单词的最终状态 h_i 至 h_j 取平均后得到，类似地，e_2 的特征向量可以由 h_k 至 h_m 取平均后得到。然后在激活操作（即 tanh）之后，分别对两个实体的特征表示使用全连接层，将它们映射到相同的维度 d^w，并得到两个目标实体的最终状态表示h_{e_1}，h_{e_2}。该步骤可以形式化表示为公式(3-1)，公式(3-2)所示：

$$h_{e_1} = W_1(\tanh(\frac{\sum_{t=i}^{j} h_t}{j-i+1})) + b_1 \tag{3-1}$$

$$h_{e_2} = W_2(\tanh(\frac{\sum_{t=k}^{m} h_t}{k-m+1})) + b_2 \tag{3-2}$$

其中$W_1, W_2 \in R^{d^w \times d^w}$是可训练的参数矩阵，$b_1, b_2 \in R^{d^w}$为偏移。

此外，根据ALBERT模型的特点，使用句首标志位“[CLS]”的最终状态表示作为句子的语义特征，并为其添加激活操作以及全连接层来得到句子的状态表示h_s，如公式(3-3)所示：

$$h_s = W_0(\tanh(h_{[CLS]})) + b_0 \tag{3-3}$$

其中$h_{[CLS]}$为句首标志“[CLS]”的状态表示，$W_0 \in R^{d^w \times d^w}$是可训练的参数矩阵，$b_0 \in R^{d^w}$为偏移。

3.3.2 知识挖掘器

正如 3.1 节所述，只有在充分理解实体的基础上，才能准确抽取出实体对中所蕴含的关系，因此本章设计的知识挖掘器旨在根据训练数据语料库建立知识图，并设计基于语义相似度的图卷积网络挖掘知识图中蕴含的知识，以此来充分理解实体，增加关系抽取的准确性。

由于来自同一语料库的句子往往描述的是相同或相似的事情，因此本章根据训练数据所在的语料库建立实体关系图 ERG 来作为训练数据的知识库，ERG 以实体为顶点，若两个实体之间存在关系就将其连接起来，因此将根据整个训练数据得到一个包含所有实体以及连接所有具有关系的实体对的实体关系图，其中两个实体间的边可看作是对实体关系的描述，因此可以用两个实体所在的句子进行表示。具体地，用$G = (V, E)$表示实体关系图，$V = \{e_1, e_2, \cdots, e_n\}$为所有的顶点（实体）集合，$e_i$ 为第 i 个实体，$E = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}$，$s_{ij}$ 表示实体 e_i 与实体 e_j 之间的边，即两个实体所在的句子。本章设计的实体关系图可以用来表示文库中实体之间的相互关联，因此可以作为数据所在语料库的知识图。

在建立好知识图后，如何高效挖掘知识图中有意义的知识并将其用于关系抽取显得尤为重要。如果简单使用 GCN 从顶点邻居中聚合知识信息，这将导致大量不相关信息的干扰。由于关系抽取是对一个语句中涉及的实体对进行关系抽取，而这个实体对的关系必然与它所在的句子描述的事实相符，因此可以从与实体对所在句子描述内容相关的其他句子中挖掘有用知识。考虑到如果两个句子的语义越相似，则这两个句子讲述的内容越相关，于是本章引入了语义相似度用来评价两个句子内容的相关性，

由此设计了一种基于语义相似度的知识挖掘器。

当抽取一个句子中两个实体 e_i, e_j 之间的语义关系时，这个句子被称为目标语句 s_{ij}。对于其中一个实体，本章通过从文库中其他包含该实体的语句中挖掘相关信息，丰富对该实体的认识。文库中包含该实体的其他语句被称为参考语句。在这个过程中，计算目标语句与每一条参考语句的语义相似度，用来评价参考语句与目标语句讲述内容是否类似。通过将语义相似度作为从参考语句中获取信息的权重，实现了对邻居信息的筛选，防止不相关或相关性较低信息对关系抽取精度的影响。

为了准确预测句子 s_{ij} 中实体 e_i, e_j 的关系类别标签，知识挖掘器需要从知识图中挖掘关于两个实体(e_i, e_j)的全局信息。为了在汇总信息的过程中有效筛选出相关性比较高的信息，降低不相关或相关程度较低信息对精度的影响，本章设计了一种软筛选方式。简单来说，分析其他同样包含 e_i 或 e_j 的句子语义，依据这些句子与当前句子 s_{ij} 语义相似程度，为这些句子赋予不同的权重，并根据这些权重丰富实体的特征。具体如下，令 $N(e_i)$为 e_i 在 ERG 中的邻居顶点集合，对于$e_k \in N(e_i)$，s_{ik} 是包含实体对(e_i, e_k)的句子。在这里本章引入了一个语义相似度 α_{ik} 来衡量句子 s_{ik} 对句子 s_{ij} 的影响权重，如公式(3-4)所示：

$$\alpha_{ik} = \frac{\exp(\cos(h_{s_{ij}} h_{s_{ik}}))}{\sum_{e_m \in N(e_i)} \exp(\cos(h_{s_{ij}} h_{s_{im}}))} \tag{3-4}$$

其中，$h_{s_{ij}}$是句子s_{ij}在神经网络中获得的语义信息的状态特征，$cos(\cdot,\cdot)$表示两个向量的余弦相似度。

为了从实体邻居中挖掘相关信息来丰富对实体的认知，本章设计了一种基于语义相似度的图神经网络，根据公式(3-4)获得的语义相似度作为权重在 ERG 中聚合邻居信息，得到经过图神经网络之后实体 e_i 的隐状态$h_{e_i}^{(GNN)}$，如公式(3-5)所示：

$$h_{e_i}^{(GNN)} = W^{(GNN)}(h_{e_i} \oplus \sum_{e_k \in N(e_i)} \alpha_{ik} h_{e_k}) + b^{(GNN)} \tag{3-5}$$

其中⊕表示拼接操作。$W^{(GNN)}$是可学习的权重矩阵，$b^{(GNN)}$是偏移。

最后，为了得到实体对的特征用于关系抽取，ERGSRE 将句子 s_{ij} 中的两个实体 e_i, e_j 经过图神经网络之后的特征拼接在一起用来表示实体对(e_i, e_j)的特征$h_{(e_i,e_j)}$，如公式(3-6)所示：

$$h_{(e_i,e_j)} = h_{e_i}^{(GNN)} \oplus h_{e_j}^{(GNN)} \tag{3-6}$$

其中，$h_{e_i}^{(GNN)}$和$h_{e_j}^{(GNN)}$分别表示实体 e_i, e_j 经过图神经网络之后的特征。

假设对于目标语句 s_{12}，知识图中有 6 条语句与 s_{12} 具有公共实体。本章将这些语句中的实体作为 ERG 的顶点，连接具有关系的实体对，构建如图 3.2 所示的 ERG 表示实体之间的联系。

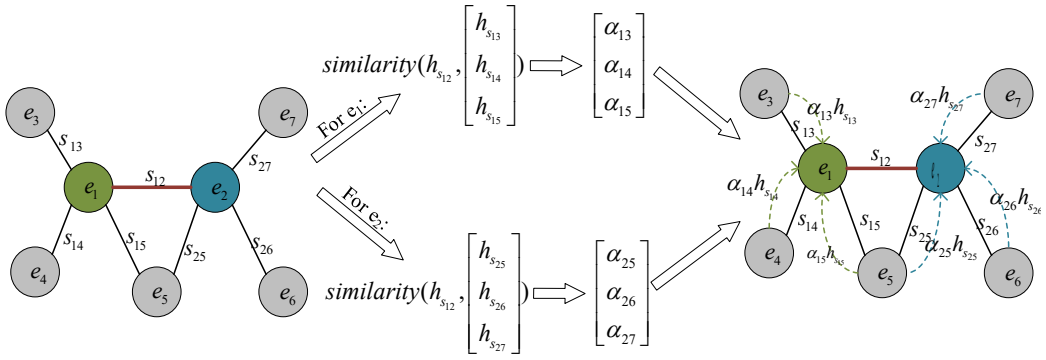

图3.2 基于语义相似度的图计算方式示意图

如图 3.2 所示，当要预测一对实体间关系时，ERGSRE 通过带权重的图神经网络从构建好的 ERG 中获取全局信息。具体地，对于 e_1，$\{e_3, e_4, e_5\}$是 e_1 的邻居集合，记为 $N(e_1)$。首先采用公式(3-4)计算 e_1 和其邻居之间的影响权重分别为 α_{13}, α_{14} 和 α_{15}。然后，采用图神经网络聚集邻居的有效信息。通过公式(3-5)，得到了$h_{e_1}^{(GNN)} = W^{(GNN)}\left(h_{e_1} \oplus \left(\alpha_{13}h_{e_3} + \alpha_{14}h_{e_4} + \alpha_{15}h_{e_5}\right)\right) + b^{(GNN)}$，$h_{e_1}^{(GNN)}$是 e_1 经过了图神经网络之后的状态表示，其中包含了从其它语句中学习到的 e_1 的相关知识。相似地，对 e_2 采取同样的操作得到了包含从 e_2 邻居中学习到的知识信息的最终状态表示$h_{e_2}^{(GNN)}$。通过上述过程，从 e_1，e_2 邻居中分别学习到了仅从 s_{12} 中无法了解到的知识，有助于加强对实体 e_1，e_2 的了解，从而可以更精确地预测这对实体之间的关系。

从上面的过程，很明显可以看出，虽然文库中所有实体构建了一张庞大的知识图，但是当抽取一条语句中实体对之间语义关系时，计算只涉及到这两个实体各自的邻居，是一种局部小体量的计算，保证了网络模型的效率。此外，考虑到二阶邻居实体顶点包含信息与待分类关系实体对的相关性较低，不能有效提高关系抽取精度，甚至会导致无效信息损害精度，所以 ERGSRE 仅使用了一层图神经网络，获取与待分类关系实体对最相关的信息，丰富对待分类实体对的了解，提高关系抽取精度。

3.3.3 关系抽取器

由于关系抽取可作为分类任务处理，因此，本章使用 softmax 分类层作为关系抽取器来预测实体对(e_i, e_j)属于关系(e_i,r,e_j)的概率。如公式(3-7)所示：

$$p(r_m \mid s) = \text{softmax}(\sigma(W * h_{(e_i,e_j)} + b)) \tag{3-7}$$

其中 softmax 分类层的输入是实体对的状态表示$h_{(e_i,e_j)} \in R^{2d^g}$，$d^g$为图神经网络中的输出特征向量维度，$W \in R^{k\times 2d^g}$是可训练的参数矩阵，$b \in R^k$为偏移项，$k$为样本数据中的关系类别数，$\sigma$为激活函数。

最后根据样本在各个关系中的概率分布，选择最大概率的关系作为目标实体对(e_i, e_j)的预测标签r：

$$r = \arg\max_{r_m}(p(r_m \mid s)) \tag{3-8}$$

其中$p(r_m|s)$表示句子中的实体对属于关系r_m的概率。

3.4 算法的训练与预测

ERGSRE 使用的损失函数为分类任务中常用的交叉熵损失函数，可表示为：

$$L = -\sum_j \hat{r} \log p(r_m \mid s, \theta) \tag{3-9}$$

其中，θ代表模型中的所有参数，$\hat{r}$代表样本的真实标签。

为了计算网络参数θ，本章使用 Adam [50]优化算法最小化交叉熵损失 L。对网络参数进行随机初始化，并采用反向传播算法进行更新。为解决过拟合问题，本章在 ALBERT 模块以及图神经网络模块都使用了 dropout 机制[51]，并将 dropout 的概率设置为 0.5。

本章提出的ERGSRE算法主要由特征提取器、知识挖掘器以及关系抽取器组成，经过对三个模块进行多次迭代训练至收敛后，停止训练，此时输入预测数据进行预测。对于一个数据集 D，在输入模型前需将其分为训练数据D_{train}、验证数据D_{val}以及预测数据D_{test}，训练数据D_{train}用于训练模型，验证数据D_{val}是训练过程中的测试集，是为了边训练边看到训练的结果，及时判断学习状态，该部分不是必须的，比例也可以设置很小。预测数据D_{test}是训练模型结束后，用于评价模型结果的测试集。整个算法的伪码如图 3.3 所示。

对图 3.3 所示的 ERGSRE 算法的伪码进行介绍，首先进行算法参数随机初始化（第 2 行），然后对训练数据进行迭代训练（第 3-13 行），其中在每次迭代时，先根据训练数据训练模型，并采用 Adam 优化算法优化训练参数（第 4-8 行），并在每次迭代后验证模型精度，打印相关信息以便及时了解模型训练情况（第 9-12 行），当模型训练结束后将测试数据集输入模型来预测测试数据的标签（第 4-17 行）。

算法 1 ERGSRE
输入：训练集D_{train}，验证集D_{val}，测试集D_{test}，总迭代次数 $epoch$，模型参数θ
输出：测试数据的标签 R
1：**begin** 2：　随机初始化模型参数θ 3：　**for** $iter \in epoch$ **do** 4：　　**for** $d_{train} \in D_{train}$，**do** 5：　　将d_{train}输入模型得到训练数据的标签 r 6：　　根据公式(3-9)计算损失 loss 7：　　计算梯度∇，反向传播梯度并优化参数θ 8：　　**end for** 9：　　**for** $d_{val} \in D_{val}$ **do** 10：　　将d_{val}输入模型得到验证数据的标签 r 11：　　计算模型精度 acc，打印输出信息 12：　　**end for** 13：　**end for** 14：　**for** $d_{test} \in D_{test}$，**do** 15：　　将d_{test}输入模型得到测试数据的标签 r 16：　　$r \rightarrow R$ 17：　**end for** 18：**end**

图3.3 ERGSRE 算法伪码示意图

图 3.4 展示了在输入数据后模型的详细计算过程，即图 3.3 中伪码的第 5、10、15 行：

算法 2 模型计算过程
输入：迭代次数 $iter$，数据 d
输出：关系标签 r
1：**begin** 2：　**if** $iter = 0$ **then** 3：　　建立实体关系图 ERG 4：　**else** 5：　　$h_{s_{ij}}, h_{e_i}, h_{e_j} \leftarrow$ 特征提取器(d) 6：　　$h_{(e_i,e_j)} \leftarrow$ 知识挖掘器($h_{s_{ij}}, h_{e_i}, h_{e_j}, SRG$) 7：　　$r \leftarrow$关系抽取器($h_{(e_i,e_j)}$)

图3.4 模型计算过程伪码示意图

其中，当首次训练时需要根据训练数据所建立实体关系图（第 2-3 行），在建图时，将实体作为顶点，若两个实体之间有联系，即在同一个句子中出现，则将其连接起来，因此，图中的边可用于描述两个实体间的联系，可以用这两个实体所在的句子来表示。建立好实体关系图后，在之后训练时，先根据特征提取器提取输入句子的语义特征$h_{s_{ij}}$

以及句子中所包含实体的特征h_{e_i}, h_{e_j}（第 5 行）。再使用基于语义相似度的图计算方式从邻居中聚合更多的实体信息，从而丰富实体的表示，具体计算过程如公式（3-4），公式（3-5）所示，之后，根据公式（3-6）将聚合后的两实体特征拼接用于表示实体对的特征$h_{(e_i,e_j)}$（第 6 行）。最后根据实体对特征利用关系抽取器进行关系分类，得到实体对间的关系 r（第 7 行）。

3.5 实验与分析

本节将对本章所提出的基于实体关系图的有监督关系抽取算法 ERGSRE 进行性能评测，通过在关系抽取数据集 Sem-Eval 2010 Task 8[52]和 TACRED[53]上与经典的以及当前最新最好的有监督关系抽取算法进行对比，来证明本章提出的最优算法 ERGSRE 可以在多种情况下提升关系抽取性能，此外还进行了消融实验，以验证实体关系图对模型的贡献。

3.5.1 实验数据集

本章在 Sem-Eval 2010 Task 8[52]和 TACRED[53]两个真实数据集上评价 ERGSRE 算法的性能。数据集的详细介绍如下文所示。

Sem-Eval 2010 Task 8：该数据集包含 9 种语义关系类型和一种人工关系类型（*Other*），不属于 9 种语义关系类型的都可认为是 *Other* 关系。这九种关系类型分别是因果关系（*Cause-Effect*）、部分与整体关系（*Component-Whole*）、包含与被包含关系（*Content-Container*）、实体目的地关系（*Entity- Destination*）、实体来源关系（*Entity-Origin*）、工具代理关系（*Instrument-Agency*）、成员集合关系（*Member-Collection*）、消息主题关系（*Message-Topic*）和产品生产者关系（*Product-Producer*）。该数据集中有 10717 个句子，每个句子包含两个实体 e_1 和 e_2，以及句子中相应的关系类型。句子中的关系是有向的，例如（e_1, *Component-Whole*, e_2）与（e_2, *Component-Whole*, e_1）被认为是不同的关系，因此在实际应用中，该数据集被认为有 2*9+1=19 种标签类型。该数据集共有 8000 训练样本以及 2717 测试样本。

TACRED：它是一个大规模的有监督关系抽取数据集，包含 106264 个样本，这些样本从新闻和 web 文本中获取，这些样本来自年度 TAC 知识库群体（TAC-KBP）挑战中使用的语料库。TACRED 中的实例涵盖了 TAC KBP 挑战中使用的 41 种关系类型（例如：*per:schools_attended* 和 *org:members* 等），如果没有关系，则标记为 *no_relation*，因此该数据集共有 41+1=42 种标签类型。数据集被划分为 68124 个训练实例、22631 个验证实例和 15509 个测试实例。

两个数据集的参数如表 3.1 所示，其中 Sem-Eval 2010 Task 8 未提供验证数据，

因此本章在训练阶段将从训练数据中随机抽取 10%作为验证数据。

表3.1有监督关系抽取数据集参数表

数据集 / 参数	Sem-Eval 2010 Task 8[52]	TACRED[53]
关系数量	19	43
训练数据样本数量	8000	68,124
验证数据样本数量	/	22,631
测试数据样本数量	2717	15,509
总样本数量	10717	106264

注：“/”表示 Sem-Eval 2010 Task 8 数据集中未提供验证数据。

3.5.2 算法性能对比指标

在分类任务中常用的性能评价指标有 *precision*、*recall* 以及 *F*1-*score*，在介绍这些指标之前，需先定义 *TP*、*FP*、*TN*、*FN* 四种分类情况，如表 3.2 混淆矩阵所示：

表3.2 混淆矩阵

真实情况	预测结果	
	正例	反例
正例	*TP*（真正例）	*FN*（假反例）
反例	*FP*（假正例）	*TN*（真反例）

假设对于某种关系 *r*，若样本标签为 *r* 记为正例，标签不为 *r* 则记为反例。其中 *TP*（真正例）表示预测为正例，真实情况也为正例；*FN*（假反例）表示预测为反例，但真实情况为正例；*FP*（假正例）表示预测为正例但真实情况为反例；*TN*（真反例）表示预测为反例真实情况也为反例。

准确率（*precision*）代表预测存在某类关系的样本中预测正确的样本所占的比例，也可称之为查准率，如公式(3-10)所示：

$$precision = \frac{TP}{TP + FP} \tag{3-10}$$

其中，*TP*，*FP* 如表 3.2 所示，分别为样本中真正例与假正例的数量。

召回率（*recall*）指在实际存在某种关系的样本中预测正确的样本所占比率，也可称之为查全率，如公式(3-11)所示：

$$recall = \frac{TP}{TP + FN} \tag{3-11}$$

其中，*TP*，*FN* 如表 3.2 所示，分别为样本中真正例与假反例的数量。

F1-score 通过准确率和召回率计算得到：

$$F1\text{-}score = \frac{2 * precision * recall}{precision + recall} \tag{3-12}$$

本章遵循数据集官方任务对衡量标准的设定。对于 Sem-Eval 2010 Task 8 和 TACRED 数据集，采用 *F*1-*score* 来衡量模型性能，*F*1-*score* 值越高，表示算法的性能越好。

3.5.3 实验参数设置

本章所提出的ERGSRE算法基于Pytorch1.2实现，对应的python版本为Python3.6，所有的实验在 Linux 环境下运行，CPU 版本为 Intel(R) Xeon(R) Gold 5115 CPU @ 2.40GHz，GPU 版本为 NVIDIA Tesla P40。

本章使用 $ALBERT_{xxlarge}$ 模型[47]对输入句子和目标实体进行特征提取。为了避免训练过程中的过度训练，采用了 dropout 策略。本章在实验中使用的所有参数设定如表 3.3 所示。

表3.3 不同数据集模型参数设置表

数据集 / 模型参数	Sem-Eval 2010 Task 8	TACRED
预训练模型(*base_model*)	$ALBERT_{xxlarge}$	$ALBERT_{xxlarge}$
学习率(λ)	$2e^{-5}$	$2e^{-5}$
训练批次大小(*batch_size*)	16	48
最大句子长度(l_{max})	128	128
总迭代次数(*epoch*)	10	10
GNN 隐状态维度(d^g)	4096	4096
Dropout 概率(*drop*)	0.5	0.5

本文的超参设置是通过多次实验取最优结果时的参数，由于 TACRED 数据集的样本远远超过了 Sem-Eval 2010 Task 8 的样本数量，并且 *no_relation* 标签占据了 TACRED 整个训练数据集的 70%以上，因此如果训练批次太小，那么很多批次中只

有 *no_relation* 训练数据，这将严重降低训练的准确性。因此，TACRED 数据集的 *batch_size* 需设置地较大。

3.5.4 对比算法简介

为了评价 ERGSRE 模型的性能，本章选择了以下几个经典算法以及最新最好的算法作为对比算法在上述两个数据集上进行对比实验，选取的对比算法如下所示：

CNN[11]：该算法于 2014 年被 Zeng 等人提出，首次使用 CNN 模型进行关系抽取。

Student-R[12]：2020 年 Zhang 等人首先设计一个二部图来发现基于整个语料库的实体和关系之间的类型约束。然后，将这种类型约束与神经网络相结合，得到知识化的模型。该模型被视为教师，通过知识提炼来生成信息丰富的软标签，并指导学生网络的优化。此外，还引入了一种注意机制来帮助学生从文本中挖掘潜在信息。

Att-BLSTM[13]：2016 年 Zhou 等人使用 Attention 与双向 LSTM 模型相结合的方式解决关系抽取问题。

R-BERT[14]：2019 年 Wu 等人将关系抽取作为下游任务，通过对预训练模型 BERT 的微调在 Sem-Eval 2010 Task 8 数据集上进行关系抽取，取得了较好结果。

$BERT_{EM}$[15]：2019 年 Livio 等人修改 BERT 输入格式，微调 BERT，并结合新的自训练方式 MTB，在 Sem-Eval 2010 Task 8 和 TACRED 数据集上取得了较大突破。

AGGCN[17]：2019 年 Guo 等人基于句子层面建立图神经网络，利用两层 GCN 模型学习句子的语法结构，并设计软剪枝策略筛选有效信息，并在 Sem-Eval 2010 Task 8 和 TACRED 数据集上进行了验证。

ERNIE[18]：2019 年 Zhang 等人从知识图谱中引入结构化知识，并将其与 BERT 中学习的语义相结合进行关系抽取。

EPGNN[19]：2019 年 Zhao 等人根据 Sem-Eval 2010 Task 8 数据集建立实体对图，并使用 GCN 学习实体对间的联系。

LST-AGCN[20]：2020 年 Sun 等人提出了一种可学习的语法传输注意图卷积网络来使网络自动学习句子的语法依赖树，以学习完整的语法结构。并在 Sem-Eval 2010 Task 8 和 TACRED 数据集上进行对比实验，在其对比算法上取得了较大提升。

以上算法中，CNN[11]和 Att-BLSTM[13]算法是两个经典算法，其余算法都是近两年具有代表性的算法，其中相比其他算法，EPGNN[19]算法的性能最好，是目前最新最好的算法。

3.5.5 实验结果与分析

1）算法在真实数据集上的性能对比

本小节对本章提出的 ERGSRE 算法在 Sem-Eval 2010 Task 8 以及 TACRED 数据集上与对比算法进行性能对比，并使用 *F*1-*score* 指标评价算法的性能，各算法的 *F*1-*score* 值如表 3.4 所示：

表3.4 ERGSRE 在真实数据集上与各算法实验结果对比表

算法	**Sem-Eval 2018 Task 8**	**TACRED**
	***F*1-*score* (%)**	
CNN[11](2014)	82.7	59.2
Student-R[12](2020)	86.8	69.6
Att-BLSTM[13](2016)	84.0	63.3
R-BERT[14](2019)	89.3	70.7
$BERT_{EM}$[15](2019)	89.5	71.5
AGGCN[17](2019)	85.7	69.0
ERNIE[18](2019)	84.3	67.9
EPGNN[19](2019)	90.2	72.4
LST-AGCN[20](2020)	86.0	68.8
ERGSRE (ours)	**91.7**	**74.6**

表 3.4 展示了本章提出的 ERGSRE 算法在 Sem-Eval 2010 Task 8 以及 TACRED 数据集上与各对比算法的 *F*1-*score* 对比，从表中可以得出以下结论：

(1) 在 Sem-Eval 2010 Task 8 数据集上，ERGSRE 的 *F*1-*score* 高于其他所有算法，与经典的基于 CNN、LSTM 的算法相比，ERGSRE 的性能提升了 7.7%-9.0%，这是由于 ERGSRE 算法相比基于传统 CNN、LSTM 的算法[11,13]，对单词以及句子的特征抽取地更加准确。与基于图神经网络[12,17,20]或预训练模型的算法[14,15]相比，ERGSRE 由于从引入了额外知识，性能提升了 2.2%-6%。与最新最好的 EPGNN 算法[19]相比，同样是根据数据所在语料库建图，但是由于 ERGSRE 的建图方式更加合理，并且基于语义相似度的图计算方式可以筛选有效信息，因此 ERGSRE 的关系抽取性能更好。

(2) ERGSRE 算法在 TACRED 数据集上的性能远远超过了其他算法，与最新最好的 EPGNN[19]算法相比 ERGSRE 的 *F*1-*score* 提升了 3.1%。由于 TACRED 数据集中的样本数量、关系类别数远多于 Sem-Eval 2010 Task 8 数据集，因此 TACRED 数据集关系抽取的难度将远远高于 Sem-Eval 2010 Task 8 数据集，但是由于 ERGSRE 算法从数据集中有效挖掘知识能力较强，且能筛选出

无效知识，即使数据集规模很大，依然能使关系抽取结果取得较大提升。

2）消融实验

本小节将通过消融实验验证实体关系图和基于语义相似度的图神经网络对于分类精度的贡献程度。本章在数据集 Sem-Eval 2018 Task 8 和 TACRED 上，设计了以下实验：

(1) ERG_no：不使用知识挖掘器，直接将特征提取器 ALBERT 中提取的实体特征输入至关系抽取器中进行关系抽取；

(2) ERG_1：使用一层 GNN 结构作为知识挖掘器，在进行图计算时，直接从顶点的一阶邻居中学习知识，即本章采用的方法。

(3) ERG_2：使用两层 GNN 结构作为知识挖掘器，在进行图计算时，从顶点的一阶和二阶邻居中学习知识。

表3.5 知识挖掘器消融实验结果对比表

算法	**Sem-Eval 2018 Task 8**	**TACRED**
	***F*1-*score* (%)**	
ERG_no	89.6	72.1
ERG_1	**91.7**	**74.6**
ERG_2	91.3	74.1

表3.6 仅使用 ALBERT 与 BERT 模型的实验结果对比表

算法	**Sem-Eval 2018 Task 8**	**TACRED**
	***F*1-*score* (%)**	
ERG_no	89.6	72.1
$BERT_{EM}$[15](2019)	89.5	71.5

表 3.5 展示了上述三组实验在 Sem-Eval 2010 Task 8 数据集以及 TACRED 数据集上的 *F*1-*score*。通过 ERG_no 和 ERG_1 实验结果对比可以发现，在加入了知识挖掘器后，关系抽取的性能明显提升，这是由于通过对数据所在语料库建图，可以将有联系的实体连接起来，经过基于语义相似度的图神经网络的计算，可以从其他相关语句中更加全面地了解实体，充分了解实体后，可以更加准确地抽取实体间的关系。通过 ERG_1 与 ERG_2 实验对比可以了解到，使用一层 GNN 结构直接从实体的一阶邻居中挖掘知识比两层 GNN 结构效果要好，这是由于顶点的二阶邻居包含信息与待分类关系实体对的相关性较低，如果从二阶邻居中学习知识，虽然通过语义相似度可以很大程度地过滤不相关信息，但不可避免地会对结果产生轻微影响，进而降低关系抽取

的精度。此外，使用两层 GNN 结构后，模型的计算时间将大幅度增加，因此本章在知识挖掘器中只使用一层 GNN 结构。

此外，通过表 3.6 中 ERG_no 与 BERT_{EM} 的 *F*1-*score* 值对比可以得到，仅使用 ALBERT 模型比使用 BERT 模型进行关系抽取的效果略好，但性能提升程度较小，因此可以说明，MMDSRE 算法性能的提升主要依靠实体关系图所挖掘的知识信息。

3.6 本章小结

本章首先对基于有监督的关系抽取算法的现有缺陷进行总结，进而引出 ERGSRE 算法的设计动机，接着展示了算法的基本框架，并介绍了算法每个模块的功能，然后详细介绍了算法的设计策略，对算法的特征抽取器、知识挖掘器以及关系抽取器的设计方法进行了详细阐述，之后对算法的训练与预测步骤进行了描述，并给出了相关伪代码，最后，通过在真实数据集上与当前最新最好的算法进行对比，证明 STRGSRE 在关系抽取方面的优势，并通过消融实验证明了知识挖掘器的有效性。

第四章 MMDSRE：基于边距度量的远程监督关系抽取算法

本章首先介绍了现有的远程监督关系抽取算法的不足，进而引出本章算法的设计动机。之后，详细论述了所提出一种新的基于边距度量的远程监督关系抽取算法MMDSRE(A **M**argin **M**etric for **D**istantly **S**upervised **R**elation **E**xtraction)。最后，在标准测试数据集上，通过仿真实验与当前流行的远程监督关系抽取算法进行了对比，实验结果表明：所提出的MMDSRE算法，其降噪方式更加有效，所抽取的关系精度更高。

4.1 算法动机

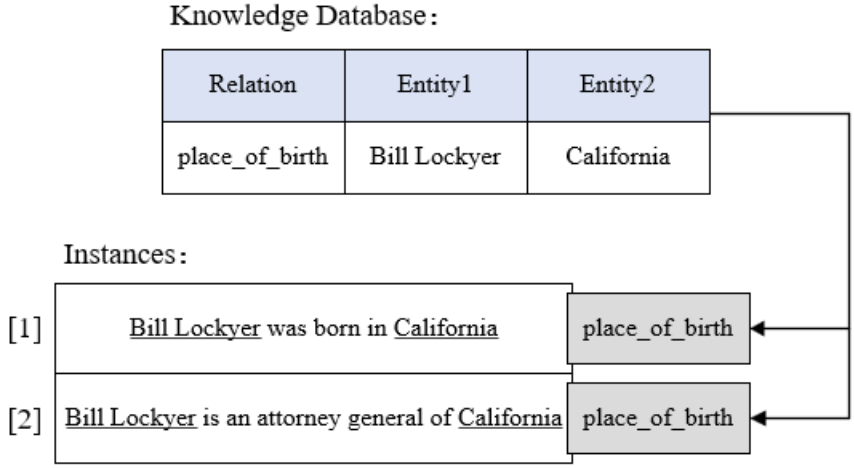

图4.1 远程监督关系抽取示意图

基于远程监督的关系抽取方法由于需要较少的人工标注工作而越来越受欢迎。其通过现有的知识库（例如常用的 FreeBase）已存在的关系三元组来从互联网语料中进行启发式对齐，即只要语料中出现两个目标实体，便将其认为是对应知识库中的关系。如图 4.1 所示，如果在知识库中存在由实体“Bill Lockyer”、“California”以及关系“place_of_birth”组成的三元组<Bill Lockyer , place_of_birth , California >。那么所有包含实体“Bill Lockyer”和“California”的句子都将被标记为具有“place_of_birth”关系。这种做法很容易产生标记错误的现象，如图 4.1 中的示例[2]，虽然句子中的两个实体不具有“place_of_birth”的关系，但是由于该句中包含“Bill Lockyer”和“California”两个实体，远程监督关系抽取方法便将这两个实体的关系标记为“place_of_birth”关系。因此远程监督的方式标注的数据包含大量错误标签，这些具有错误标签的数据被称为噪声，远程监督关系抽取的难点在于如何正确识别出噪声数据，避免其影响模型的训练。

通过对现有的基于远程监督的关系抽取算法的研究与分析，基于选择性的远程监督算法虽然设计简单，但是由于不能解决句包所有句子都是噪声的情况，具有较大局限性，因此本章采用降噪的方式，现有的基于降噪的远程监督关系抽取算法普遍存在以下缺陷：1）特征提取不完全：现有的方法大都使用 PCNN 算法作为特征提取器提取实体以及句子特征，但是由于 PCNN 中的核心模型 CNN 受卷积核大小限制，难以从较长句子中准确提取特征；2）噪声识别精度有待提高：现有的降噪方式采用强化学习或基于匹配的方法识别噪声，但是由于强化学习消耗大量计算资源且很难寻找最优决策，而基于匹配的方法往往需要先验知识，因此当前缺乏一种简单高效的识别噪声的方式；3）难以从噪声数据中挖掘有用信息：现有方法在识别出噪声数据后往往选择将噪声数据丢弃，但是如果能够充分利用噪声数据，将会对关系抽取的精度有很大提升。

本章针对以上三点不足，提出一种基于边距度量降噪的远程监督关系抽取算法 MMDSRE，MMDSRE 算法利用 PCNN 与图神经网络相结合的方式作为特征提取器，可以充分发挥图神经网络的优势，以弥补 PCNN 在特征提取方面的不足。在识别噪声时，本章设计并提出一种基于边距度量的降噪方式（4.3.2 节所示），可以自动地准确识别出噪声数据。最后，MMDSRE 算法将噪声数据作为负样本，与正确标注的数据进行对比训练以提升关系抽取的精度。

4.2 算法框架

基于降噪方法的远程监督关系抽取算法的关键在于如何准确提取句子的语义信息，并精确识别出噪声数据，有效利用噪声数据提升关系抽取的精度，因此本章利用图神经网络学习句子结构来丰富句子语义信息，此外，为精确识别出噪声数据，本章欲设计出一种可以自动识别噪声的噪声识别器，最后，为充分利用噪声数据中的有用信息，本章将噪声数据与正确标注的数据进行对比训练，以学习正负样本在类别空间的分布特征，并准确抽取目标实体对的关系。基于以上思想，本章设计并提出了一种基于边距度量的远程监督关系抽取算法 MMDSRE（A **M**argin **M**etric for **D**istantly **S**upervised **R**elation **E**xtraction），算法的模型框架如图 4.2 所示。

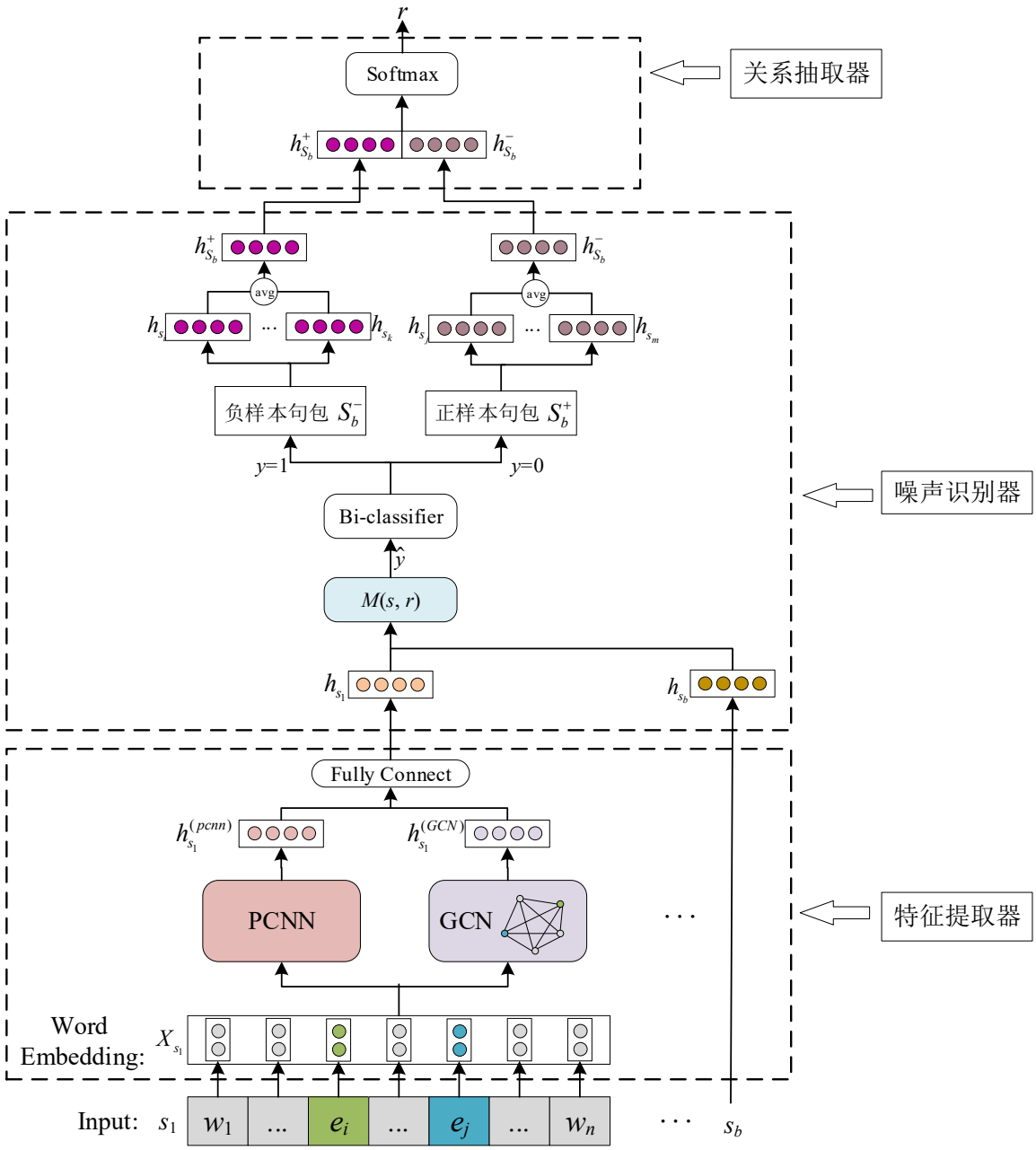

图4.2 MMDSRE 算法模型框架示意图

如图 4.2 所示，本章提出的 MMDSRE 算法模型主要由特征提取器、噪声识别器以及关系抽取器组成，各模块功能如下所示：

特征提取器：本章采用的特征提取器主要由 PCNN[30]和 GCN[48]组合而成，利用 PCNN 提取句子中单词的相对位置信息以及语义信息，同时引入 GCN 补充句子的语法信息并进一步丰富其语义信息，最后将 PCNN 及 GCN 学习的句子表示相结合来作为句子的特征表示。

噪声识别器：本章设计的噪声识别器是根据特征提取器提取的句子特征识别出噪声数据，并将噪声作为负样本生成负样本表示，同时根据正样本特征生成正样本表示。

关系抽取器：根据句包的正样本与负样本表示对给定的句包中的实体对进行关系分类，抽取出其中蕴含的关系。

4.3 算法主要设计策略

4.3.1 特征提取器

在进行关系抽取时，首先需使用特征提取器提取句子的语义特征，而远程监督关系抽取算法需要一种轻量型特征提取器来快速准确提取句子特征，为后续噪声的识别提供基础。本文第三章基于 ALBERT 的特征提取器以及引入知识的方式虽然对句子的特征学习较为准确，但是训练模型所耗费的时间较长，难以满足远程监督关系抽取的需要，因此本章需设计一种轻量型的特征提取器。本章首先使用词嵌入技术将一个句子转换成一个矩阵，其中包含单词嵌入和位置嵌入。然后，使用分段卷积神经网络（PCNN）[30]根据单词和位置嵌入获得最终句子表示。此外，为学习句子的结构信息，丰富句子语义表示，将句中每个单词作为顶点建立全连接图，并利用 GCN 学习句子的拓扑结构挖掘其语法信息，并以此丰富句子的语义表示。

给定一个包含 b 个句子的句包 S_b，$S_b = \{s_1, s_2, \ldots, s_b\}$，这 b 个句子都包含了相同的两个实体 e_1, e_2，远程监督关系抽取的目的是预测出 S_b 中包含的实体对间的关系 $(e_1,r,e_2), r_i \in R = \{r_1, r_2, \ldots, r_k\}$。每个句子 s_i 都是由一系列单词组成，$s_i = \{w_1, w_2, \ldots, w_n\}$，$n$ 句子中单词个数。在一个句子中，每个单词 w_i 首先需要使用词嵌入技术被映射到一个 d^w 维向量 v_i 中。此外，相对位置是关系抽取的一个重要特征，它可以为下游神经模型提供丰富的位置信息，本章采用了 Zeng 等人[30]提出的位置特征(PFs)来指定目标实体对，使模型更加关注靠近目标实体的词。PFs 明确地描述了每个单词 w_i 与两个目标实体 e_1 和 e_2 之间的相对距离。对于第 i 个单词，使用一个随机初始化的权重矩阵将相对位置特征投影到表示头部和尾部实体的两个向量中，即 $p_i^{e_1}$ 和 $p_i^{e_2}$。每个单词 w_i 的最终表示 x_i 是单词嵌入和两个位置嵌入的连接，即 $x_i=[v_i; p_i^{e_1}; p_i^{e_2}]$。因此，句子的初始表示如公式(4-1)所示：

$$X_s = \{x_1, x_2, \cdots, x_n\} \tag{4-1}$$

其中，x_i 为单词的特征表示，n 为句子的长度。

PCNN：MMDSRE 算法采用 PCNN[30]学习实体与句子的位置关系以及语义信息，PCNN 主要由一维卷积和分段最大池化组成。一维卷积是权重矩阵 $W^{(pcnn)}$ 和输入矩阵 X_s 之间的运算，如公式(4-2)[30]所示：

$$m_i = W^{(pcnn)T} x_{(i-w+1):i} \tag{4-2}$$

其中 $W^{(pcnn)}$被视为卷积的滤波器，x_i 是与句子中第 i 个词相关联的输入向量，一般来说，将 $x_{i:j}$ 看作是从 x_i 到 x_j 的拼接，w 是滤波器的大小，那么卷积是将向量 W 与 X_s 中每个 w-gram 的点积得到另一个向量 m_i。

在一个句子中，由公式（4-2）得到的 m_i 的个数一共为 n-w+1 个，在本章中，将对每个句子使用元素填充使得 m_i 的个数与句子长度 n 相等。由此得到的卷积结果为一个特征映射矩阵$M = \{m_1, m_2, \cdots, m_n\}$。之后，采用分段最大池化方法捕捉句子的结构信息。卷积层之后，每个特征映射矩阵 M_i 根据两个实体的位置分为三个部分$\{M_{i1}, M_{i2}, M_{i3}\}$。然后，对这三部分分别执行最大池操作，即$p_{ik} = \max(M_{ik})$，将最终得到三段池化结果拼接得到句子的表示，如公式(4-3)[30]所示：

$$h_s^{(pcnn)} = [p_{i1}; p_{i2}; p_{i3}] \tag{4-3}$$

其中，p_{ik}表示由第 i 个词的第 k 段池化结果，$h_s^{(pcnn)}$表示使用 PCNN 模型学习到的句子特征。

GCN：为了挖掘句子的语法结构信息，丰富句子语义，使得特征提取器所提取的特征更加准确，本章在特征提取器中引入了 GCN[48]模型提取每个句子的特征。为此，本章将句子中的每个单词作为顶点建立全连接图，并将每个单词的表示作为顶点的表示进行图卷积运算，计算方式如公式(4-4)[48]所示：

$$h_s^{(GCN)} = GCN(X_s, A) = \sigma(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}X_s W^{(GCN)}) \tag{4-4}$$

其中，X_s如公式（4-1）所示为句子 s 中的单词表示的组合，$W^{(GCN)}$为图滤波器用来挖掘句子结构信息，如 2.2.2 节所述，$\tilde{A}$为添加了自连接的邻接矩阵，$\widetilde{D}$为度矩阵，σ为激活函数。

为了充分获取句子的语义信息，本章将 PCNN 以及 GCN 得到句子表示$h_s^{(pcnn)}$以及$h_s^{(GCN)}$拼接在一起作为句子的特征，使用该方式获得的句子特征不仅包含了 PCNN 捕获的局部特征，也包括了 GNN 学习的结构特征，如公式（4-5）所示：

$$h_s = \sigma(W_3(h_s^{(pcnn)} \oplus h_s^{(GCN)}) + b_3) \tag{4-5}$$

其中 W_3 为可训练的权重矩阵，b_3 为偏移，$\sigma(\cdot)$为激活函数。

根据上述步骤，可到的每个句子 s_i 的最终特征表示，对于一个包含 b 个句子的句包，将每个句子表示组合得到整个句包的表示 H：

$$H = \{h_{s_1}, h_{s_2}, \cdots, h_{s_b}\} \tag{4-6}$$

其中h_{s_i}表示句子 s_i 的特征表示。

4.3.2 噪声识别器

噪声识别器将根据每个句子的特征来判断句子是否正确表示了目标实体中的关系，如果句子不能正确表达句包所携带的标签关系，则认为该句子为噪声数据，因此噪声识别器可用二分类器表示，如果目标句子是噪声，则标记为 1，否则标记为 0。但是，由于对有噪声的数据没有任何明确的监督，很难知道二分类器的分类结果是否正确，因此本章设计了一种基于边距度量的监督方式来训练二分类器。

假设训练数据 D_{train} 由正确标记和错误标记的两种数据组成，错误标记的样本是指样本描述的内容与其携带的标签不匹配的样本，而正确标记的样本有一个与描述事实相匹配的标签。一些正确标记的样本可能是易学的，即很容易学习到句子的特征与关系标签之间的映射关系，另一些由于描述的是罕见的事件，可能是“难以学习”的。一般来说，D_{train} 中容易和困难学习的正确标记样本都能提高模型的泛化能力，而错误标记的样本则会影响泛化能力。本章的目标是通过观察样本间训练动态的差异来识别数据中的错误标记数据。

如何确定训练样本是否有助于或不利于模型泛化，即如何确定数据是正确标记或者错误标记？本章使用了一种基于训练样本边距的度量方式，这是许多机器学习算法的公认概念[54]。使用边距度量的优势在于：1）它在训练过程中可以简单有效的计算出来；2）它们可以自然地跨样本分解，从而有可能估计每个数据点对泛化的贡献。假设$(s,r) \in D_{train}$是训练集中的一个样本，其中 s 表示句子，r 表示其关系标签，$z(s)$是一个 *logits* 向量（每个 *softmax* 的输出），表示样本属于标签 r 的概率，边距 $M(s,r)$表示标签类别的 *logits* 值比其他类别的 *logits* 值大多少，对于边距的计算，如公式(4-7)[54]所示：

$$M(s,r) = z_r(s) - \max_{i \neq r} z_i(s) \tag{4-7}$$

其中$z_r(s)$表示训练样本的标签r所对应的 *logits* 值，而$max_{i \neq r} z_i(s)$表示非标签类别中的最大 *logits* 值。对于一个样本，若正确标记，则其标签类别的 *logits* 值必然大于其他类别的 *logits* 值，即边距 $M(s,r)$为正值，若样本错误标记，则 $M(s,r)$为负值。

对于 MMDSRE 的噪声识别器，给定训练数据中的一个样本(s, r)，首先需要根据特征提取器提取的句子特征h_s计算出每个类别的 *logits* 值，如公式(4-8)所示：

$$Z = \text{softmax}(Wh_s + b) \tag{4-8}$$

其中$W \in R^{k \times d^h}$是可训练的参数矩阵，$b \in R^k$为偏移项，k 为样本数据中的关系类别数，d 为句子的特征维度。

然后从 Z 中分解出样本标签的 *logits* 值$z_r(s)$以及其他最大的 *logits* 值$max_{i \neq r} z_i(s)$，之后根据公式（4-7）计算边距 $M(s,r)$，若 $M(s,r)>0$，则标记为正样本（$\hat{y}=0$），否则标记为负样本（$\hat{y}=1$），如公式（4-9）表示：

$$\hat{y} = \begin{cases} 1 & if\ M(s,r) < 0 \\ 0 & if\ M(s,r) > 0 \end{cases} \tag{4-9}$$

由此，可根据期望标签$\hat{y}$对噪声识别器进行训练，以此来识别数据中的噪声，识别噪声的过程如公式（4-10），公式（4-11）所示：

$$p(y_i \mid s) = \text{softmax}(W^{(Denoise)} * h_s + b^{(Denoise)}) \tag{4-10}$$

$$y = \arg\max_{y_i}(p(y_i \mid s)) \tag{4-11}$$

其中，$W^{(Denoise)} \in R^{2 \times d^h}$为噪声识别器的权重矩阵，$b^{(Denoise)} \in R^{2 \times d^h}$为偏移项，$p(y_i|s)$为句子 s 的噪声标签为 y_i（y_i=0 或 1）的概率。

最后可根据期望标签$\hat{y}$和预测的噪声分布$p(y_i|s)$训练噪声识别器，由于噪声识别器为二分类器，可使用交叉熵损失函数作为噪声识别器的损失函数，如公式(4-12)所示：

$$L^{(Denoise)} = -\sum_j \hat{y} \log p(y_i \mid s, \theta) \tag{4-12}$$

其中θ为算法中的参数，$\hat{y}$为噪声的期望标签，$p(y_i|s)$为句子 s 的噪声标签为 y_i（的概率。

对于包含 b 个句子的句包$S_b = \{s_1, s_2, \ldots, s_b\}$，经过噪声识别器后会将其分为两部分，即正样本句包$S_b^+ = \{s_i | y_i = 0\}$和负样本句包$S_b^+ = \{s_j | y_j = 1\}$，其中$y_k$=0 或 1，为噪声识别器中的二分类器的分类结果。

4.3.3 关系抽取器

利用噪声识别器识别出噪声数据后，如何有效利用这些噪声数据尤为重要，以往的远程监督关系抽取算法大都选择将噪声数据直接丢弃，选择使用正确标注的数据训练关系抽取器。但是这些噪声数据中往往包含着可以利用的信息，如果能够有效利用，便能够提升关系抽取器的精度。因此本章将噪声数据作为负样本，通过正负样本的对比训练使两者在关系标签空间的分布距离更远，以防噪声数据的混淆影响数据的正确

分类。

为了得到正样本句包的特征以及负样本句包的特征，本章采用取平均的方式，将两个句包中的句子特征分别平均后得到正样本句包特征$h_{S_b^+}$及负样本句包表示$h_{S_b^-}$，如公式(4-13)所示：

$$\begin{cases} h_{S_b^+} = \frac{1}{k}\sum_i^k h_{s_i}, s_i \in S_b^+ \\ h_{S_b^-} = \frac{1}{m}\sum_j^m h_{s_j}, s_j \in S_b^- \end{cases} \tag{4-13}$$

其中，k 为正样本句包中的句子数量，m 为负样本句包中的句子数量。

在得到正样本句包表示$h_{S_b^+}$以及负样本句包表示$h_{S_b^-}$后，分别使用公式(4-14)及公式(4-15)计算其句包的条件概率：

$$p(r_i \mid S_b^+) = \text{softmax}(\sigma(Wh_{s_b^+} + b)) \tag{4-14}$$

$$p(r_i \mid S_b^-) = \text{softmax}(\sigma(Wh_{s_b^-} + b)) \tag{4-15}$$

其中$p(r_i|S_b^+), p(r_i|S_b^-)$分别表示正样本句包以及负样本句包中的实体对关系为$r_i$的概率，$W \in R^{k \times d^h}$是可训练的参数矩阵，$b \in R^k$为偏移项，$k$ 为样本数据中的关系类别数，σ为激活函数。

MMDSRE 中同样使用交叉熵损失函数分别计算正负样本句包的损失，如公式(4-16)，(4-17)所示：

$$L^+ = -\sum_i^k \hat{r} \log p(r_i \mid S_b^+, \theta) \tag{4-16}$$

$$L^- = -\sum_i^k \hat{r} \log(1 - p(r_i \mid S_b^-, \theta)) \tag{4-17}$$

其中，L^+表示正样本句包的损失，L^-表示负样本句包的损失，θ代表模型中的所有参数，$\hat{r}$代表句包中的目标实体对的标签。

最后，将正负样本的损失相加得到关系抽取器的损失函数 $L^{(Extract)}$：

$$L^{(Extract)} = L^+ + L^- \tag{4-18}$$

4.4 算法的训练与预测

MMDSRE 算法在最终训练时将噪声识别器以及关系抽取器中的损失相结合进行联合训练，因此，最终的损失函数为：

$$L = L^{(Denoise)} + L^{(Extract)} \tag{4-19}$$

其中 $L^{(Denoise)}$和 $L^{(Extract)}$分别为噪声识别器和关系抽取器的损失。

为了计算网络参数θ，本章使用 Adam [50]优化算法最小化损失 L。对网络参数进行随机初始化，并采用反向传播算法进行更新。为解决过拟合问题，本章在各个模块都使用了 dropout 机制[51]，并将 dropout 的概率设置为 0.5。

在测试阶段，对于每个输入句包，在使用噪声识别器识别出噪声数据并得到正负样本句包表示之后。利用公式(4-14)对正样本句包进行关系分类。

MMDSRE 算法主要由特征提取器、噪声识别器以及关系抽取器三个模块组成，在训练过程中，通过对三个模块的多次迭代组合训练，使得特征提取器可以准确提取输入句包中的句子特征，噪声识别器能够精确识别正负样本，并利用正负样本对特征提取器进行训练，使得其对于正样本能够准确预测出其类别，同时对于容易混淆的负样本也能准确识别出来。与 3.4 节中的有监督关系抽取算法 ERGSER 算法类似，MMDSRE 也将数据分为三部分，分别为训练数据D_{train}、验证数据D_{val}以及预测数据D_{test}。同样在训练过程中使用验证数据D_{val}测试训练效果并及时输出。算法的伪代码如图 4.3 所示.

算法 3 MMDSRE 算法伪码
输入：训练集D_{train}，验证集D_{val}，测试集D_{test}，总迭代次数 $epoch$，模型参数θ
输出：测试数据的标签 R
1：**begin** 2：　随机初始化模型参数θ 3：　**for** $iter \in epoch$ **do** 4：　　**for** $d_{train} \in D_{train}$，**do** 5：　　将d_{train}输入模型得到训练数据的标签 r 6：　　根据公式(4-19)计算损失 loss 7：　　计算梯度∇，反向传播梯度并优化参数θ 8：　　**end for** 9：　　**for** $d_{val} \in D_{val}$ **do** 10：　　将d_{val}输入模型得到验证数据的标签 r 11：　　计算模型精度，打印输出信息 12：　　**end for** 13：　**end for** 14：　**for** $d_{test} \in D_{test}$，**do** 15：　　将d_{test}输入模型得到测试数据的标签 r 16：　　$r \rightarrow R$ 17：　**end for** 18：**end**

图4.3 MMDSRE 算法伪代码示意图

对图 4.3MMDSRE 算法的伪代码进行简要说明，首先进行参数随机初始化（第 2

行），然后对训练数据进行迭代训练（第 3-13 行），其中在每次迭代时，先根据训练数据训练模型，并采用 Adam 优化算法优化训练参数（第 4-8 行），并在每次迭代后验证模型精度，打印相关信息以便及时了解模型训练情况（第 9-12 行），当模型训练结束后将测试数据集输入模型来预测测试数据的标签（第 4-17 行）。

在数据输入模型后，需经过特征提取器、噪声识别器以及关系抽取器才能得到最终的标签（图 4.3 中的算法第 5 行），下面将对模型的三个模块实现细节进行介绍：

正如 4.3.1 节所述，特征提取器由 PCNN 模型和 GCN 模型组合而成，通过单词与实体的相对位置以及句子的拓扑结构等准确提取句子的语义信息，具体步骤如图 4.4 的伪代码所示。

算法 4 MMDSRE 特征提取器实现算法
输入：数据 d
输出：句子特征h_s
1：**begin** 2：　**for** $S_b \in d$ do 3：　　**for** $s \in S_b$ do 4：　　$x_i \leftarrow [v_i; p_i^{e_1}; p_i^{e_2}]$ //根据单词的词嵌入及位置特征得到每个单词的表示 5：　　$X_s \leftarrow \{x_1, x_2, \dots, x_n\}$ //将单词特征组合得到句子的初始表示 6：　　$h_s^{(pcnn)} \leftarrow PCNN(X_s)$ //利用 PCNN 模型得到句子的特征 7：　　$h_s^{(GCN)} \leftarrow GCN(X_s)$ //利用 GCN 模型句子的特征$h_s^{(GCN)}$ 8：　　$h_s \leftarrow combine(h_s^{(GCN)}, h_s^{(GCN)})$ //将两个特征组合得到最终句子特征 9：　　**end for** 10：　**end for** 11：**end**

图4.4 MMDSRE 特征提取器实现算法伪代码示意图

对图 4.4 中的伪码进行简要说明，输入数据 d 是由一个个句包S_b组成，而在特征提取阶段，需要对句包中的每个句子s进行处理，对于每个句子首先需要根据词嵌入以及单词与实体的相对位置得到每个单词的特征 x_i（第 4 行），其次利用单词特征根据公式（4-1）得到整个句子的初始表示（第 5 行），然后分别使用 PCNN 模型以及 GCN 模型根据句子的单词位置、语法结构等学习句子的特征（第 6-7 行），最后将两个模型学到的特征根据公式（4-5）组合得到最终句子的特征表示h_s（第 8 行）。

从特征提取器中得到句子的特征h_s之后，需要利用噪声识别器根据h_s识别句子是否为噪声，在训练阶段，需要依据边距度量的方法（公式 4-7）标记噪声句子，之后训练噪声识别器，在验证或测试阶段，直接使用噪声识别器识别噪声数据。噪声识别器的实现过程如图 4.5 中的伪代码所示。

对图 4.5 中的算法伪码进行简要说明，首先在训练过程中，需要对噪声识别器进行有监督训练，因此本章使用了基于边距度量的方式依据公式（4-7，4-8，4-9）标记

噪声数据的期望标签$\hat{y}$（第 3-4 行），并用噪声识别器预测句子的噪声标签 y（第 5 行），最后根据期望标签$\hat{y}$和预测标签 y 训练噪声识别器（第 6 行）。在验证或测试阶段，由于数据没有标签，难以根据边距度量的方式判断噪声数据，因此需使用噪声识别器识别噪声数据并进行标记（第 7-8 行）。

算法 5 MMDSRE 噪声识别器实现算法
输入：句子特征h_s，*process*(*train*, *val* or *test*)
输出：噪声标记 *y*
1：**begin** 2：　**if** *process* = *train* **do** 3：　　$Z \leftarrow softmax(Wh_s + b)$ //计算样本 logits 值 4：　　$\hat{y} \leftarrow z_r(s) - \max_{i \neq r} z_i(s)$ //根据边距度量的方法计算期望标签$\hat{y}$ 5：　　$p(y_i\|s) \leftarrow softmax(W^{(Denoise)} * h_s + b^{(Denoise)})$ //用噪声识别器预测句子噪声标记 6：　　$L^{(Denoise)} \leftarrow -\sum_j \hat{y} \log (p(y_i\|s))$ //计算 loss 7：　**else if** *process* = *val* or *test* **do** 8：　　$y \leftarrow argmax(softmax(W^{(Denoise)} * h_s + b^{(Denoise)}))$ //预测句子的噪声标记 9：**end**

图4.5 MMDSRE 噪声识别器实现算法伪代码示意图

在区分出噪声数据后，为了有效利用噪声数据来提升关系抽取器的精度，MMDSRE 在训练阶段将噪声数据作为负样本，正确标记的数据作为正样本，通过正负样本的对比训练学习两者在关系标签空间的分布，以防噪声数据对关系抽取器的影响。而在验证和测试阶段，直接使用正确标记的样本进行关系抽取。关系抽取器的实现过程如图 4.6 中的伪代码所示：

算法 6 MMDSRE 关系抽取器实现算法
输入：句子噪声标记 y，句子特征h_s
输出：句包的关系标签 *r*
1：**begin** 2：　**if** *process* = *train* **do** 3：　　生成正样本句包表示$h_{S_b^+}$及负样本句包表示$h_{S_b^-}$。 4：　　依据正负样本句包表示计算条件概率$p(r\|S_b^+)$和$p(r\|S_b^-)$ 5：　　计算 *loss*，训练关系抽取器 6：　**else if** *process* = *val* or *test* **do** 7：　　生成正样本句包表示$h_{S_b^+}$ 8：　　利用$h_{S_b^+}$预测句包的关系标签 *r* 9：**end**

图4.6 MMDSRE 关系抽取器实现算法伪代码示意图

对图 4.6 中的算法伪码进行简要说明，在训练过程中由于需要正样本与负样本对比训练，因此需根据句子的噪声标记以及句子特征分别生成正样本句包表示$h_{S_b^+}$及负

样本句包表示$h_{S_b^-}$（第 3 行），然后利用正负样本的句包表示如公式（4-14，4-15）所示计算条件概率$p(r|S_b^+)$和$p(r|S_b^-)$（第 4 行），之后根据条件概率和句包真实标签计算 $loss$，如公式（4-16，4-17，4-18）所示，并训练关系抽取器（第 5 行）。在进行验证或测试时，使用正样本进行关系预测，因此只需生成正样本句包表示$h_{S_b^+}$（第 7 行），并根据$h_{S_b^+}$预测句包的关系标签 r（第 8 行）

上述算法 4、5、6 即为模型各个模块的实现过程，由此可得到整个模型的实现过程如图 4.7 所示。

算法 7 模型计算过程
输入：句包 S_b，模型参数θ
输出：句包关系标签 r
1：**begin** 2：　初始化模型参数θ 3：　根据算法 4 抽取句包中句子的特征 4：　根据算法 5 识别句包中的噪声句子 5：　根据算法 6 抽取句包中实体的关系 r 6：**end**

图4.7 模型计算过程伪代码示意图

如图 4.7 所示的伪代码为整个模型的计算过程，再输入数据后首先需要进行参数随机初始化（第 2 行），之后将数据输入特征提取器提取句子的语义特征（第 3 行），然后利用噪声识别器识别噪声数据（第 4 行），最后通过关系抽取器来抽取实体的关系 r（第 5 行）。

4.5 实验与分析

为了评估本章提出的 MMDSRE 算法，并将该算法与对比算法进行比较，本节将在一个流行的基准数据集 NYT(New York Times)上进行实验，用于远监督关系提取。在本节中，首先介绍数据集和评测标准，其次，给出了实验设置，最后，将 MMDSRE 的性能与几种最先进的方法进行比较。

4.5.1 实验数据集

NYT：该数据集是远程监督标准数据集，它通过将 Freebase 知识库中的关系与纽约时报语料库（NYT）对齐而形成。使用斯坦福命名实体标记器在文档中寻找涉及的实体，并与 Freebase 实体的名称进一步匹配。训练集由 2005 年和 2006 年的纽约时报新闻组成，测试集由 2007 年的纽约时报新闻组成。该数据集有 53 个关系类，其中包含一个特殊的关系 NA（表示实体对之间没有关系）。训练数据集包括 522611 个句子

和 281270 个实体对。测试数据集包含 172448 个句子和 96678 个实体对。

4.5.2 算法性能对比指标

与之前的远程监督关系抽取工作[7]类似，本章使用准确率-召回率（P-R）曲线、和 top-N 精度(P@N)作为 MMDSRE 在 NYT 测试数据集的度量标准。这些标准的详细介绍如下：

准确率-召回率（P-R）曲线：P-R 曲线是刻画准确率(precision)和召回率(recall)之间关系的一种曲线，其中，准确率和召回率的概念如 3.5.2 节所示。一般而言，准确率和召回率是一对矛盾的度量，准确率越高则召回率越低，准确率越低，召回率往往越高。为衡量算法的综合性能，可对其预测结果进行排序，越靠前边的样本是算法预测为正例概率最大的样本，按照此顺序把样本逐个作为正例进行预测，每次将计算得到的准确率为纵轴，召回率为横轴作图，就得到了准确率-召回率(P-R)曲线。在进行性能衡量时，若一个算法的 P-R 曲线完全“包住”了另一个算法的 P-R 曲线完全，则认为前者的性能优于后者。如果两个算法的 P-R 曲线发生了交叉，一般很难直观上比较两个算法的优劣，此时往往选择比较 P-R 曲线下面积的大小，它在一定程度上表示了算法在准确率和召回率上取得相对“双高”的比例，因此 P-R 曲线下面积越大，表示算法性能越好。

top-N 精度(P@N)：为了降低由于知识库语料不充分导致一些本身存在关系的实体被标记错误的样本（False Negative 样本）对实验结果的影响，人工评估的方法被广泛应用于远程监督的评价标准中。由于 P-R 曲线中随着召回率值不断上升，False Negative 样本会逐渐增多，使得算法预测正确的示例被错误评判，因此常选择一小部分数据用于测试，该方法通过手动统计前 N 个概率最高的关系事实的精确度来判断算法性能，这样含有大量 False Negative 样本的数据就会被排除在外，避免错误评判的问题。P@N 值越高说明算法预测的关系准确度越高，算法性能越好，P@N 值的计算方式如公式（4-20）所示。

$$\mathrm{P@N} = \frac{1}{N}\sum_{i=1}^{N} pre_i \tag{4-20}$$

其中pre_i指将所有样本的 *precision* 值按照从大到小的顺序排序后的第 i 个值，样本的 *precision* 值计算方式如公式(3-10)所示。

4.5.3 实验参数设置

与第三章的实验环境一样，本章所提出的 MMDSRE 算法基于 Pytorch1.2 实现，对应的 python 版本为 Python3.6，所有的实验在 Linux 环境下运行，CPU 版本为 Intel(R)

Xeon(R) Gold 5115 CPU @ 2.40GHz，GPU 版本为 NVIDIA Tesla P40。

表4.1 MMDSRE 算法参数表

算法参数	参数值
学习率(λ)	0.1
训练批次大小(*batch_size*)	160
最大句子长度(l_{max})	128
总迭代次数(*epoch*)	60
隐状态维度(d^h)	230
词嵌入维度(d^w)	50
Dropout 概率(*drop*)	0.1
CNN 卷积核大小(m)	3

为了与对比算法进行公平合理的比较，MMDSRE 的大多数超参数设置遵循对比算法的设置，具体实验参数设置如表 4.1 所示。

4.5.4 对比算法简介

为了评价 MMDSRE 算法的性能，本章选择了以下几个经典算法以及最新最好的算法作为对比算法在 NYT 数据集[29]上进行对比实验，本章所选取的对比算法如下所示：

PCNN [30]：2015 年 Zeng 等人利用卷积神经网络的分段池化方式提取句子语义，并将远程监督关系抽取任务作为多实例学习方法处理。

PCNN+ATT[31]：2016 年 Liu 等人使用 PCNN 模型作为句子特征提取器，之后利用选择性注意力机制选择正确标注的样本以缓解噪声的影响。

HG[38]：2019 年 Duan 等人建立混合图模型学习句子的特征，然后利用选择性注意力机制选择正样本进行关系抽取。

Seg[39]：2020 年 Li 等人使用 PCNN 模型和注意力机制学习句子语义特征，并设计选择性门控网络筛选正确标注的样本。

PCNN+RL[42]：2018 年 Qin 等人利用 PCNN 模型学习句子的特征，并设计强化学习的方法从句包中移除噪声数据。

DCRE[44]：2020 年 Shang 等人将句包中的每个句子与其对应标签进行相似度匹配来识别噪声，之后采用聚类的方法修正噪声数据的标签以有效利用噪声数据的信息。

PCNN+RU[45]：2020 年 He 等人利用强化学习识别句包中的噪声数据后，将噪声数据作为对应关系类别的负样本，与正样本(即标注正确的数据)一起训练，以提升分

类器的精度。

以上算法中，PCNN [30]和 PCNN+ATT[31]算法是两个经典算法，其余算法都是近两年具有代表性的算法，其中相比其他算法，PCNN+RU[45]算法的性能最好，是目前最新最好的算法。

4.5.5 实验结果与分析

1） MMDSRE 算法与对比算法 P-R 曲线对比

本章将 MMDSRE 算法与上述对比算法在远程监督数据集 NYT 上进行对比，如图 4.8 中的 P-R 曲线图所示，根据 4.5.1 节对 P-R 曲线的介绍，算法在曲线下的面积越大，性能越好。

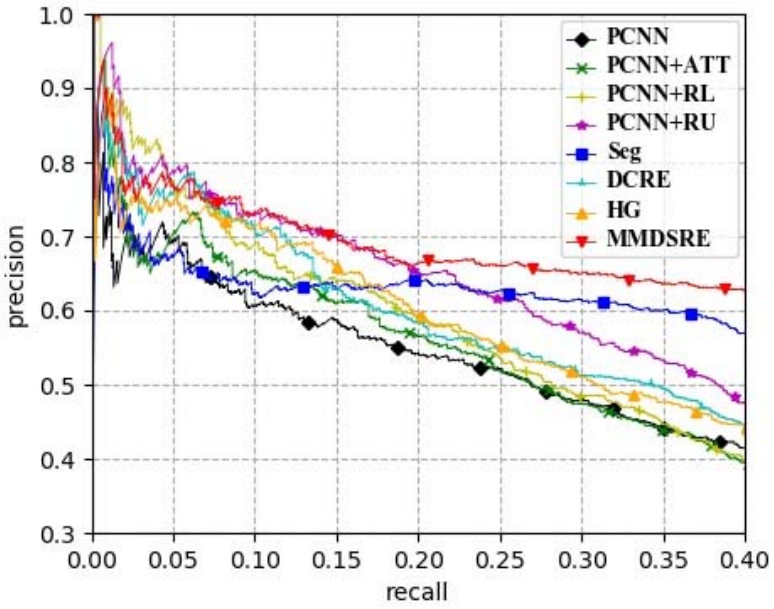

图4.8 MMDSRE 与对比算法 P-R 曲线对比示意图

从图 4.8 中的实验结果中可以明显得出以下结论：

(1) 基于选择性的算法中，Seg 算法由于其选择性门控网络的作用可以缓解一个句包中所有句子都是噪声的情况，因此性能相较于其他三者（PCNN、PCNN+ATT，HG）较好。HG 算法由于使用了图神经网络作为特征提取器，提取的句子特征更加准明确，因此性能优于 PCNN 和 PCNN+ATT 算法。

(2) 基于降噪的算法如 DCRE、PCNN+RL、PCNN+RU 整体上性能要优于基于选择性的算法，这是因为基于降噪的算法通过对噪声的识别，可以有效减小噪声的影响，对于句包中全是噪声的情况可以直接将其预测为“NA”类别，因此，准确率较高。而在这三种基于降噪的方式中，由于 PCNN+RL 算法在识

别出噪声数据后直接选择丢弃，损失了有用信息，因此其性能相较于 DCRE 和 PCNN+RU 算法略低。对比 DCRE 算法和 PCNN+RU 可以明显看出，后者的性能优于前者，这是由于 DCRE 算法使用聚类的方式为噪声数据重新修正标签时，难以衡量标签的准确性，可能导致再次引入噪声数据，而 PCNN+RU 算法将噪声作为负样本可以在训练过程中使正负样本在关系空间的距离更远，提升关系抽取器的精度。

(3) 对比本章提出的 MMDSRE 算法和其他对比算法，可以明显看出 MMDSRE 算法的性能均高于其他算法，由于 MMDSRE 采用降噪的方式将全是噪声的句包预测为“NA”类别，可以解决基于选择性方法的缺陷。与对比算法中最好的算法 PCNN+RU 相比，MMDSRE 算法由于特征提取器使用了 GNN 可以抽取句子结构信息，提取的特征更加完整，并且采用基于边距的降噪方式相较于强化学习的方式更加简便合理，因此 MMDSRE 的性能要由于 PCNN+RU。

2） 取不同 N 值时 MMDSRE 算法与对比算法的 P@N 指标对比

此外，与之前的工作[7]类似，本章使用 P@N 指标衡量算法的性能，P@N 指的是测试集中前 N 个最高概率的关系提取结果的精度，本章选取前 100，200，300 个关系事实的精度以及他们的平均值进行性能评测，如表 4.2 所示。

表4.2 N 取 100，200，300 时不同算法的 P@N 值对比表

P@N(%) / 算法	N=100	N=200	N=300	Mean
PCNN[30](2015)	72.3	69.7	64.1	68.7
PCNN+ATT[31](2016)	76.2	73.1	67.4	72.2
HG[38](2019)	79.1	75.6	71.2	75.3
Seg[39](2020)	88.7	84.3	80.5	84.5
PCNN+RL[42](2018)	77.4	74.2	70.9	74.2
DCRE[44](2020)	84.4	80.3	75.9	80.2
PCNN+RU[45](2020)	91.6	86.9	83.0	87.2
MMDSRE(ours)	**93.2**	**87.8**	**83.9**	**88.3**

如表 4.2 所示，当 N 值分别取 100，200，300 时，本章所提出的 MMDSRE 算法取得了较高的 P@N 值，与 PCNN 算法相比，MMDSRE 算法的 P@N 值平均提升了 28.6%，即使与当前最好的 PCNN+RU 算法相比，MMDSRE 算法的 P@N 值平均也提

升了 1.3%，这是由于 MMDSRE 可以更加有效地识别出句包中的噪声数据，从而使噪声对预测结果地影响程度减小。这充分说明了 MMDSRE 算法所采用的策略更加有效，更能准确地识别出句包中的噪声数据并准确提取实体对的关系。

3） 消融实验

为充分展示 MMDSRE 算法的噪声识别器的性能，本章在 NYT 数据集上对 MMDSRE 算法、DCRE[44]算法以及 PCNN+RU[45]算法中的噪声识别方式进行对比，分别从三种方式识别的噪声数据中随机选取 100 条数据并手动识别其准确性，为了实验公平性，三者都使用 PCNN 模型作为特征提取器，噪声识别器的精度如表 4.3 所示，为了消除随机性影响，表中的数据均是进行 5 次随机选取数据后取平均值得到的。

表4.3 不同算法的噪声识别精度对比表

算法	噪声识别精度
DCRE[44](2020)	0.835
PCNN+RU [45](2020)	0.876
MMDSRE(ours)	**0.923**

如表 4.3 所示，噪声识别精度指的是从三种方式识别的噪声数据中随机选取 100 条数据并手动识别其为噪声的准确性。对该结果的分析如下所示：

(1) 三种算法中，DCRE 算法[44]采用的基于匹配的方式的噪声识别精度最低，这是由于该方法在计算句包中的每个句子与其对应标签的相似度后，若相似度低于阈值则认为该句子为噪声语句，而对于阈值的设置需要大量先验知识，局限性较大。

(2) PCNN+RU 算法[45]中的基于强化学习的降噪方式完全可以通过强化学习机制训练，性能优于基于匹配的方法，但是由于强化学习中的延迟奖励机制，难以寻找最优决策，因此噪声识别精度也有待提高。

(3) 本章设计的基于边距度量的噪声识别器可以通过如公式（4-7）所示的边距度量方式识别出噪声数据，并以此训练噪声分类器，基于边距度量的噪声识别方式能够充分利用数据的分布特征快速准确识别出噪声，与 DCRE 以及 PCNN+RU 的噪声识别方式相比，MMDSRE 识别噪声的精度分别提升了 10.9% 和 5.8%。

此外，为测试特征提取器中 GCN 的有效性，基于上述噪声识别精度对比实验，本章通过使用 MMDSRE 的特征提取器（PCNN 与 GCN 相结合的方式）提取句子特征并采用 MMDSRE 的噪声识别器识别的噪声精度如表 4.4 所示：

表4.4 使用不同特征提取器的噪声识别精度对比表

特征提取器	噪声识别精度
PCNN	0.923
PCNN+GCN	0.936

如表 4.4 所示，在特征提取器中加入 GCN 后，与原精度相比噪声识别精度提升了 1.4%，这是由于使用 GCN 提取句子特征时，可以根据图神经网络学习句子的结构信息，以此丰富句子的语义，提取的语义更加准确，更有助于后续噪声数据的识别。

4.6 本章小结

本章首先对基于远程监督关系抽取算法的现有缺陷进行总结，进而引出 MMDSRE 算法的设计动机，接着展示了算法的基本框架，并介绍了算法每个模块的功能，然后详细介绍了算法的设计策略，对算法的特征抽取器、噪声识别器以及关系抽取器的设计方法进行了详细阐述，之后对算法的训练与预测步骤进行了描述，并给出了相关伪代码，最后，通过在真实数据集上与当前最新最好的算法进行对比，证明 MMDSRE 在远程监督关系抽取方面的优势，并通过消融实验证明了噪声识别器的有效性。

第五章 与 OpenNRE 的集成

OpenNRE[55]是清华大学自然语言处理实验室推出的一款开源的神经网络关系抽取工具包，包括了多款常用的关系抽取模型。本章将本文提出的 ERGSRE 算法和 MMDSRE 算法集成至 OpenNRE 中，可以对句子中的实体对一键式关系抽取并提升 OpenNRE 工具的性能。

5.1 OpenNRE 工具介绍

OpenNRE 是一个开源的且可扩展的关系抽取工具包，它提供了一个统一的框架来实现用于关系提取（RE）的神经网络模型。通过实现典型的 RE 方法，OpenNRE 不仅允许开发人员训练自定义模型以从纯文本中提取结构化的关系事实，而且还支持研究人员进行模型快速验证。此外，OpenNRE 提供了基于 TensorFlow 和 PyTorch 的各种功能性 RE 模块，以保持足够的模块化和可扩展性，从而可以更加方便地将新模型集成到框架中。此外，该工具包可以在各种场景中提取关系事实，以及将提取的事实与 Wikidata 知识库对齐，这可以使各种下游知识驱动的应用程序受益（例如，信息检索和问题解答）。

OpenNRE 的设计目标是实现系统封装、操作效率、模型扩展性和易用性之间的平衡。其整体架构如图 5.1 所示，它由分词（Tokenization）、组件（Module）、编码器（Encoder）、模型（Model）以及框架（Feamework）5 个模块组成，各个模块的功能如下：

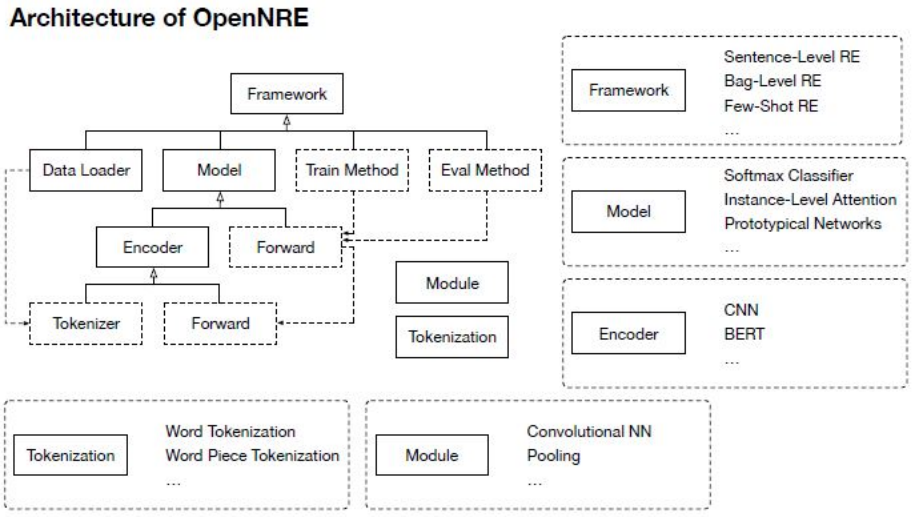

图5.1 OpenNRE 工具的整体架构

分词（Tokenization）：该模块负责将输入文本划分为多个单词，对于组合词或有前后缀的词，将其拆分为基本单词输入。

组件（Module）：该模块由用于模型实现的各种功能性神经网络组件组成，如基本的池化操作和激活函数等。使用这些原子模块来构建和部署系统可使系统具有高度的可扩展性。

编码器（Encoder）：编码器用于将文本编码到相应的词嵌入空间中，以提供语义特征。OpenNRE 实现了基于分词和组件的 BaseEncoder 类，它可以提供文本分词和词嵌入的基本功能。通过扩展 BaseEncoder 类和实现特定的神经网络结构，可以实现各种特定的编码器。

模型（Model）：一些开发人员可能不需要实现以及验证自己的模型，他们的主要目标是可以快速训练和部署一些特定模型。为此，OpenNRE 集成了几种典型的关系抽取模型（如 PCNN 等），此外工具包中还包括一些特殊基本算法，例如注意力机制、强化学习等，以供用户无需了解所有技术细节，迅速部署训练模型。

框架（Feamework）：框架模块主要负责集成上述四个模块并支持各种功能（包括数据处理、模型训练、模型优化和模型评估）。

OpenNRE 适用于各种 RE 场景，包括句子级 RE、句包级 RE、文档级 RE 和小样本 RE。为了完成一条完整的结构化信息提取流水线，OpenNRE 在一定程度上具备了面向实体应用的能力，如命名实体识别（NER）等。这些应用场景的示例如图 5.2 所示。

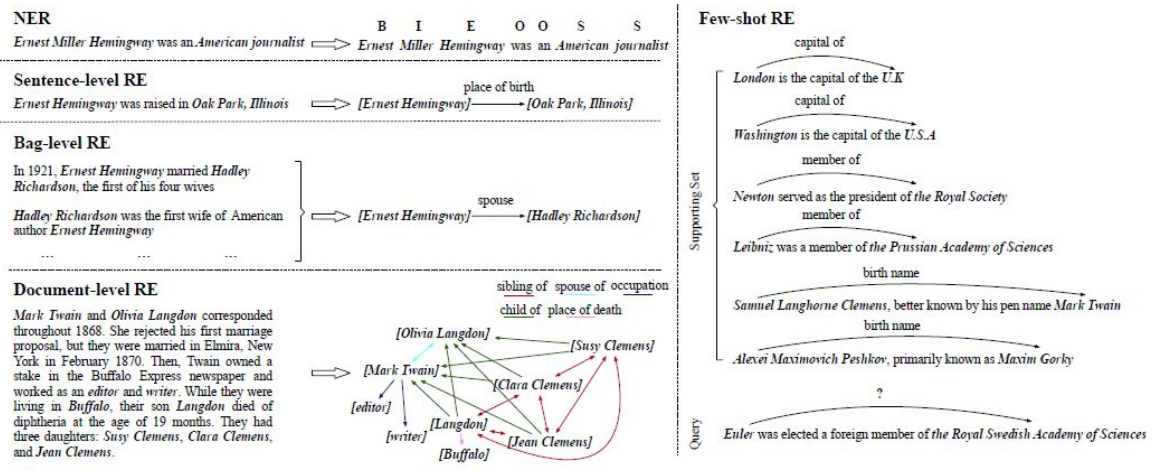

图5.2 OpenNRE 中所有应用场景的应用示例

为了从纯文本中提取结构化信息，需要从文本中提取实体，然后预测实体之间的关系，命名实体识别（NER）的目标就是从文本中识别并提取实体；句子级的关系抽取（Sentence-level RE）是从单个句子包含的两个实体中抽取关系；句包级关系抽取

（Bag-level RE）是利用远程监督对齐知识库与文本，将具有相同实体对的文本组成句包，对整个句包进行关系抽取；文档级关系抽取（Document-level RE）是挖掘整个文档中涉及的实体间的复杂关系，这些实体可能是句内的也可能是句间的；小样本关系抽取（Few-shot RE）是根据每个关系中的极少量训练数据训练模型，来抽取句子中实体对之间的关系。

本文所提出的 ERGSRE 算法和 MMDSRE 算法分别是从单个句子、句包中抽取实体关系，因此分别属于句子级的关系抽取和句包级的关系抽取。本章将本文提出的最优算法 ERGSRE 算法和 MMDSRE 算法分别集成于 Open-NRE 的 Sentence-level RE 和 Bag-level RE 中，以此来提升 OpenNRE 工具包的性能。

5.2 ERGSRE 算法的集成

OpenNRE 当前集成的句子级的关系抽取算法采用的是基于 CNN[11]，LSTM[13]等方式的算法，其主要采用 CNN、LSTM 等基础模型学习实体以及句子的语义来进行关系分类，由于 CNN、LSTM 这些模型很难准确提取较长句子的语义，因此这类方法的关系抽取的精度有待提高。从 3.5.3 节的实验部分可知，本文提出的 ERGSRE 算法性能远远高于 OpenNRE 工具中集成的关系抽取算法，因此，本文将 ERGSRE 算法集成于 OpenNRE 工具中以提升该工具的性能。

在与 OpenNRE 工具集成时，为保证输入一致性，采用 OpenNRE 的分词模块对句子进行分词处理，然后使用 ERGSRE 的特征提取器作为 OpenNRE 的 Encoder 学习文本的特征表示，最后将噪声识别器和关系抽取器集成与 model 模块用于关系抽取，最后使用 OpenNRE 的 Framework 模块进行算法的整体训练与预测。集成示意图如图 5.3 所示：

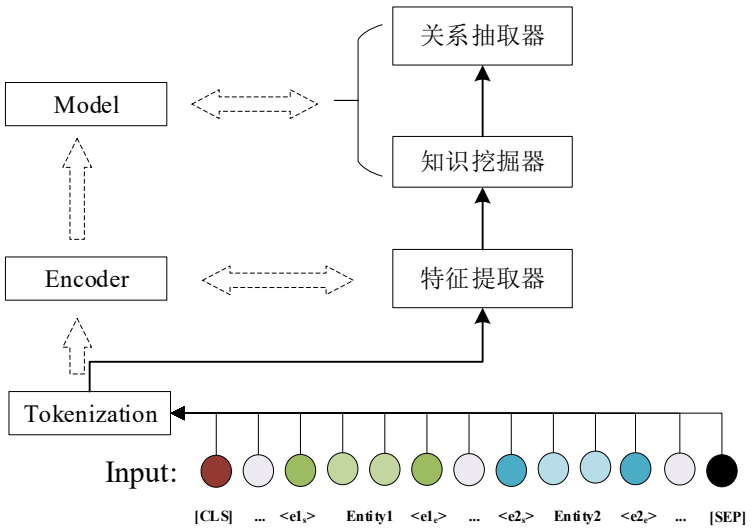

图5.3 ERGSRE 与 OpenNRE 集成示意图

如图 5.3 所示，图中的实线箭头表示输入数据后，模型的执行过程，双向虚线箭头表示 ERGSRE 模型的各模块与 OpenNRE 工具的集成方式，特征提取器集成至 OpenNRE 的 Encoder 模块中，知识挖掘器和关系抽取器集成至 Model 模块。

将 ERGDSRE 的各模块与 OpenNRE 的对应模块集成在一起后，即可使用 OpenNRE 工具进行句子级别的关系抽取。此时使用 OpenNRE 执行关系抽取的流程如图 5.4 所示：

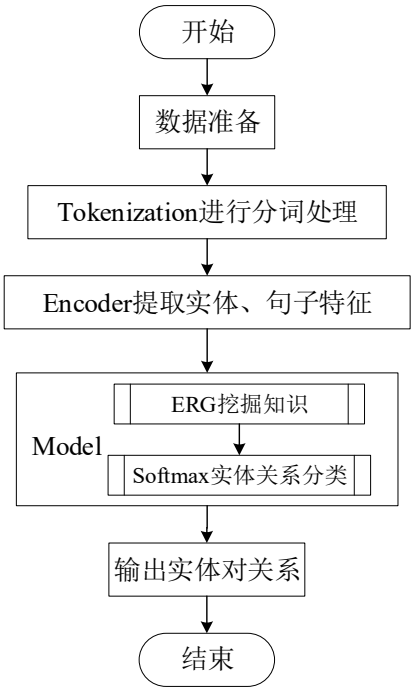

图5.4 嵌入 ERGSRE 后 OpenNRE 进行句子级别关系抽取流程示意图

如图 5.4 所示，将 ERGNRE 与 OpenNRE 集成后进行句子级别的关系抽取时，首先进行数据准备，之后使用分词工具对输入数据进行处理，然后输入至 Encoder 模块中提取实体以及句子的语义特征，之后使用 Model 模块进行知识的挖掘以及关系分类，最终输出实体对的关系。

使用 OpenNRE 进行句子级别的关系抽取时，仅需要调用 OpenNRE 的 *model.infer* 接口即可预测目标句子中的实体关系，如下示例所示：

```
model.infer({'text': 'He was the son of Máel Dúin mac Máele Fithrich, and grandson of the high king Áed Uaridnach (died 612).', 'h': {'pos': (18, 46)}, 't': {'pos': (78, 91)}})
```

其中，*'text'*指文本内容，*'h'*指第一个实体，*'t'*指第二个实体，*'pos'*指实体位置，即实体的第一个字符到最后一个字符的位置。

使用 OpenNRE 原本的算法执行上述命令时，OpenNRE 的输出结果为：

```
('father', 0.5108704566955566)
```

该输出表示最终预测得到的实体关系为'*father*'关系，置信度为 0.5108704566955566。

在嵌入 ERGSRE 算法后，通过如图 5.4 的流程，最终输出的结果为：

```
('father', 0.723503192347433)
```

通过嵌入 ERGSRE 算法后的输出结果对比可以发现，对于同一个句子，ERGSRE 算法能够给出更高置信度的预测。此外，为对比嵌入 ERGSRE 算法前后 OpenNRE 工具性能，本文使用 Sem-Eval 2010 Task 8 数据集测试了 OpenNRE 的精度，对比结果如表 5.1 所示：

表5.1 嵌入 ERGSRE 算法前后 OpenNRE 工具的精度对比表

ERGSRE 算法	**算法精度**
嵌入 ERGSRE 前	0.73
嵌入 ERGSRE 后	0.78

如表 5.1 所示，将 ERGSRE 算法嵌入至 OpenNRE 工具后，使得 OpenNRE 在 Sem-Eval 2010 Task 8 数据集上的关系抽取精度提升了 6.8%，这是由于 ERGSRE 算法能够挖掘知识以丰富实体对信息，进而提升关系抽取的精度。

5.3 MMDSRE 算法的集成

有监督关系抽取算法需要大量的人工标注的训练数据，这些人工标注的数据即昂贵又耗时，因此具有较大的局限性。Mintz 等人[8]引入了远程监督，通过对齐知识库和文本，自动标注文本数据。远程监督虽然带来了足够的标注数据，但是也引入了错误标注的问题，考虑到一个实体对可能在不同的句子中出现多次，并且其中一些句子很有可能表达实体对之间的关系。因此 Riedel 等人[29]将涉及相同实体的句子聚和在一起组合一个句包，如图 5.2 所示，综合一个句包中不同句子的特征可以提供更可靠的信息，并得出更准确的预测结果。基于句包级的关系抽取广泛应用于各种远程监督关系抽取方法中。

OpenNRE 中集成的句包级关系抽取算法为 PCNN[30]和 PCNN+ATT[31]算法，这些基于选择性的算法通过筛选句包中能正确表达实体间关系的句子进行关系抽取，而忽略标注错误的数据。但是对于一个句包中的句子都标注错误的情况无法处理，因此本节将基于降噪的 MMDSRE 算法集成于 OpenNRE 中，通过识别句包中的噪声数据，若一个句包中的数据全部标注错误，则将其标记为“NA”关系，这种方式可有效解决基于选择性算法的缺陷。对于错误标记的数据还可以将其作为负样本与正确标注的正样本数据一起训练，提升关系抽取的准确性，并提高 OpenNRE 工具的性能。

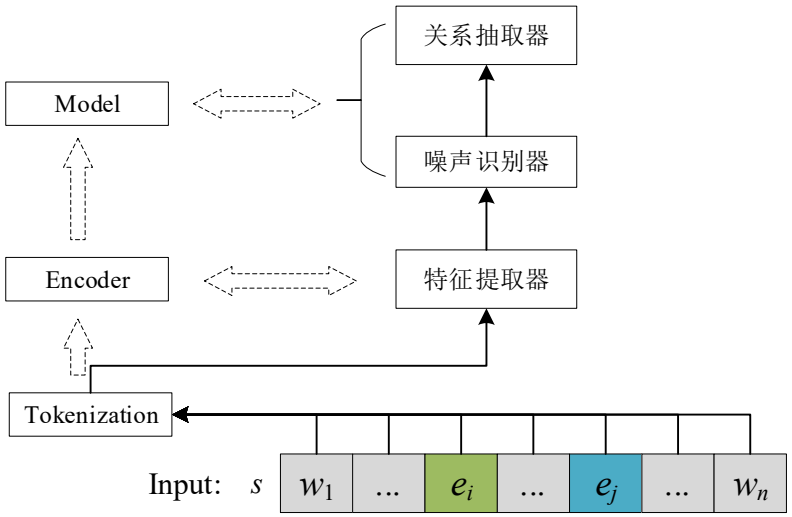

图5.5 MMDSRE 算法与 OpenNRE 工具的集成示意图

与 5.2 节类似，将 MMDSRE 算法嵌入至 OpenNRE 时，首先使用 MMDSRE 的分词模块进行分词处理，对于编码器 Encoder 模块，使用 MMDSRE 的特征提取器提取句子的语义特征，随后将噪声识别器和关系抽取器集成至 Model 模块，最后使用 OpenNRE 的 Framework 模块进行算法的整体训练与预测。MMDSRE 算法与 OpenNRE 工具的集成示意图如图 5.5 所示。

如图 5.5 所示，与图 5.3 类似，图中的实线箭头表示输入数据后，模型的执行过程，双向虚线箭头表示 MMDSRE 模型的各模块与 OpenNRE 工具的集成方式，特征提取器集成至 OpenNRE 的 Encoder 模块中，噪声识别器和关系抽取器集成至 Model 模块。

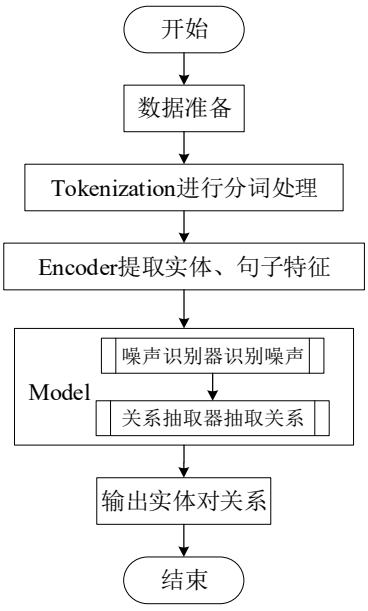

图5.6 嵌入 MMDSRE 后 OpenNRE 进行句包级别关系抽取流程示意图

将 MMDSRE 算法集成至 OpenNRE 的 Bag-level RE 之后，当输入数据后，OpenNRE 的执行过程如图 5.6 所示。

如图 5.6 所示，将 MMDSRE 与 OpenNRE 集成后进行句包级别的关系抽取时，首先进行数据准备，之后使用分词工具对输入数据进行处理，然后输入至 Encoder 模块中提取实体以及句子的语义特征，之后使用 Model 模块进行噪声识别以及关系分类，最终输出实体对的关系。

使用 OpenNRE 进行句包级关系抽取时，同样调用 *model.infer* 接口，但与句子级关系抽取不同的是，此时‘text’输入的是一个句包，而不是一个单独的句子，输入示例如下所示：

```
model.infer({{'text': 'In 1921, Ernest Hemingway married Hadley Richardson, the first of his four wives.', 'h': {'pos': (9, 25)}, 't': {'pos': (34, 51)}}, {'text': 'Hadley Richardson was the first wife of American author Ernest Hemingway.', 'h': {'pos': (57, 73)}, 't': {'pos': (0, 17)}}})
```

与 5.2 节类似，使用 OpenNRE 原本的算法执行上述命令时，OpenNRE 的输出结果为：

```
('father', 0.4608704566955566)
```

该输出表示最终预测得到的实体关系为'*father*'关系，置信度为 0.5108704566955566。

嵌入 MMDSRE 算法后，根据图 5.6 所示的流程执行后，OpenNRE 最后的输出为：

```
('spouse', 0.5734768937665873)
```

通过嵌入 MMDSRE 算法后的输出结果对比可以发现，对于同一个句包，MMDSRE 算法能够给出更高置信度的预测。此外，为对比嵌入 MMDSRE 算法前后 OpenNRE 工具性能，本文使用 NYT 数据集测试了 OpenNRE 的精度，对比结果如表 5.2 所示：

表5.2 嵌入 MMDSRE 算法前后 OpenNRE 工具的精度对比表

MMDSRE 算法	**算法精度**
嵌入 MMDSRE 前	0.65
嵌入 MMDSRE 后	0.82

如表 5.2 所示，将 MMDSRE 算法嵌入至 OpenNRE 工具后，使得 OpenNRE 在 NYT 数据集上的关系抽取精度提升了 26.2%，这是由于 MMDSRE 算法具有更准确的噪声识别精度，并且在识别出噪声后，可以将噪声作为负样本，通过正负样本学习的方式提升了远程监督关系抽取算法的性能。

5.4 本章小结

本章首先对开源关系抽取工具包 OpenNRE 进行了介绍，然后将本文设计的 ERGSRE 算法与 MMDSRE 算法分别集成至 OpenNRE 的句子级关系抽取和句包级关系抽取，以提升 OpenNRE 工具的性能。

第六章 总结与展望

6.1 论文工作总结

随着互联网上非结构化文本数据的日益增多，人们迫切需要建立一个高效的可以自动提取非结构化文本数据的信息抽取系统。关系抽取任务是从给定的文本数据中抽取实体的关系，是信息抽取最重要的环节。本文对已有的关系抽取算法进行深入研究，并分别提出一种高效的有监督关系抽取算法和一种高效的远程监督关系抽取算法。

在有监督关系抽取算法的研究中，本文针对现有算法的不足，设计了一种可以存储数据所在语料库的知识的实体关系图，为了筛选与挖掘有用知识，本文提出了一种基于语义相似度的图神经网络来从实体关系图中挖掘与实体相关的知识，以丰富实体信息，提升关系抽取的准确性。同时，本文在真实的有监督关系抽取数据集上进行仿真实验，实验结果表明，本文提出的基于实体关系图的有监督关系抽取算法 ERGSRE 性能均好于当前最新最好的有监督关系抽取算法。

在远程监督关系抽取算法的研究中，本文使用图神经网络（GCN）与分段卷积神经网络（PCNN）相结合的方式作为特征提取器充分提取句子的语义关系，然后设计了一种新的基于边距度量的噪声识别方式来识别远程监督关系抽取数据集中的噪声，最后将噪声作为负样本，与正确标记的正样本数据一起训练关系抽取器以提升关系抽取器的精度。最后在远程监督关系抽取真实数据集上进行仿真实验，实验表明，本文设计并提出的基于边距度量降噪的远程监督关系抽取算法 MMDSRE 是当前最新最好的远程监督关系抽取算法。

此外，本文深入研究了关系抽取的应用场景，并将本文设计的 ERGSRE 算法和 MMDSRE 算法集成至开放的关系抽取工具包 OpenNRE 中，以提升 OpenNRE 工具的性能。

6.2 未来工作展望

本文提出的 ERGSRE 算法和 MMDSRE 算法虽然都取得了最好的效果，但是这两个算法对于训练数据的依赖过高，当算法加入新的训练任务后，很容易忘记之前的训练任务，而终身学习机制可以有效解决这一现象，因此如何引入终身学习机制以提升模型的抗遗忘能力是作者未来的研究方向。

参考文献

[1] Mollá D, Vicedo J L. Question answering in restricted domains: An overview[J]. Computational Linguistics, 2007, 33(1): 41-61.

[2] Gan L X, Wan C X, Liu D X, et al. Chinese entity relationship extraction based on syntactic and semantic features[J]. Journal of Computer Research and Development, 2016,53(2): 284-302.

[3] Chu C, Wang R. A survey of domain adaptation for neural machine translation[J]. arXiv preprint arXiv:1806.00258, 2018.

[4] 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, 2004: 415-422.

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

[6] Cui M, Li L, Wang Z, et al. A survey on relation extraction[C]. Proceedings of China Conference on Knowledge Graph and Semantic Computing. Springer, Singapore, 2017: 50-58.

[7] Shi Y, Xiao Y, Niu L. A brief survey of relation extraction based on distant supervision[C]. Proceedings of International Conference on Computational Science, Springer, Cham, 2019: 293-303.

[8] 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.

[9] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics on Interactive poster and demonstration sessions, 2004: 22-es.

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

[11] 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: technical papers, 2014: 2335-2344.

[12] Zhang Z, Shu X, Yu B, et al. Distilling knowledge from well-informed soft labels for neural relation extraction[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05): 9620-9627.

[13] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]. Proceedings of the 54nd Annual Meeting of the Association for Computational Linguistics, 2016: 207-212.

[14] Wu S, He Y. Enriching pre-trained language model with entity information for relation classification[C]. Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019: 2361-2364.

[15] Soares L B, FitzGerald N, Ling J, et al. Matching the blanks: Distributional similarity for relation learning[C]. Proceedings of the 57nd Annual Meeting of the Association for Computational Linguistics, 2019: 2895-2905.

[16] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE transactions on neural networks and learning systems, 2020, 32(1):4-24.

[17] Guo Z, Zhang Y, Lu W. Attention Guided Graph Convolutional Networks for Relation Extraction[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 241-251.

[18] Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced language representation with informative entities[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1441-1451.

[19] Zhao Y, Wan H, Gao J, et al. Improving relation classification by entity pair graph[C]. Proceedings of the Asian Conference on Machine Learning, 2019: 1156-1171.

[20] 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(05): 8928-8935.

[21] Sam R C, Le H T, Nguyen T T. Relation extraction in Vietnamese text using conditional random fields[C]. Asia Information Retrieval Symposium. Springer, Berlin, Heidelberg, 2010: 330-339.

[22] Zhao S, Grishman R. Extracting relations with integrated information using kernel methods[C]. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics, 2005: 419-426.

[23] 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.

[24] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[25] Yang S M, Yoo S Y, Jeong O R. DeNERT-KG: Named Entity and Relation Extraction Model Using DQN, Knowledge Graph, and BERT[J]. Applied Sciences, 2020, 10(18): 6429.

[26] Bizer C, Lehmann J, Kobilarov G, et al. DBpedia-A crystallization point for the Web of Data[J]. Journal of web semantics, 2009, 7(3): 154-165.

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

[28] Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.

[29] Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text[C]. Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2010: 148-163.

[30] 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.

[31] 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.

[32] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv: 1706.03762, 2017.

[33] Tran T T, Le P, Ananiadou S. Revisiting unsupervised relation extraction[J]. Proceedings of the 58nd Annual Meeting of the Association for Computational Linguistics, 2020:7498-7505.

[34] Sun A, Grishman R, Sekine S. Semi-supervised relation extraction with large-scale word clustering[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: human language technologies. 2011: 521-529.

[35] Xiao Y, Jina Y, Cheng R, et al. Hybrid Attention-Based Transformer Block Model for Distant Supervision Relation Extraction[J]. arXiv preprint arXiv: 2003.11518, 2020.

[36] Vashishth S, Joshi R, Prayaga S S, et al. Reside: Improving distantly-supervised neural relation extraction using side information[J]. arXiv preprint arXiv:1812.04361, 2018.

[37] 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.

[38] Duan S, Gao H, Liu B, et al. A Hybrid Graph Model for Distant Supervision Relation Extraction[C]. Proceedings of European Semantic Web Conference. Springer, Cham, 2019: 36-51.

[39] Li Y, Long G, Shen T, et al. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05): 8269-8276.

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

[41] Jia W, Dai D, Xiao X, et al. ARNOR: Attention regularization based noise reduction for distant supervision relation classification[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1399-1408.

[42] Qin, P., Xu, W., Wang, W.Y.: Robust distant supervision relation extraction via deep reinforcement learning.[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 2137-2147.

[43] Chen T, Wang N, He M, et al. Reducing Wrong Labels for Distantly Supervised Relation Extraction With Reinforcement Learning[J]. IEEE Access, 2020, 8: 81320-81330.

[44] Shang Y, Huang H Y, Mao X L, et al. Are Noisy Sentences Useless for Distant Supervised Relation Extraction?[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05): 8799-8806.

[45] He Z, Chen W, Wang Y, et al. Improving neural relation extraction with positive and unlabeled learning[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05): 7927-7934.

[46] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[47] Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv 2019, arXiv:1909.11942.

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

[49] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2015, 29(1): 2181-2187.

[50] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[51] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

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

[53] Zhang Y, Zhong V, Chen D, et al. Position-aware attention and supervised data improve slot filling[C]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 35-45.

[54] Pleiss G, Zhang T, Elenberg E R, et al. Identifying mislabeled data using the area under the margin ranking[J]. arXiv preprint arXiv:2001.10528, 2020.

[55] 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.

致谢

三年研究生生活一晃而过，在这三年生活中经历了许多的事情、学习到了很多的知识，这三年的研究生生活使我终生难忘！首先要感谢的是我的研究生导师李雁妮教授，从论文选题、论文开题答辩、中期审查、论文的反复修改和最后的论文答辩，李老师都始终给予我了很多悉心指导和殷切鼓励。李老师在我的三年研究生生活中倾注了许多汗水和心血，李老师的刻苦科研、严于律己、认真做事的态度和精神是我人生前行的标杆，它激励着我不断进取、不断创新，这是一笔极其珍贵的精神财富。

然后，我要感谢企业导师刘志鹏高级工程师，在我的工程实践学习中，他不断地对我进行教导，无私地教授我工程领域知识和技巧，使我的工程实践能力得到极大的提升，为我以后走入工作岗位打下了坚实的基础。

特别感谢实验室为我提供帮助的王陟、焦一源师兄，姚凯程、胡代旺、刘佳伟、史家辉师弟，李朝霞、吕鹏帆师妹以及张文辉、费航、郝小慧同学。是他们为我的研究生生活提供了欢乐、积极向上的学习氛围，感谢他们三年来的陪伴，在我科研和生活上遇到困难时，感谢他们的帮助与关怀，以及对我的倾力相助。在此，还要特别感谢焦一源博士师兄，是他与我一起在关系抽取问题一起攻克难关，做了大量的工作。

除此之外，要感谢研究生期间给我上课、教我知识、给我帮助的每一位老师，感谢你们在我成长之路上的帮助，感谢你们的无私教诲，感谢母校为我所提供的优良学习环境与资源。还要衷心感谢在百忙之中对我论文进行评阅和参与答辩的各位专家评审，辛苦了。

此外，感谢我的女朋友刘静，感谢她在我研究生三年以来的陪伴与鼓励，在我遇到困难时，是她在我身边支持我、鼓励我，让我勇敢面对挑战，在我取得成绩时，也有她陪我一起开心、庆祝。三年的异地生活更加奠定了我们坚实的感情基础，正是因为有她的陪伴，我的人生旅途才更加精彩。

最后，我要深深地感谢我的父母。感谢他们的养育之恩，更感谢他们多年来对我学业和生活的支持与关怀，他们是我能顺利完成研究生学业生涯的基础与关键，是我的精神上的支柱。

我将用我今后更加努力地工作与学习，来感恩与回报所有曾给予我帮助与支持的老师、同学及家人！再次衷心地感谢你们！

作者简介

1. 基本情况

杨文成，男，河南三门峡人，1996 年 8 月出生，西安电子科技大学计算机科学与技术学院计算机技术专业 2018 级硕士研究生。

2. 教育背景

2014.08～2018.07 南京邮电大学，本科，专业：网络工程
2018.08～　　西安电子科技大学，硕士研究生，专业：计算机技术

3. 攻读硕士学位期间的研究成果

3.1 发表学术论文

[1] Li Y, **Yang W**, Zhong Z, et al. NOLGP: A Novel Optimized Large-Scale Graph Partitioning Algorithm[C]. 2019 15th International Conference on Computational Intelligence and Security (CIS). IEEE, 2019: 127-131. (EI: 20192507062904)

3.2 参与科研项目及获奖

[1] 国家自然科学基金面上项目，海量深网数据源的自动发现与集成研究, 2015.1~2018.12, 已结题, 项目参与者。
[2] 西安中科微精光子制造科技有限公司合作项目，对外轮廓图形内部填充算法研究，2019.5~2019.12，已完成，项目核心参与者。
[3] 中国信息通信研究院项目，一站式深度学习平台研发，2019.7~2019.10，已结束，项目主要参与者。
[4] 获校一等奖学金，2018~2019 年度。
[5] 获校一等奖学金，2019~2020 年度。
[6] 获校一等奖学金，2020~2021 年度。

西安電子科技大學
XIDIAN UNIVERSITY

地址：西安市太白南路2号
邮编：710071
网址：www.xidian.edu.cn