分 类 号 学号 M201671801

学校代码 10487 密级



硕士学位论文

基于实体类型嵌入与循环分段残差网络的关系抽取研究

学位申请人：赵恢强

学科专业：信息与通信工程

指导教师：王玉明 副教授

答辩日期：2019年5月

A Dissertation Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

Research of Relation Extraction Based on Entity Type Embedding and Recurrent Piecewise Residual Networks

Candidate： Huiqiang Zhao

Major： Information and Communication Engineering

Supervisor：Prof. Wang Yuming

**Huazhong University of Science and Technology**

**Wuhan, Hubei 430074, P. R. China**

**May, 2019**

**独创性声明**

本人声明所呈交的学位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，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

**学位论文版权使用授权书**

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保 密□，在\_\_\_\_\_年解密后适用本授权书。

不保密□。

（请在以上方框内打“√”）

学位论文作者签名： 指导教师签名：

日期： 年 月 日 日期： 年 月 日

摘 要

作为信息抽取的核心任务，有效的关系抽取是信息爆炸时代的一大挑战。本文在进行关系抽取这一基础性研究时发现，现有的模型都无法有效的抽取长而复杂或无直接对应关系触发词的句子中的实体对关系，而这种句子在自然语言文本中是普遍存在的，这使得关系抽取的效果遇到了瓶颈。

本文深入调研并分析了国内外在关系抽取方面的研究。在融合更多更深次信息的思想启发下，为了丰富模型的表示，提出**实体类型嵌入(Entity Type Embedding, ETE)**，ETE和词嵌入、位置嵌入一起作为模型的联合嵌入层；为了更精细地编码模型的嵌入表示，精心设计了**循环分段残差网络(Recurrent Piecewise Residual Networks, RPRN)**结构，RPRN能有效获取句子上下文中的潜在表示。本文在进行关系抽取建模时还引入了对抗训练，从而进一步增强模型的泛化能力。

为了更好地评估模型，本文进行实验时采用了在关系抽取领域广泛使用的纽约时报公开数据集。然后，将本文提出的**(ETE-RPRN)**与其他学者的模型从各种评估指标上进行了定量比较，并对不同模型抽取出的不同结果进行了定性分析。实验结果显示，本文提出的胜过现有最先进的模型。

本文提出的关系抽取模型有效解决了现存模型所遇到的困境，从而大大提高了关系抽取的整体效果，这对文本数据的信息化建设具有重大的意义。同时抽取的关系事实是构建语义检索、问答系统等应用的重要基础。

**关键词：**关系抽取；实体类型嵌入；循环神经网络；残差网络

Abstract

As a core mission in information extraction, effective relation extraction is a great challenge in the information explosion era. In this paper, we found that existing models usually fail in extracting correct relation of entity pair effectively when the sentence is long and complex or there is not the firsthand trigger word of corresponding relation, whereas such sentences are ubiquitous in texts formed of natural language, which makes the effects of relation extraction hit a bottleneck.

In this paper, we penetrate into investigating and anlyzing the researches on relation extraction at home and abroad. Inspired by the idea of fusing more and deeper information, we propose **Entity Type Embedding(ETE)** to enrich the representation of the model, ETE, word embedding and position embedding serve as the joint embedding of our model; For encoding the representation of the embedding subtly, we devise an architecture of **Recurrent Piecewise Residual Networks(RPRN)** meticulously, which may acquire the latent representation underlying the context of a sentence. We introduce adversrial training when modeling relation extraction as well, which further enhance the generalization capacity of our model.

For the sake of evaluating the model preferably, the public dataset of New York Times, extensively employed in the field of relation extraction, is adopted when conducting experiments in this paper. Subsequently, our proposal(**ETE-RPRN**) is compared with the models of other scholars quantitatively from diverse evaluation metrics, and then different results extracted from different models are analyzed qualitatively. Experiment results show that our proposal outperforms the state-of-the-art models.

Our proposal about relation extraction effectively solve the dilemma of existing models, in turn improving the overall effect of relation extraction, which is of great significance to the information construction of text data. Simutaneously, the extracted relation facts are key foundation of building applications such as semantic search, question answering system and so on.

**Keywords**：Relation Extraction, Entity Type Embedding, Recurrent Neural Networks, Residual Networks

# 目 录

[摘 要 I](#_Toc3789988)

[Abstract II](#_Toc3789989)

[目 录 III](#_Toc3789990)

[略缩语 V](#_Toc3789991)

[1 绪论 1](#_Toc3789992)

[1.1 课题背景及来源 1](#_Toc3789993)

[1.2 课题现状及分析 2](#_Toc3789994)

[1.3 研究内容及意义 5](#_Toc3789995)

[1.4 本文的组织结构 7](#_Toc3789996)

[2 相关理论及技术概述 8](#_Toc3789997)

[2.1 远程监督 8](#_Toc3789998)

[2.2 循环神经网络 10](#_Toc3789999)

[2.3 残差网络 14](#_Toc3790000)

[2.4 注意力机制 18](#_Toc3790001)

[2.5 对抗训练 18](#_Toc3790002)

[2.6 StanfordNLP 19](#_Toc3790003)

[2.7 Tensorflow 20](#_Toc3790004)

[2.8 本章小结 20](#_Toc3790005)

[3 关系抽取研究与算法设计 21](#_Toc3790006)

[3.1 困境分析 21](#_Toc3790007)

[3.2 模型架构设计 22](#_Toc3790008)

[3.3 联合嵌入 23](#_Toc3790009)

[3.4 RPRN编码器 25](#_Toc3790010)

[3.5 注意力选择器 28](#_Toc3790011)

[3.6 本章小结 31](#_Toc3790012)

[4 实验及模型评估 32](#_Toc3790013)

[4.1 实验数据预处理 32](#_Toc3790014)

[4.2 实验设置 33](#_Toc3790015)

[4.3 评估结果 35](#_Toc3790016)

[4.4 案例分析 38](#_Toc3790017)

[4.5 本章小结 39](#_Toc3790018)

[5 总结与展望 40](#_Toc3790019)

[5.1 论文总结 40](#_Toc3790020)

[5.2 未来工作展望 41](#_Toc3790021)

[致 谢 42](#_Toc3790022)

[参考文献 43](#_Toc3790023)

# 略缩语

|  |  |
| --- | --- |
| **ETE** – Entity Type Embedding | 实体类型嵌入 |
| **RPRN** – Recurrent Piecewise Residual Networks | 循环分段残差网络 |
| **DARPA -** Defense Advanced Research Projects Agency | 国防高级研究计划局 |
| **NIST -** National Institute of Standards and Technology | 国家标准与技术研究所 |
| **MUC –** Message Understanding Conference | 消息理解会议 |
| **ACE –** Automatic Content Extracting | 自动内容抽取 |
| **OCR –** Optical Character Recognition | 光学字符识别 |
| **CRF -** Conditional Random Field | 条件随机场 |
| **NYT** – New York Times | 纽约时报 |
| **MIML –** Multiple Instance Multiple Label | 多实例多标签 |
| **POS** – Part Of Speech tagging | 词性标注 |
| **SRL** – Sematic Role Labeling | 语义角色标注 |
| **NLP** – Natural Language Processing | 自然语言处理 |
| **CNN** – Convolutional Neural Networks | 卷积神经网络 |
| **RNN** – Recurrent Neural Networks | 循环神经网络 |
| **LSTM –** Long Short Term Memory | 长短时记忆 |
| **GRU** – Gated Recurrent Unit | 门控循环单元 |
| **NER –** Named Entity Recognition | 命名实体识别 |
| **KB –** Knowledge Base | 知识库 |
| **BP** –Backpropagation | 反向传播 |
| **BPTT –**Backpropagation Through Time | 通过时间反向传播 |
| **ResNet –** Residual Networks | 残差网络 |
| **VGGNet** –Visual Geometry Group Network | 视觉几何群网络 |
| **ILSVRC** – Imagenet Large Scale Visual Recognition Challenge | 大规模视觉识别挑战赛 |
| **COCO –** Common Objects in Context | 上下文公共对象 |
| **ReLU –** Rectified Linear Units | 整流线性单元(激活函数) |
| **CPU –** Central Processing Units | 中央处理器 |
| **GPU –** Graphics Processing Units | 图形处理器 |
| **TPU –** Tensor Processing Units | 张量处理器 |
| **ASIC** **–** Application Specific Integrated Circuits | 专用集成电路 |
| **DNN** **–** Deep Neural Networks | 深度神经网络 |
| **AI** **–** Artificial Intelligence | 人工智能 |
| **Json –** JavaScript Object Notiation | JavaScript对象符号(一种轻量级的数据交换格式) |
| **MSE –** Mean Square Error | 均方误差 |
| **BEP –** Break-Event Point | 平衡点 |
| **AUC–** Area Under the Curve | 曲线下面积 |

# 绪论

信息科技持续不断的发展，使得人们的工作和生活变得愈来愈便捷。然而面对当今信息爆炸的数据洪流，要想从中获得有价值的信息则变得举步维艰。世界上的许多知识是用自然语言文字记录的，但如何以这种形式有效地加以利用是一项重大挑战[1]。本课题研究的关系抽取正是信息抽取[1][2]的一项核心而又艰巨的任务，同时又是构建语义搜索[3]、问答系统[4]等知识图谱[5]相关应用的重要基础。

近年来，关系抽取的研究越来越受到国内外研究学者的关注，在这方面的研究也取得了比较显著的进展。但是由于关系抽取这个任务本身的复杂性，现有的模型依然存在许多的局限，关系抽取的效果差强人意。

本课题就是要研究并解决目前的关系抽取模型所面临的困境，如无法有效的抽取长而复杂或无直接对应关系触发词的句子中涉及的实体对关系。在本章中，首先，简要介绍了本课题的背景及来源；其次，概述了国内外在关系抽取方面的研究，并剖析了关系抽取当前所面临的困境及本课题要解决的重难点问题；然后，简明扼要的阐述了本课题研究的主要内容是什么，以及研究的意义在哪里；最后，说明了下本文的行文组织结构。

## 课题背景及来源

由于网络应用爆炸式增长，互联网中每时每刻都有海量的数据在不断地生成，这标志着大数据、云计算的时代已经来临。这一方面给人们的生活方式带来了极大的变革，另一方面，却也使得人们从这些海量数据中获取有效信息变得愈发困难。因此我们亟需一种能够从海量的自然语言文本中抽取有用信息的技术，从而更加充分发挥这些文本语料的价值，信息抽取正是应上述之需而诞生的。

关系抽取是信息抽取中极为重要而又复杂的任务，然而当前已有的模型在关系抽取上的效果却不尽人意，在稍微复杂的语料实例中往往不能抽取出有效的实体对关系，这极大地限制了关系抽取在大规模自然语言文本语料上的应用，如以关系抽取为基础的语义检索、问答系统等知识图谱相关应用。因此，关系抽取的研究依然任重而道远。基于此背景，本课题展开了关系抽取这一基础性研究，以解决现有模型遇到的困境，从而提高关系抽取的整体效果。

本课题来源于教育部社科研究重大委托课题“科研管理信息化建设”，并受到了教育部哲学社会科学繁荣计划建设项目（2017-2018）、华中科技大学双一流建设项目（2017-2018）的资助。

## 课题现状及分析

### 国内外研究概况

由美国国防高级研究计划局(DARPA)资助的最后一届消息理解会议(**MUC**)首次引入关系抽取任务[6]。美国国家标准技术研究院(NIST)组织的自动内容抽取(ACE)评测会议中提出的一项重要任务就是实体关系识别[7]。ACE评测会议中关系语料的来源包括新闻专线、新闻广播、通过OCR识别的报纸、新闻组、电视访谈、网络日志等等。和MUC比较起来，关系识别任务在ACE评测会议中的数据规模、语种数目等有比较大的提升。MUC和ACE会议定义的关系抽取任务存在很大的局限性，其中的实体只能是人名、地名、组织名、机构名等命名实体，定义的关系类型也非常少，比如位置区域关系、国籍关系、任职关系等。在MUC与ACE评测会议之后，SemEval[8]是信息抽取领域出现的又一个影响力很大的会议，该会议吸引了众多高校等研究机构学者参与到其中。SemEval-2007[9]定义了普通名词或名词性短语之间的语义关系，但提供的语料库规模较小，SemEval-2010[10]任务8在SemEval-2007基础上进行了丰富和完善，引发了普通名词或名词性短语之间语义关系抽取的研究新高潮。

关系抽取的研究愈来愈受到国内外研究学者的关注，且其研究的方法也是层出不穷。这些方法概括起来主要包括基于传统有监督的方法、基于远程监督的方法与基于神经网络的深度学习方法这三大类。

1. 基于传统有监督的方法

在关系抽取中，目前有监督的方法能抽取更有效的特征，其准确率和召回率都更高，因此有监督的方法受到了国内外研究学者更多的关注。

有监督的关系抽取方法得要有事先标注好的训练语料，然后再从标注语料中自动学习要训练的关系对应的抽取模式。典型的方法有：Miller, Scott等人[11]的基于句法解析增强的方法, Kambhatla, Nanda等人[12]的基于逻辑回归的方法,基于核的方法[13],包括Zhao and Grishman等人[14]的解析树核、Bunescu and Mooney等人[15]的子序列核等, Culotta, Aron等人[16]的基于条件随机场(CRF)的方法。然而在关系抽取中,阻碍这些方法的主要问题在于人工标注训练语料耗时耗力。

1. 基于远程监督的方法

针对传统有监督方法存在的上述局限性，Mintz et al[17], Zhang et al [18], Krause et al [19]等人将远程监督的思想应用到了关系抽取的建模过程中。他们将纽约时报(NYT)新闻语料与大规模知识库FreeBase[20](包含7300多个关系和超过9亿个实体)进行实体对齐。远程监督假设为：一个包含两个实体的句子蕴含了该实体对在知识库当中的语义关系，然后将这个句子作为该实体对在知识库中所蕴含关系的训练正例。作者利用远程监督方法来自动标注训练语料，然后抽取语料特征并训练关系抽取学习器，有效解决了关系抽取在大规模训练语料上的标注问题。后续众多研究学者从不同的方面对远程监督方法进行了改进。例如，Takamatsu等人[21]提出了实体对齐技术的改进方案，缓解了噪声数据带来的性能下降问题；Yao等人[22]提出了基于无向图模型的关系抽取方法；Riedel等人[23]则增强了远程监督的假设，与 Mintz的相比错误率减少了31%，提高了关系抽取的整体效果。

上述基于远程监督的假设均是一个实体对只对应一种关系。然而实际中许多实体对具有多重关系。如“Steve Jobs founded Apple”和“Steve Jobs is the CEO of Apple”就表达了实体对之间的多重关系。为解决在很多实际场景中基于上述假设的不合理性，Hoffmann等人[24]提出了多实例多标签(MIML)方法来建模关系抽取，该方法考虑了一个实体对存在多重关系的场景。类似地，Surdeanu等人[25]也提出了多实例多标签+贝叶斯网络(MIML+BN)来抽取自然语言文本语料中实体之间的语义关系。关于生成负样本的这一严重缺陷，Bonan等人[26]提出了在实体对级别上仅仅从正例与未标记样例中学习的算法，Angeli等人[27]提出了一种利用少量精心选取的实例对远程监督关系抽取器进行局部监督的方法。

1. 基于深度学习的方法

应用传统机器学习的方法尽管已经取得了很不错的效果，但这些方法过于依赖词性标注(POS)[28]、语义角色标注(SRL)[29]等自然语言处理(NLP)[30]任务。这些任务在经过NLP工具处理后存在大量的错误，而这些错误在关系抽取中被不断传播放大，进而影响关系抽取的整体性能。近些年来，由于深度学习的迅速崛起，尤其是其在计算机视觉领域的革命性突破，越来越多的研究学者倾向于用深度学习的方法来建模关系抽取。值得关注的一点是，深度学习允许由多个处理层组成的计算模型学习具有多个抽象级别的数据表示[31]。

Socher等人[32]提出采用递归神经网络来建模关系抽取，该方法考虑了句子的句法结构信息，但没有考虑两个实体在句子中的位置及语义信息。Zeng et al[33], Nguyen et al[34],Xu et al[35]等人将卷积神经网络(CNN)运用于关系抽取任务中，该方法同时考虑实例的词向量和词的位置向量，在关系抽取中较好地考虑到了句子中的实体信息。之后，Santos等人[36]设计了一种新的卷积神经网络结构，并使用新的损失函数来建模关系抽取，更加有效地区分了各种类型的关系。

深度学习方法在人工标注的语料库上效果非常地显著，但是类似于基于特征的方法，基于深度学习方法的关系抽取模型也同样面临着一个比较严重的问题，即人工标注语料耗时耗力。为此，Zeng等人[37]提出通过将CNN网络结构与远程监督结合起来进行关系抽取，这和基于特征的方法相比，在很大程度上解决了基于深度学习的关系抽取模型在大规模语料上的应用问题，其在NYT数据集上进行的关系抽取效果非常地显著。Miwa等人[38]提出端到端的神经网络，该网络用双向LSTM[39]来建模关系抽取。Peng et al[40]、Song et al[41]等人探究了一种基于图LSTM的通用关系抽取框架，该框架很容易扩展到跨句子的n元关系抽取。为了有效缓解噪声数据带来的性能下降问题，Lin等人[42]将选择性注意力机制应用到了关系抽取中，该方法可以根据指定关系为袋子中的每个实例分配权重，通过训练学习从而使有效实例的权重不断提高，而带噪声的实例权重不断下降。在其他相同情况下，与不带注意力机制的模型相比，该方法的整体效果提升较大。类似地，Wang et al[43]、Du et al[44]等人提出一种新的多级别的注意力机制用于关系抽取。Wu et al[45]、Qin et al[46]等人将对抗模型引入到关系抽取中，从而进一步增强了模型的泛化能力。此外，也有改善子句模式的自训练[47]模型、全局关系嵌入[48]模型等其他各种各样的关系抽取方法。

### 重点难点问题

本课题对关系抽取这一基础性任务进行了研究，为提高关系抽取的效率，提出和设计了新的关系抽取模型，并将提出的模型应用到科研文档中，以构建一个语义检索系统。本课题在进行关系抽取的研究与实现过程中，重点问题主要为以下几点：

1. 大规模语料的标注

语料标注在关系抽取建模过程中是一个非常重要的问题，尤其是对于大规模语料，人工标注耗时耗力，这就亟需一种自动标注语料的技术。本课题是在关系抽取研究中广泛采用的NYT数据集上进行实验的，该数据集规模庞大，因此本文采用了远程监督[17]来解决语料标注的问题。

1. 噪声实例的影响

应用远程监督来自动标注自然语言文本语料在解决了大规模语料的标注问题的同时，也给关系抽取模型带来了很多噪声实例，从而影响关系抽取的效果。为了缓解噪声实例带来的性能下降问题，本文采用了选择性注意力机制[42]。

1. 模型泛化能力

很多时候模型在训练集上能取得较好的结果，而在测试集上往往性能会有所下降甚至很差，这通常是由于过拟合造成的。一般可以采用正则化、提前终止、dropout等策略来解决，本文在进行关系抽取建模时，除了采用这些策略之外，还利用了对抗训练[45]，即在模型训练过程中，对嵌入层表示一个微小的反馈扰动，从而进一步增强模型在测试集上的泛化能力。

本课题深入探索并分析了国内外研究学者在关系抽取方面的研究工作，发现目前已有的研究方法存在很多的局限性，这些正是本课题要攻克的难点所在，其中主要分为以下两点：

1. 无直接关系触发词的句子的关系抽取

在自然语言文本语料中，普遍存在类似图1-1第一个句子的情况，即句子中不存在要抽取关系的直接触发词，而已有的关系抽取模型在这种情景下往往是无法抽取出有效关系的。

1. 长而复杂的句子的关系抽取

对于图1-1第二个句子这样长而复杂的句子，由于两个实体之间的距离隔得比较远，现有的方法对实体对上下文语义关系的学习比较困难，也常常不能提取出这类句子中实体之间的正确关系。



图1-1 关系抽取模型例子

## 研究内容及意义

### 研究内容

本课题研究的关系抽取是信息抽取的一项核心任务和重要环节。信息抽取旨在从大规模非结构化或半结构化的自然语言文本中抽取结构化信息，其主要任务是识别句子中的实体、抽取实体之间的语义关系以及其中涉及的事件[49]。关系抽取主要是从自然语言文本语料中抽取实体对的语义关系，然后表示成三元组的形式以构成关系事实，如：<Jobs, founder, Apple>表示乔布斯是苹果公司的创始人这样一个关系事实。这些关系事实是构建其他上层应用如语义检索、问答系统等的重要基础[50]。本课题研究内容的应用范围并不局限于科研文本，对于包括新闻文本、百科文本等在内的文本数据也同样是适用的。

对于关系抽取的基础性研究，为了方便与国内外其他研究学者论文中的方法进行比较，本文是在NYT公开数据集进行实验与模型评估的，进而验证本文提出的模型的有效性。本课题的研究内容主要有以下几点：

1. 解决无直接关系触发词的关系抽取的困境

现有的关系抽取方法通常采用词嵌入[51][52]及位置嵌入[53]作为模型的分布式表示层，而这种表示存在1.2.2小节所述的第一个难点，即无法有效的抽取无直接关系触发词的句子中的关系。为了解决这一困境，本文提出实体类型嵌入(ETE)。ETE和词嵌入、位置嵌入一起作为模型的联合嵌入层。

1. 处理长而复杂的句子的关系抽取的问题

在应对长而复杂的句子时，现有的模型也常常不能提取出正确的关系，如图1的第2个例子。我们认为，以上抽取失败是由于缺失深层语境下的隐藏信息导致的。为解决这一问题，我们在提出的ETE的基础上，精心设计了循环分段残差网络(RPRN)结构。RPRN网络结构深而复杂，能够有效编码句子上下文中的潜在信息，并学习到句子上下文中深层次的语义信息。

### 研究意义

本课题深入探索并分析了国内外研究学者在关系抽取方面的研究，提出ETE，并精心设计了RPRN网络结构。本文提出的模型有效解决了当前关系抽取研究长期以来面临的困境，如无法很好地抽取无直接关系触发词或长而复杂的句子中的实体对关系。本课题开展的这一基础性研究，极大地提高了关系抽取的整体性能，这对关系抽取的研究与发展具有重要的意义。

通过机器学习相关技术自动抽取自然语言文本语料中蕴含的关系事实，人们可以更加有效地获取文本中涉及的各主客体之间的语义关系，有效地避免了以人工阅读的方式从不断生成的海量文本中提取关系事实等结构信息，这对自然语言文本语料的信息化建设具有重大的意义。同时抽取的关系事实是构建语义检索、问答系统等上层应用的重要基础。

本课题提出的关系抽取模型，不仅对本文要构建的语义检索系统提供了基础，而且在一定程度上为今后关系抽取方面的研究指明了方向，本课题的研究具有广泛的实际应用和较高的理论研究价值。

## 本文的组织结构

本文共计包含五个章节，各章节的内容简述如下：

第一章：绪论。在本章节中，首先，简要介绍了本课题的背景及来源；其次，概述了国内外研究学者在关系抽取方面的研究，并剖析了关系抽取当前所面临的困境及本课题要解决的重难点问题；然后，简明扼要的阐述了本课题研究的主要内容以及研究的意义，最后描述了下本文的组织结构。

第二章：相关理论及技术概述。本章主要介绍了在进行关系抽取研究中涉及的关键理论，以及在建模过程中用到的核心技术。包括远程监督、循环神经网络、残差网络、选择性注意力机制、对抗训练等重要理论，还有自然语言处理工具StanfordNLP、深度学习框架Tensorflow等相关技术。

第三章：关系抽取研究与算法设计。本章深入分析了关系抽取研究中遇到的难点问题，并针对现有方法遇到的困境，提出ETE并设计RPRN网络结构。然后概述模型的总体设计架构，最后详述模型的各个核心模块及算法设计。

第四章：实验及模型评估。本章首先概述了一下实验数据预处理的大体流程；接下来，简单介绍本文所设计的模型的超参数的设置及模型评估的相关指标；然后，以表格和P-R曲线图的形式对我们提出的模型和现有的模型在各项评估指标上进行比较；最后，对模型进行关系抽取的不同结果做一个具体的案例分析，以从不同角度验证本文所提出的模型的有效性。

第五章：总结与展望。本章节简要总结了本课题开展的关系抽取的研究工作，并对其未来的发展和研究做了进一步的探讨和展望。

# 相关理论及技术概述

本课题开展的关系抽取这一基础性研究，主要基于机器学习相关理论与算法来建模。本文是在NYT大规模公开数据集上进行的实验，首先为了解决人工标注自然语言文本语料耗时耗力的问题，本文应用了远程监督。利用循环神经网络(RNN)[54]进行序列学习的优势，以及残差网络[55]深度编码模型表示的能力，本文设计了一种新的神经网络结构RPRN作为关系抽取编码器。为进一步增强模型在测试数据集上的泛化能力，在关系抽取建模过程中还引入了对抗训练[45]。

关系抽取的建模依赖于命名实体识别(NER)[56]任务，本文采用的是StanfordNLP[57]工具来完成这一任务。在关系抽取的设计过程中，我们应用了深度学习的相关方法，具体采用的框架是google推出的Tensorflow[58]。

## 远程监督

关系抽取就是要抽取句子中包含的实体之间的语义关系，然后表示成如<Jobs, founder, Apple>这样的三元组形式。在传统有监督的关系抽取中，首先要对训练语料进行人工标注，因为关系抽取器模型本身是无法给要训练的实体对关系打标签的。然而，人工标注语料存在耗时耗力的问题，这极大地限制了大规模语料在关系抽取方面的应用。因此，Mintz等人[17]提出利用远程监督来自动标注大规模语料。

### 远程知识库

应用远程监督来自动标注文本语料需要一个远程的知识库(KB)[59]。知识库是一种用于存储被计算机系统使用的复杂结构化或非结构化信息的技术[60]，包括专家系统所使用的规则集，文本中蕴含的关系事实、涉及的事件等等。本课题研究的关系抽取采用的远程监督知识库为FreeBase[20]，如下表2-1所示：

表2-1 FreeBase知识库-关系事实

|  |  |  |
| --- | --- | --- |
| **relation** | **subject\_entity** | **object\_entity** |
| /people/person/nationality | Sanath Jayasuriya | Sri Lanka |
| /location/location/contains | Russia | Arkhangelsk |
| **…** | **…** | **…** |

### 命名实体识别

命名实体识别(NER)是自然语言处理中一项非常重要的任务[61]，也是本课题要开展的关系抽取研究的重要基础。NER的主要目的是识别自然语言文本语料中的人名、地名、机构名、组织名、日期时间、专有名词等具有特殊含义的实体。在关系抽取中要应用远程监督，首先就要识别出文本语料中的实体对，然后再结合知识库来自动标注。NER识别句子中的实体对如下所示：

**Steve Jobs** was the co-founder and CEO of **Apple** and formerly **Pixar**.

在上述实例中，使用NER可以识别出三个实体，这三个实体在该句子的上下文中可以组成两对实体，接下来就是自动标注实体对在对应实例中的语义关系。

### 实体链接

识别出了自然语言文本语料中涉及的实体后，就要采用远程监督来自动标注训练语料。在自动标注语料时，需要应用实体链接[62] ，如图2-1所示，即将训练语料中所提及的实体对链接到大规模远程KB中。

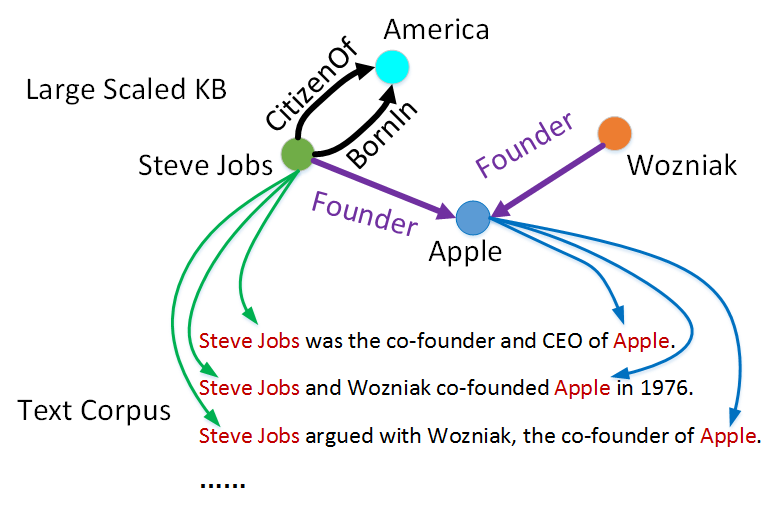


图2-1 远程监督实例

远程监督假设为：一个句子包含的实体对蕴含了该实体对在知识库中对应的关系，然后在模型训练时，将该句子作为其实体对对应关系的训练正例。最初的远程监督假设一个实体对只对应一种关系，但在自然语言文本语料中，大量存在一个实体对拥有多种关系的场景。因此，后续有很多的研究学者从不同角度对远程监督进行了改进，以适应其在大规模语料上的应用。

## 循环神经网络

循环神经网络(RNN)是一种处理序列数据的网络结构[63]，例如处理可变长度的自然语言文本序列。相较于全连接的前馈神经网络只有相邻层之间的神经元有连接而言，如图2-2所示的RNN的网络结构的每一个隐藏层中的神经元也有连接，这使得RNN在几个时间步内可以共享相同的参数，有效避免了在不同位置上重复学习相同的特征规则。



图2-2 多层循环神经网络

对于任何一个图灵可计算的函数而言，都可以通过图2-2这样一个有限维的RNN结构来模拟计算[63]，从这一点来考虑的话，RNN可以说是万能的。在RNN的设计过程中，主要包含以下几个重要的设计模式：

1. RNN模型每一个时间步都有输出，并且隐藏层神经元之间有如图3所示的循环连接；
2. RNN模型每个时间步都会产生一个输出，只有从一个时间步的输出到下一个时间步的隐藏层神经元有循环连接；
3. RNN模型隐藏层神经元之间有循环连接，单个输出会在读完序列之后产生。

### 通过时间反向传播

反向传播(BP)算法自20世纪70年代被引入以来，作为一种神经网络模型的优化算法，一直是深度学习的主力[64]。而RNN是一种具有长期记忆能力的序列学习网络模型，普通的BP算法并不适用于这种模型的优化。在RNN模型中，采用的是通过时间反向传播(BPTT)[65]的神经网络优化算法，这是一种更有效的关于在RNN中传递参数的方法。在RNN的前向传播中，以特定的初始状态为起点，对于每一个从t=1到t=T的时间步，应用如下方程式来进行更新：

(2-1)

(2-2)

(2-3)

(2-4)

其中对于每一个时间步t，表示输入，表示隐藏层，表示输出，为模型输出向量，W、U、V表示权重矩阵，b、c为偏置向量。模型的损失函数为：

(2-5)

为在给定到t时刻的序列及t时刻的目标下的负对数似然，模型的总损失是所有时间步损失的累积之和。如下所示公式为通过BPTT优化算法来计算RNN中反向传播的梯度：

(2-6)

(2-7)

(2-8)

(2-9)

(2-10)

(2-11)

为对角矩阵，是关于t时刻与隐藏层神经元i关联的Jacobian，指的是包含这个元素。

### RNN变体

在普通RNN模型的训练过程中，存在一个众所周知的问题，及梯度弥散和梯度爆炸[66]。梯度弥散和梯度爆炸本质上可以归结为梯度不稳定问题，其主要原因是神经网络在优化时进行反向传播的过程中，同一矩阵经过了多次相乘，梯度倾向于消失或超过某个阈值，导致后续的神经元无法更新参数，神经网络的优化无法正常进行。

为了解决RNN在训练过程中遇到的上述问题，一些学者提出了RNN相关的一些变体结构。如极为经典的LSTM网络，其结构原理如图2-3所示。



图2-3 LSTM结构原理图

LSTM结构设计的巧妙之处在于：除了RNN拥有的外部循环外，内部结构还有自循环，这种结构具有学习长期依赖的能力。其主要由3个门来控制，分别为输入门，遗忘门，输出门。如下所示公式表示其结构内部各状态的输出：

(2-12)

(2-13)

(2-14)

(2-15)

(2-16)

(2-17)

其中、、为输入门的权重矩阵，、、为遗忘门的权重矩阵，、、为输出门的权重矩阵，、、分别为3个门的偏置向量，为t时刻隐藏状态的输出向量。

LSTM的结构看起来比较复杂，那么该结构的所有构成都是必须的吗？有没有一种机制可以动态控制模型的遗忘功能呢？相较于LSTM，GRU[67]是一种更加简化的门控循环神经网络结构， 其结构原理如图2-4所示。



图2-4 GRU结构原理图

从图2-4与图2-3的对比中可以看出，GRU在原LSTM结构的基础上将门控单元简化成了2个，即更新门和重置门，其公式分别表示为：

(2-18)

(2-19)

其中，、表示更新门的权重矩阵，、表示重置门的权重矩阵，和分别表示两个门的偏置向量，隐藏层状态向量为，由以下公式表示：

(2-20)

(2-21)

GRU结构已经不再显示包含记忆单元，但它至少和LSTM一样有效，并且收敛速度更快。与LSTM的最大不同在于，GRU中的单个门控单元可以同时控制遗忘功能和状态更新。

LSTM、GRU等门控机制的RNN模型都能够有效解决梯度弥散的问题，但处理不了梯度爆炸的问题。梯度截断(Gradient clipping)是一种非常流行且有效的处理梯度爆炸问题的方法[68]，在模型训练过程中，该方法通过截断梯度来使得反向传递的梯度值总是小于给定的阈值。

## 残差网络

深度学习冠之以“深度”一词，以区别于早期的浅层神经网络模型。浅层网络模型最大的缺陷在于其很难学出具有不同抽象级别的数据表示，而深层模型则通过堆叠多层次网络结构有效地解决了这一难题。然而，越深的网络模型越难以训练和优化，甚至有时候训练出的效果还不如浅层模型。

为了解决深层模型训练及优化难的问题，Kaiming He等人[55]提出了残差网络(ResNet)深度模型。作者使用了深度达152层的ResNet在ImageNet数据集[69]进行评估，虽然该网络结构的深度是VGGNet-19模型的8倍，但其复杂度却比VGGNet-19低，并且在ImageNet测试集上的效果达到了3.57%的错误率，从而赢得了ILSVRC和COCO 2015比赛[[1]](#footnote-1)中多项任务的第一名。

ResNet的基本单元是卷积池化模块，然后以此为基础进行残差学习，进行残差学习的网络结构由恒等映射块来构造，块中的单元通过捷径来连接，最后由这些基本残差构造块通过堆叠的方式来设计网络。

### 深度卷积网络

深度卷积网络在图像分类方面取得了一系列的突破[70]，其主要由一系列的卷积、池化以及经过激活函数的非线性变换等操作构成。卷积的特性是输入表示的潜在模式比整个输入小且出现在不同区域，池化的则是通过再次抽样来获取主要特征而不失去原始输入的表征。

卷积操作的原理示意图如图2-5所示，每一层的卷积操作都会对应一组过滤器，然后用该组过滤器中的每一个分别与上层输出进行点乘操作。通过不同层次的卷积操作，输入表示被编码成了具有不同抽象级别的局部潜在特征。

卷积操作之后会进行池化操作，即将一层神经元簇的输出组合成下一层的某个神经元。池化操作的类别有很多种，如最大池化、平均池化等等。顾名思义，最大池化就是使用前一层每一个簇的最大值作为该层当前神经元的值，其操作原理如图2-6所示；以此类推，平均池化即使用前一层每一个簇的平均值作为该层当前神经元的值。当然，卷积和池化操作后还会经过激活函数处理，如sigmoid、ReLU[71]等激活函数，激活函数的作用就是对上一层的输入做一个非线性变换。



图2-5 卷积操作原理



图2-6 最大池化操作原理

### 残差学习

假定将作为一个由几个堆叠的层(不一定是整个网络)匹配的潜在映射，其中是这些堆叠层的第一个的输入。假定一个复杂的函数能够由多个非线性层来渐进近似，从一定程度上来说，它能够由残差函数来渐进近似，即如公式2-22所示：

(2-22)

在以上公式中假定输入和输出具有相同的维度。并不是要堆叠层去近似，而是显示地让这些层近似，原始函数因此而变成了。尽管两种形式都要能够渐进近似期望的函数，但学习的难度却有很大的不同。残差学习不仅训练优化的复杂度低，而且效果极好。残差函数是残差基本构造块的基础。

### 通过捷径的恒等映射

上述提法可以通过拥有快捷连接的前馈神经网络来实现。快捷连接，即在构建块中跳过一或多个层。在残差学习中，快捷连接就是执行简单的恒等映射，然后将该层的输出添加到下一堆叠层的输出中。残差网络的残差学习基本构建块可以由公式2-23的形式来定义：

(2-23)

其中和分别表示所考虑层的输入和输出向量，表示要学习的残差映射，对于图2-7构建块中的2层，，为ReLU激活函数，为简化表示偏置项被忽略了。操作通过快捷连接和元素级相加来执行。公式2-23既没有引入额外的参数，也没有增加计算的复杂性，这不仅在实践中极具吸引力，而且在残差网络与普通网络的比较中也相当重要。公式2-23中的和必须具有相同的维度，如果不是这种情况的话，我们可以通过快捷连接执行一个线性投影来匹配维度，如公式2-24所示的变换：

(2-24)

当然，在公式2-23中也可以使用方阵，但对于处理退化问题来说恒等映射是足够的，故一般只用于维度匹配上。

### ResNet架构

ResNet基于普通的深度卷积网络，由堆叠的残差构造块组成，并在其中加入了快捷连接。ResNet架构与其他深度神经网络模型的对比如图2-7所示(左：VGG-19，中间：34层的普通深度卷积网络，右：34层的ResNet)。值得注意的是，和VGGNet相比，ResNet有更少的过滤器和更低的复杂度，且其在众多的机器学习比赛项目中取得了更好的效果。

当输入和输出具有相同的维度的时候可以直接使用公式2-23，当维度在增加的时候，需要考虑以下两点：

1. 捷径同样采用恒等映射操作，为了扩张维度，会再附加额外的零项填充，这并不会带来其他的参数开销；
2. 公式2-24中的投影捷径用于匹配维度。

对于以上这两点而言，当快捷连接跨了两个大小的特征映射时，它们所执行的步幅都是2。

ResNet受深度所蕴含的重大意义的驱动，通过残差学习、恒等映射、快捷连接等手段构造基本块，再由基本块以堆叠的方式来构造网络结构，使得学习更好的网络就像堆叠更多的层一样那么容易，网络的训练与优化也变得不是那么的困难，这对各种机器学习任务在大规模数据集的广泛应用具有重要意义。



图2-7 残差网络与其他深度神经网络结构对比

## 注意力机制

在关系抽取的建模过程中，通常采用远程监督的方法来解决标注训练语料耗时耗力的问题，但远程监督会出现错误标记的问题，错误标记的实例称为噪声，噪声实例会使关系抽取的整体性能恶化，从而影响模型效果的提升。因此要有一种机制能从实体对对应的袋子中挑选有效的句子作为特定关系与该实体对的训练正例。最初的远程监督假设是将袋子中的所有句子都作为训练正例，即对袋子中实例做平均处理，经选择器后的输出向量表示为:

(2-25)

其中b表示相应袋子的实例个数。这种方式存在大量噪声实例参与训练过程，使得关系抽取的性能难以提升。针对这个问题，Riedel等人[23]提出从实体对对应袋子中选出最有效的一个作为相应关系的训练正例，用该方法选择的输出向量V为：

(2-26)

其中B为选择函数，参数为袋子中实例的索引，用于选择标记矩阵对应列标记向量，R为关系矩阵，bias为偏置向量，表示第i个袋子的标记。该方法选择了最有效的实例作为训练正例，极大地缓解了噪声实例带来的性能下降的影响，但同时也损失了大量有效的句子，进而限制了关系抽取模型的整体效果。近些年很多学者在采用深度学习方法解决各种问题时应用了注意力机制，Yankai Lin等人[42]也将这种机制引入到了关系抽取的建模过程中。

## 对抗训练

尽管深度神经网络在各种不同的领域取得了巨大的成功，但其对输入数据中的噪声倾向于太过自信。一个神经网络模型往往在训练集上可能表现出很好的效果，但在测试集上的预测却不尽人意，这通常是由于过拟合造成的，这类问题很多情况下都可以采用正则化的策略来处理。对抗训练(Adversarial Training)就是一种通过生成对抗噪声到训练数据中的正则化分类算法[45]，以增强模型在测试集上的泛化能力。使用对抗训练编码实例的计算图如图2-8所示。

其中表示第i个训练实例，表示关于的对抗扰动。对于一个实体对，关系的概率及损失函数表示如下：

(2-27)

(2-28)



图2-8 使用对抗训练编码实例的计算图

其中为投影矩阵，为偏置向量，表示关系个数，为给定的参数集合。对抗训练会在训练过程产生持续的微小扰动，然后将产生的对抗噪声加入到词嵌入层。引入对抗训练后模型的损失函数表示如公式2-29所示，表示当前值的一个固定副本，因为从计算的角度来看，公式2-30是比较难处理的。

(2-29)

(2-30)

## StanfordNLP

StanfordNLP是斯坦福NLP组研发的一套NLP软件，它为主要的计算语言学问题提供了统计NLP、深度学习NLP和基于规则的NLP，这些工具将人类语言技术的需求结合到了应用中，并在工业界、学术界和政府中被广泛使用[57]。

本课题使用StanfordNLP来进行关系抽取模型训练前的数据预处理相关工作，主要包括用其中的POS Tagger工具对原始自然语言文本语料进行分词，然后用NER工具来识别语料中的命名实体，其示意如图2-9所示。



图2-9 命名实体识别实例

## Tensorflow

由google大脑团队研发的深度学习框架Tensorflow，是一种大规模、异构环境下运行的机器学习系统[72]，不仅大量用于机器学习方面的研究，而且在计算机视觉、语音识别、自然语言处理等计算领域被广泛应用。

数据流图是Tensorflow中一个非常核心的设计模式，可以用其来建模任何的计算任务。数据流图是由节点和边构成的有向图，节点代表计算操作，也可表示输入数据的起点或输出数据的终点，边表示两节点之间的输入输出关系。顾名思义，张量(Tensor)与流(Flow)是Tensorflow中极为重要的两个概念。张量可以简单的理解为多维数组，流即为张量在计算图各节点上流动并通过节点上的计算而相互转化的过程。图2-10是Tensorflow中一个训练管道的数据流图的原理图，其中包含了用于读取输入数据、预处理、训练和检查点状态的子图。Tensorflow将数据流图的各节点映射到集群中多台机器上，其中一台机器跨多个计算设备，包括多核CPUs、通用GPUs和定制的被称为TPUs的ASICs，被分配到不同计算设备的各节点可以并行异步的进行计算。这种架构为应用程序开发人员提供了很大的灵活性：在其之前的“参数服务器”的设计中，系统内置了共享状态管理，使得开发者可以尝试新的优化和训练算法。



图2-10 Tensorflow中一个训练管道的数据流图

本课题开展的关系抽取研究利用深度神经网络来建模，至于具体的建模工具则采用目前最流行的Tensorflow。Tensorflow的计算图模型为本课题的研究及所提出的模型算法的具体实现提供了极大的便利和灵活性。

## 本章小结

本章主要概述了在进行关系抽取的研究和实现过程中所涉及的关键理论和相关技术。首先，介绍了在关系抽取建模过程中基于的重要理论，包括用于自动标注大规模语料的远程监督，进行序列学习的循环神经网络，学习语料深层次语义信息的残差网络，用于缓解噪声实例带来的性能下降问题的注意力机制，以及在测试集上提升泛化能力的对抗训练；最后，简述了用于关系建模的机器学习相关框架，包括用于自然语言处理的StanfordNLP工具，实现深度神经网络的Tensorflow框架。

# 关系抽取研究与算法设计

本课题充分调研并分析了国内外研究学者在关系抽取方面的研究，发现了现有方法都普遍存在的问题，这些问题严重阻碍了关系抽取效果的提升。针对已有方法所遇到的困境，本文提出**实体类型嵌入(ETE)**，ETE将和词嵌入、位置嵌入一起作为模型的联合嵌入层，此外，还精心设计了**循环分段残差网络(RPRN)**，已解决现有模型学习语料深层次语义信息的困难。

本章首先会对已有方法所遇到的困境进行深入分析，包括出现的困境是什么、为什么会有这样的困境以及解决的思路；然后，介绍下模型的整体架构设计；最后，分别阐述模型设计过程中的核心构成要素。

## 困境分析

关系抽取作为信息抽取的核心任务和重要环节，目的是为了抽取自然语言文本语料中涉及的实体对关系以构成如<Steve Jobs, founder, Apple>这样的关系事实，然后便可以利用这些关系事实去构建语义检索、问答系统等知识图谱相关的应用。现有的方法在应对如下这类句子(其中红色加粗词的是利用NER技术识别出的命名实体，而紫色加粗的斜体词为关系触发词)时抽取出来的关系效果良好。

**Steve Jobs** was the ***co-founder*** and ***CEO*** of **Apple** and formerly **Pixar**.

通过观察上述句子可以发现，首先这个句子比较简短，其次在这个句子中出现了要抽取的实体对关系对应的触发词，这就为该类句子中涉及的实体对关系的正确抽取提供了重要依据。然而，基于我们在NYT公开数据集上所进行的实验，现有的模型往往无法在以下具有挑战性的情况下抽取出有效的关系。

1. 句子中没有对应关系的直接触发词，如1.2.2小节的图1-1中第一个句子所示的情况；
2. 句子长而复杂并且两个实体之间的距离相隔比较远，如1.2.2小节的图1-1中第二个句子所示的情况。

我们认为现有模型在以上情况下抽取失败是由于缺失深层语境中隐藏的潜在信息造成的。为解决这一困境，我们提出一种基于联合嵌入和新的神经网络结构的方法。更具体地说，提出**实体类型嵌入(ETE)**，并将其和词嵌入、位置嵌入一起作为模型的联合嵌入层；此外，设计了**循环分段残差网络(RPRN)**的结构，和现有关系抽取模型所用的神经网络结构相比，RPRN大大增加了模型的深度，且其内部结构也是精心设计的，这样使其能有效编码句子上下文中深层次的潜在信息。

## 模型架构设计

受到融合更多更深次信息的思想启发，本文将提出的ETE与词嵌入、位置嵌入一起作为模型的联合嵌入层，以丰富模型的嵌入表示，从而有助于抽取句子所涉及实体对中潜在的隐晦关系；而设计的RPRN结构则弥补了现有关系抽取浅层模型存在的缺陷，即学出的特征太粗粒度以致于无法获取句子中丰富的潜在语义。根据以上提出的两点，本文设计的模型的整体设计架构如图2-1所示。

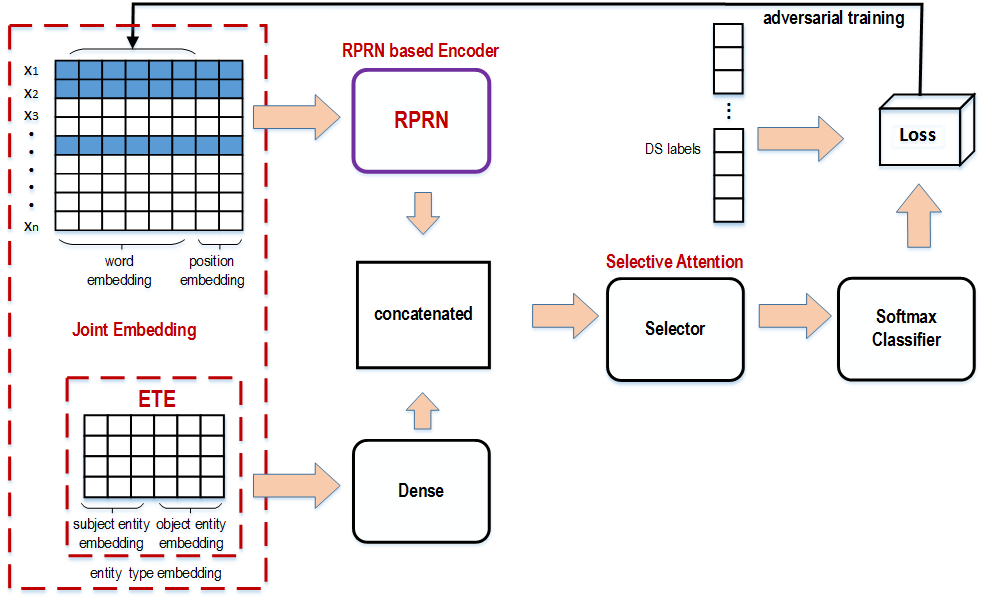


图3-1 提出的关系抽取模型整体设计架构

基于提出的关系抽取模型，本文设计了如图3-1所示的网络模型架构，其主要可以分为以下3个核心模块：

1. 嵌入层：由提出的ETE与词嵌入、位置嵌入一起作为模型的联合嵌入层，以丰富模型的分布式表示；
2. 编码层：对于词嵌入和位置嵌入，使用我们设计的RPRN进行编码，而对于提出的ETE则使用稠密的全连接网络(Dense)来编码；
3. 注意力层：为缓解噪声实例带来的性能下降问题，我们在模型训练过程中应用了选择性注意力机制。

以图3-1中的实例作为例子，最上面的表表示词嵌入与位置嵌入的联合，其中词嵌入是预训练好的，而位置嵌入是随机初始化的，下面的表表示我们提出的ETE，这是通过我们的算法1进行训练得到的，并将和词嵌入、位置嵌入一起作为模型的联合嵌入层。词嵌入与位置嵌入的联合表示将经过设计的RPRN网络结构进行编码，并将和ETE的编码输出进行加权连接。编码之后，拥有相同实体对的实例会被放到同一袋子中，然后，基于注意力机制的选择器将会尽可能从袋子中选择训练关系的相应正例。最后，会得到一个表示训练关系潜在模式的向量表示，最后一层的分类器将会基于这个向量表示输出最可能的关系。

## 联合嵌入

本文设计的关系抽取模型的嵌入层由词嵌入、位置嵌入及所提出的ETE一起联合构成，这是模型的一种分布式向量表示，其表示是否充分直接影响了关系抽取模型最后的整体效果。

### 词嵌入

词嵌入是将词从高维的词空间映射到低维的向量空间的一种分布式表示，语义相似的词在这个向量空间中相对也更加地靠近。由此可以看出，词嵌入可以表达词之间的语义关系。在我们提出的关系抽取模型中，使用的是skip-gram[52]预训练以得到词向量，但是预训练得到的词向量仅仅作为词嵌入的初始化值，之后再模型训练过程中会动态更新。

### 位置嵌入

为了有效利用句子中每个词的位置信息，位置嵌入[53]被应用到关系抽取的建模过程中。上面提到的位置信息指的是句子中的每一个词与相应的实体对之间的相对距离，如图3-2所示，然后将每一个词映射到一个向量空间中。此外，每一个词对应的两个向量都是随机初始化的，然后随着模型的继续训练而不断地动态更新。



图3-2 句子中每个词和相应实体的相对位置

### 实体类型嵌入

由3.1小节分析的困境可知，现有方法在所述情景中失败的主要原因是缺少深层语境中隐藏的潜在信息造成的，那有没有什么信息可以弥补这一缺陷呢？我们认为，实体类型信息在关系抽取中至关重要。假如<Steve Jobs, founder, Apple>这一关系事实在不同的实例中都有体现，现有模型通常只能从简单明了的实例中抽取出这一有效的关系，而在困境所述情景中会失败，如果有一种机制可以学出实例中实体对对应的实体类型，从直观上来讲，该类型信息可以有效避免错判和漏判的情况。

基于以上想法，我们提出ETE。但ETE的结构与词嵌入、位置嵌入的大相径庭，故不能直接将ETE和词嵌入、位置嵌入一起简单地连接起来作为联合嵌入层。我们采取的方案是将ETE作为表示层的一部分单独拿出来并使用Dense网络结构进行编码，如图3-1最下面的表所示。

给定某一实体对应的实体类型序列，训练时要最小化以下目标函数来学习ETE表示：

(3-1)

(3-2)

(3-3)

其中，为目标实体类型，表示大小为c的窗口内的上下文实体类型，和是分别作为目标和上下文时实体类型E的嵌入，负实体类型样本来自平滑的unigram[73]概率分布，即噪声分布：，表示实体类型在训练数据中出现的频率，为平滑参数()。

为了自适应学习率，目标函数应用AdaGrad来优化，给定目标上下文的实体类型对(和)以及k个来自于噪声分布的负样本()，梯度约为：

(3-4)

梯度被计算后，梯度下降就会执行更新、，以及。在训练过程执行参数更新后，我们就可以得到实体类型嵌入，即提出的ETE。ETE核心算法的伪代码如算法1所示。

|  |
| --- |
| 算法1：ETE嵌入表示训练 |
| for i = 1, … , n do      for j = -c, … , -1, 1, … , c do  从 分布中抽取k个负样本  使用AdaGrad执行更新、，以及  end for  end for |

其中，实体类型信息是借助远程监督来进行弱标注的，举一个简单的例子，假设有两个句子：“乔布斯是苹果的创始人”、“乔布斯手里拿着一个苹果...”，这两个句子中“苹果”的实体类型是不一样的，知识库中存在这个实体的相关类型：“公司”、“水果”等等。远程监督会用知识库中的这些类型对这两个句子的实体进行一个弱标注，但并不清楚哪个句子的“苹果”具体对应哪一种类型，而是我们模型在训练时为这些类型分配权重，比如训练“创始人”这一关系时，第一个句子中的“苹果”分配“公司”类型的权重就会很大，这是通过模型训练学习出来的。

在我们的关系抽取模型中，词嵌入的维度为，位置嵌入的维度为，然后它们连接在一起将一个训练实例转变成了一个矩阵，表示语料中句子的最大长度，表示联合词嵌入与位置嵌入的表示层的维度，矩阵将会喂给RPRN编码器。ETE的表示维度为，ETE表示为，表示语料中一个实体最多的类型个数。

## RPRN编码器

由于ETE被引入到联合嵌入层并被独立编码，我们的关系抽取模型的编码器就包含了两个部分：一个是用于编码词嵌入、位置嵌入的联合表示的RPRN编码器，另一个是用于编码ETE的Dense结构编码器。RPRN的结构设计原理如图3-3所示。



图3-3 RPRN的结构原理

由3.1节分析的困境可知，对于复杂的句子，关系抽取模型要想获取深层语境下隐藏的潜在表示是很困难的。受抽象分层思想的启发，本文设计了RPRN的网络结构，该结构主要包含三个部分：进行序列学习的循环结构模块，用于对序列特征表示进行深度分级抽象的残差模块，以及采样特征以获取主特征的分段池化模块。下面几节将详细阐述各模块的设计与原理。

### 序列学习

对于自然语言文本、语音等可变长度的序列，如何共享不同位置的参数信息，以及如何处理不同距离的依赖，是极为重要而又较难解决的问题。RNN网络模型的出现有效解决了序列学习的困难，然而对于长期依赖的问题，普通RNN处理的效果依然不尽如人意。为解决这一问题，有研究学者就提出了各种RNN的变体，如经典的LSTM、GRU等门控机制的循环网络单元。

本文在对关系抽取进行建模时，利用双向RNN来学习句子的上下文信息，其中使用GRU作为循环单元。GRU与LSTM都能较好地解决长期依赖问题，但GRU由于比LSTM少一个门控单元而更加简化，优化收敛的速度都更快，且GRU在我们的关系抽取模型中确实取得了更好的效果。GRU主要由更新门和重置门来控制，在我们的模型中，其表示分别如公式3-5和3-6所示。

(3-5)

(3-6)

其中，为句子中第个位置的词所对应的词嵌入与位置嵌入的联合表示，、表示更新门的权重矩阵，、表示重置门的权重矩阵，和分别表示两个门的偏置向量，当前位置隐藏层状态向量为，这是由上一时刻更新和保留的信息共同决定的，公式3-7与3-8为其推导过程。

(3-7)

(3-8)

在本模型中，GRU作为双向RNN的基本循环单元，通过横向随着序列延伸并纵向堆叠的方式来进行序列学习，如图3-1右下角部分所示。

### 残差结构

深度神经网络(DNNs)是极其强大的机器学习模型，它在一些诸如语音识别、视觉对象识别等困难的问题上依然有着非常优秀的表现。通常来说，网络结构越深，模型的表示能力就越强，但是随着网络层数的加深，模型也将变得越难以优化，甚至有些情况的表现还不如浅层模型。因此，ResNet[55]神经网络模型被提出，从而有效解决了深度网络模型优化难的问题。

为了学出深层语境下隐藏的潜在表示，本文在进行关系抽取建模时，设计的RPRN网络结构借鉴了ResNet的优秀思想，但RPRN中的残差结构与ResNet中的有所不同，我们模型的残差基本构造块的原理如图3-4所示。在RPRN中，由残差基本构造块得出的残差函数如公式3-9所示。

(3-9)



图3-4 残差基本构造块的原理

其中，为1维卷积函数，为激活函数，在恒等映射中，基本构造块的大小为，基本构造块的个数为，在残差结构之前有一个卷积层，残差块后就会执行分段池化，以在各段上采样主要特征。

### 分段池化

为了有效抽取实例中实体对的语义关系，我们利用分段池化来学习相应上下文中的结构化信息，这是一种极为有效的结构化学习[74]的方法。在分段池化中，我们引入隐码信息，表示为，其嵌入表示为，其中，m表示每一个批量的大小，n表示语料中句子的最大长度。执行分段池化操作后得到的结果为，其推导如公式3-10所示，其中表示模型编码器的隐藏层的大小。

(3-10)

关于公式3-10中的参数做以下说明：表示调整张量形状的函数，是执行分段池化操作的函数，表示扩展张量中的某一维的函数，且其值为1以保持该张量的整体维度不变，为查找隐码嵌入的函数，表示能使分段池化操作能有效执行而定义的一个常量，是残差模块的输出张量。图3-3中残差结构之后的分段池化操作的实现原理如图3-5所示。



图3-5 分段池化操作的原理

作为RPRN结构的一部分，分段池化层是为了从之前层表示的特征编码中采样出主要特征。分段池化，在我们的关系抽取的上下文中，表示根据实体对在句子中的位置而将当前实例分成三段，然后每一段分别执行最大池化操作。

至于ETE，编码之后使用的最大池化操作，其隐藏层的大小为。最大池化的结果会与分段池化的进行连接作为关系抽取器的编码器，其输出将会喂给下一阶段的实例选择器。

## 注意力选择器

在应对大规模自然语言文本语料时，远程监督方法通常被应用到关系抽取的建模中来，以解决手动标注语料耗时耗力的问题。然而，该方法同时也伴随着错误标记的问题，错误标记的实例被称为噪声。在关系抽取模型训练时，具有相同实体对的实例会被放到同一个袋子中。

远程监督最初的假设是将一个袋子中的所有实例都作为正例，这使得在模型训练过程中存在大量的噪声。为解决这一问题，Riedel等人[23]采取的是将袋子中最有可能的一个实例作为训练正例，这极大地缓解了噪声实例带的性能下降问题，但同时也丢弃了大量有效的实例，从而阻碍了关系抽取性能的进一步提升。因此，Lin等人[42]将选择性注意力机制应用到关系抽取中。

本文在进行关系抽取的研究时，也是用选择性注意力机制来解决噪声标注的问题，其原理如图3-6所示，其中，表示同一袋子中的第i个实例，表示第i个实例对应的特征潜在编码，表示由注意力机制学出的权重参数。经选择器得到的输出向量推导如公式3-11所示。

(3-11)

其中，表示挤压函数，用于去掉一个张量中所有所有维度值为1的维度，表示矩阵相乘操作的函数，表示扩展张量中的某一维的函数，且其值为1以保持该张量的整体维度不变，表示编码器的输出，表示关系矩阵，为语料的真实标记，表示偏置向量，表示袋子的大小。

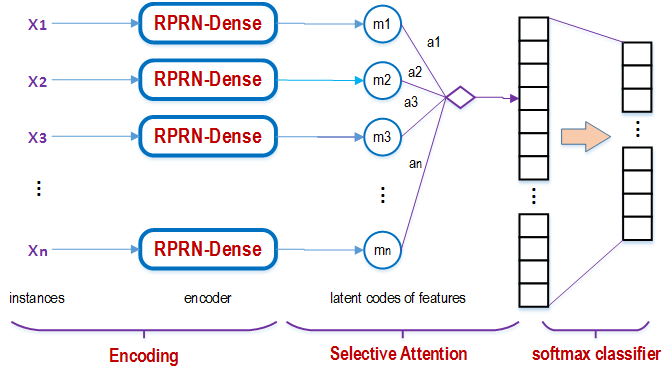


图3-6 选择性注意力机制

利用上述的选择性注意力机制，那些确实表达了其实体对在知识库中对应关系的句子就能够被挑出来作为正例喂给模型训练，而那些没有表达对应关系的则被当成噪声。更具体的说，这种选择性注意力机制保证了有效实例的权重会上升，而噪声实例的权重会下降，从而极大缓解了噪声带来的性能下降的问题。

我们模型的最后一层采用softmax来分类输出最可能的关系，损失函数使用的是交叉熵损失，其表示及其梯度分别如公式3-12和3-13所示，公式3-14则表示在给定实例下输出第k个关系的概率，其中为实例个数，为总共训练的关系数，和则分别表示实例和相应标记。

(3-12)

(3-13)

(3-14)

此外，在模型训练优化的过程中，会不断地给嵌入层中的词嵌入表示一个反馈扰动，即进行对抗训练，其原理及推导已在第2章2.5节描述，此处不再详述。实验表明，这种方式可以增强模型在测试集上的泛化能力。

至此，我们整个模型从嵌入表示到最后的分类输出总共构成了12层的深度神经网络，模型的核心训练流程如算法2的伪代码所示。

|  |
| --- |
| 算法2：ETE-RPRN模型训练核心算法 |
| 输入：训练语料，超参数，预训练的词向量  使用*算法1*训练  初始化参数，以及  ，， ，，    for to do      for to do  在上应用公式3-5到3-8  在上应用公式3-9  在上应用公式3-10        foreach do  尽最大可能从袋子中挑选出正例来  在上应用公式3-11  end for    更新参数    end for  end for  输出：抽取的关系 |

关于算法2中涉及的符号，我们接下来将做一些简要说明。作为训练语料，主要包含词、位置及实体类型信息，即分别表示为、及，表示模型在整个训练过程中的参数，表示位置向量，这是随机初始化的，表示实体类型向量，这个由算法1先预训练，再随着模型的继续训练而不断被更新，在算法2中直接用到的超参数包括最大轮数和批量大小。此外，表示隐藏层的输出，表示最后一层的输出，为整个模型的损失函数。

## 本章小结

本章主要对关系抽取这一基础性研究进行了分析并精心设计了我们自己提出的模型。首先，详细剖析了国内外研究学者在关系抽取方面所遇到的困境，包括困境是什么，为什么会有这样困境，以及解决的思路是什么；然后从整体上阐述了模型的架构设计及各主要构成要素；接下来，详细论述了模型的几个主要构成模块的设计，具体的来说，第一，联合嵌入层的设计，包括词嵌入、位置嵌入以及我们提出的ETE，并给出了ETE设计的详细推导及训练算法的伪代码，第二，我们设计的RPRN网络结构，其中包括进行序列学习的循环单元GRU，残差结构的设计和原理说明，以及分段池化的作用及原理阐释，第三，应用的注意力选择器，主要对其作用、推导及原理做了详细的阐述；最后，简单描述了对抗训练在本模型的作用及运作机制，以及给出了我们整个模型训练的核心算法的伪代码和相关说明。

# 实验及模型评估

前两章详细论述了关系抽取模型研究、算法设计思路。根据所提模型的设计与实现，本文在NYT公开数据集上进行了实验。这一章会介绍实验数据预处理流程以及实验的设置，并基于所做的实验进行模型评估与案例分析，在评估时会将本文提出的模型与其他研究学者的模型在多项评估指标上进行比较，并将已有模型抽取失败而本模型抽取成功的具体案例挑出来分析，以进一步验证本模型的有效性。

## 实验数据预处理

对原始的自然语言文本语料进行建模，第一步就是要对数据进行预处理，即对原始数据做清理、标注和变换等工作，使其满足模型训练所需的格式规范。为方便与其他研究学者的模型进行比较，本模型是在NYT公开数据集上进行实验和验证的，在模型训练之前，进行数据预处理的流程大致如图5-1所示。

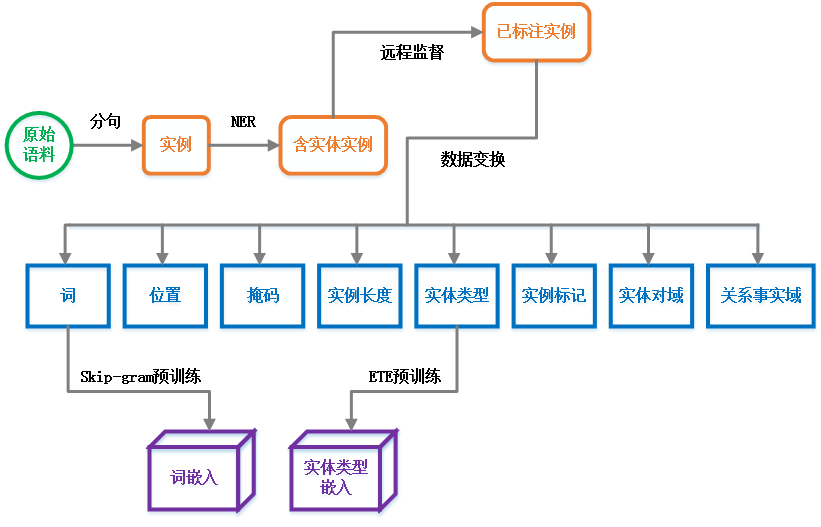


图5-1 数据预处理流程

从最原始的语料开始，首先经过分句处理会得到很多的实例，然后对每一个实例使用StanfordNLP工具来识别其中的命名实体，接下来，使用远程监督来自动标注每一个实例，最后，对已标注的实例进行数据的变换，以得到模型训练所需的如词、位置、掩码、实例长度等各种形式的信息。除此之外，我们模型训练时要用的词嵌入及实体类型嵌入是先分别经过Skip-gram预训练和ETE预训练后再喂给模型的。

## 实验设置

本课题开展的关系抽取研究是在NYT公开数据集上进行的，该数据集在关系抽取模型中被广泛使用。其中，NYT的训练集包含522611个句子、281270个实体对以及18252个关系事实，测试集包含172448 个句子、96678个实体对和1950个关系事实。学习器在经过预处理的语料上进行训练，再用训练好的模型去抽取测试集上每个句子中涉及的实体对之间的语义关系，抽取的过程中识别的每一个实体对(表示成：sub\_entity\_guid#obj\_entity\_guid)会得到相应关系的分数，哪一个分数高该实体对就在所在的实例中蕴含了对应关系，抽取的结果再进一步处理成三元组的形式已构成关系事实，这些关系事实可以存储到图形数据库或以RDF[[2]](#footnote-2)的形式存在，并利用D3js这样的可视化工具对存储的关系事实进行可视化以构建知识图谱相关应用。

从前一章所设计的关系抽取模型可以看出，本文提出的模型是基于深度神经网络结构的，这类模型在训练的时候涉及到大量的矩阵运算，而GPU在这方面具有天然的优势。因此，为了加快模型的训练速度，本模型是在实验室配备的用于AI研究的GPU服务器上进行训练的，训练前可以通过参数指定利用的GPU核数，本模型指定的是8个核，从而大大节省了训练的时间开销。

### 超参数设置

本模型训练时的超参数设置如表5-1所示，其中，最佳超参数是通过交叉验证与网格搜索的方式来指定的。

表5-1 关系抽取模型超参数设置

|  |  |  |
| --- | --- | --- |
| 参数名称 | 参数值 | 描述 |
| word\_dim | 50 | 词嵌入的维度 |
| pos\_dim | 5 | 位置嵌入的维度 |
| et\_dim | 12 | ETE的维度 |
| hidden\_size | 230 | 编码词嵌入与位置嵌入的隐藏层的大小 |
| et\_hidden\_size | 80 | 编码ETE的隐藏层的大小 |
| ib\_num | 3 | RPRN中恒等映射块的大小 |
| learning\_rate | 0.5 | 优化器的学习率 |
| drop\_out | 0.5 | 训练时丢弃某些神经元的比率 |

### 度量指标

对训练模型的泛化性能进行评估，不仅需要有效可行的实验估计方法，还需要衡量学习器泛化能力的评价标准，即性能度量(Performance Measure)。在对不同的模型进行比较时，使用不同的性能度量往往会导致不同的结果，这也揭示了学习器的好坏是相对的，模型的好坏不仅取决于数据和算法，还决定于任务需求。在预测任务中，给数据集,其中是示例的真实标记。要评价学习器的泛化性能，就要将真实标记与模型预测结果作比较。

1. 错误率与精度

分类任务常用的性能度量方法是错误率和精度，对数据集D，错误率以及精度的定义分别如公式5-1与5-2所示。

(5-1)

(5-2)

对于数据分布D和概率密度p(·),错误率与精度的一般形式与MSE相同，即将求和符号换成积分的形式。

1. 准确率、召回率和F1

错误率和精度并不能满足所有任务的需求。如在信息检索中，我们常常关注的是“在检索出来的信息中有多少是用户真正感兴趣的”以及“在用户所有感兴趣的信息中有多少能够被检索出来”，准确率(Precision)与召回率(Recall)正是应此类需求而被提出的。以二分类为例，根据模型预测结果与真实标记的组合可将样本分为真正例(TP)、假正例(FP)、假反例(FN)、真反例(TN)四种情况，其关系以混淆矩阵表示如下：

表5-2 分类结果混淆矩阵

|  |  |  |
| --- | --- | --- |
| 真实标记 | 预测结果 | |
| 正例 | 反例 |
| 正例 | TP | FN |
| 反例 | FP | TN |

根据以上混淆矩阵，准确率P与召回率R的定义如公式5-3所示：

(5-3)

通常来说，准确率与召回率是一对矛盾的度量，如何在准确率与召回率之间找到平衡就成为了一个难点。平衡点(BEP)表示“准确率=召回率”时的取值，但BEP太过简化，最常用的还是如公式5-4所示的F1[75]度量。

(5-4)

其中，samples\_num表示样本总数。在特定场景中，对准确率和召回率的偏好程度不同。为了表示对P和R的不同偏好，可以引入参数表示偏好程度。

## 评估结果

关系抽取模型的评估采用F1与AUC[76]值这两个指标来进行评估。每训练完一轮，就会对模型进行评估，如果评估的指标高于前一次保存的最好的，则最好的指标就会被更新并继续训练，如果连续20轮都没得到更好的结果，训练就终止以防止过拟合。模型训练的参数会保存起来，以便在测试时使用。表5-3表示本文进行的各关系抽取模型的评估结果，其中加粗的为本文提出的模型，灰色的为待比较的基模型。

表5-3 关系抽取模型评估结果

|  |  |  |
| --- | --- | --- |
| **模型名称** | **F1** | **AUC值** |
| **ete\_rprn\_att\_ad** | **0.4427** | **0.3965** |
| **ete\_rprn\_att** | **0.4409** | **0.3935** |
| **ete\_pcnn\_att\_ad** | **0.4443** | **0.3898** |
| **ete\_pcnn\_att** | **0.4416** | **0.3882** |
| **rprn\_att\_ad** | **0.4198** | **0.3649** |
| **rprn\_att** | **0.4135** | **0.3568** |
| pcnn\_att\_ad | 0.4209 | 0.3518 |
| bilstm\_att | 0.4020 | 0.3467 |
| pcnn\_att | 0.3991 | 0.3414 |
| pcnn\_one | 0.3984 | 0.3280 |

### 基准模型

为了和所提出的模型进行比较，本文采用了四个经典且最先进的模型进行比较，以下是本文采用的四个基准模型，图5-1为这些模型的P-R曲线。

* pcnn\_one(Zeng等人[37])将袋子中最可能的一个作为正例用于训练；
* pcnn\_att(Lin等人[42])应用选择性注意力机制到关系抽取建模中，该方法有效利用了袋子中大量有效的实例；
* bilstm\_att(Miwa等人[38])提出一种新的通过堆叠双向树结构的LSTM的端到端的神经网络模型来建模关系抽取；
* pcnn\_att\_ad(Wu等人[45]，Qin等人[46])在关系抽取建模时引入对抗训练，这进一步增强了模型的泛化能力。



图5-1 四个基模型的P-R曲线

### 本文提出的模型

本文提出的关系抽取模型的核心贡献主要包含两部分，分别为提出的ETE嵌入表示及设计的RPRN神经网络结构。为了更加充分地说明这两部分各自以及联合起来的有效性，本文基于这两个贡献设计了6个模型，分别与对应的基准模型进行比较，其说明如下所示：

* **ete\_pcnn\_att**在嵌入层应用提出的ETE嵌入表示，ETE将和词嵌入、位置嵌入一起作为模型的联合嵌入层；
* **ete\_pcnn\_att\_ad**除了在嵌入层应用ETE之外，还引入了对抗训练到模型中，以便与已有的对抗模型进行比较；
* **rprn\_att**利用本文设计的RPRN网络结构，RPRN用于编码词嵌入、位置嵌入的联合表示；
* **rprn\_att\_ad**在利用了RPRN网络结构的基础上，还引入了对抗训练到模型中，以便于和已有的对抗模型进行比较；
* **ete\_rprn\_att**联合了本文提出的ETE嵌入表示和RPRN网络结构，已验证这两部分联合起来的有效性；
* **ete\_rprn\_att\_ad**除了联合提出的ETE和RPRN之外，还将对抗训练引入到了模型中，以便于和已有的对抗模型进行比较。

为了验证ETE的有效性，图5-2给出了上述前两个模型分别与对应基模型的P-R曲线，为了验证RPRN的有效性，图5-3给出了中间两个模型分别与对应基模型的P-R曲线，为了验证ETE与RPRN联合起来的有效性，图5-4给出了最后两个模型分别与对应基模型的P-R曲线。



图5-2 采用**ETE**的模型与对应基模型之间的P-R曲线



图5-3 利用**RPRN**的模型与对应基模型之间的P-R曲线



图5-4 联合**ETE**与**RPRN**的模型与对应基模型之间的P-R曲线

表5-3给出了不同模型在F1及AUC两个指标上的一个直观的评估结果，而图5-1到图5-4进一步显示了基准模型之间以及本文提出的模型与对应基模型的P-R曲线的比较。实验结果显示我们的模型在各项评估指标上获得了很大的性能提升。更具体地说，各个模型之间的比较揭示了以下观测结果：

1. 图5-1显示各基准模型之间的性能差异并不是很大；
2. 在图5-2的两个子图中，分别将两个应用了ETE的模型与对应的基模型进行比较，从中可以看出，ETE极大地提升了关系抽取的性能；
3. 对于图5-3的两个子图，分别将利用了RPRN的两个模型与对应基模型进行比较，发现RPRN也能较好的提升模型的效果；
4. 图5-4中的两个子图均联合了ETE与RPRN，并分别与对应的基模型比较，并揭示了两个的联合是非常有效的，且**ete\_rprn\_att\_ad**模型获得了最好的关系抽取效果。

## 案例分析

从前一小节中定量的评估结果可以看出，本文提出的模型在各项评估指标上明显胜过现有最先进的模型。接下来，我们将进一步细致地分析不同模型关系抽取的一些具体案例。表5-4明确的列出了关系抽取结果的几个案例，并将我们提出的一个模型与对应的基准模型进行了比较。

表5-4 关系抽取案例

|  |  |  |
| --- | --- | --- |
| **Sentence** | **Extraction Result** | |
| **ete\_pcnn\_att** | **pcnn\_att** |
| There was good old **Sanath Jayasuriya**, the 37-year-old left-handed batsman for **Sri Lanka**, wal-loping the hard ball rising off the slippery grass,sending seven shots over the fences, a home run tous, a six to cricket fans. | /people/person  /nationality | NA |
| Donskoi, only 36 years old, unknown outside of **Arkhangelskand** perhaps better off for it, would stand little chance in a real campaign to be the leader of a country as sprawling, complex and deeply troubled as **Russia**. | /location/location  /contains | NA |
| In 1948, Rabbi Kret came to **New York City**, and with the help of Rabbi Yosef Eliyahu Henkin, of blessed memory, he was hired as the rabbi of the Old Broadway Synagogue in the West Harlem neighborhood of **Manhattanville**. | /location  /neighborhood  /neighborhood\_of | NA |
| But Justice Michael R. Ambrecht of State Supreme Court in Manhattan said that as a professional BASE-LRB- Bridge, Antenna, Span, Earth-RRB- jumper, Mr. Corliss, who has parachuted from the Eiffel Tower, the Golden Gate Bridge and the Petronas Towers in **Kuala Lumpur**, **Malaysia**, was experienced and careful enough to jump off a building without endangering his own life or anyone else ’s. | /location  /administrative  \_division/country | /people/person  /nationality |

表5-4中第一个句子的实体对“Sanath Jayasuriya”和“Sri Lanka”在所在句子中确实表达了“nationality”关系，但这一关系比较隐晦，没有直接关系触发词，仅仅根据上下文很难抽取出有效的关系出来，很容易理解，基模型(pcnn\_att)不能抽取出该关系，而我们提出的模型(ete\_pcnn\_att)由于利用了上述两个实体对应的实体类型信息，即“person”和“country”，便成功地抽出了正确的关系。此外，第二个和第三个句子可以与第一个句子进行类似地分析。至于第四个句子，没有利用实体类型信息的模型对所在句子的上下文感到困惑，反而识别出了“nationality”这一错误关系，而我们的模型再一次抽出了正确的关系。

无论是5.2小节给出的定量评估结果以及直观的P-R曲线图的比较，还是5.3小节中关于具体案例的定性分析，都清楚地表明了本文提出的ETE嵌入表示与RPRN网络结构的有效性。

## 本章小结

本章主要基于前两章设计与实现的模型与现有的最先进的模型做了对比，在各项指标上进行了定量评估，并分析了具体的案例。首先，介绍了下关系抽取模型中实验数据预处理的一般流程；接下来，简单说明了在NYT公开数据集上进行实验的一些设置，包括一些重要超参数的设置，以及模型的度量指标；然后给出了本文提出的模型与基准模型在各项指标上的定量评估结果，以及我们不同模型与对应基模型的P-R曲线图的比较；最后，列出了我们的一个模型与对应基模型进行关系抽取的具体结果，并定性分析了造成结果差异的原因。

# 总结与展望

## 论文总结

本课题主要针对当前关系抽取这一基础性研究所遇到的困境，如无法有效抽取长而复杂或无直接对应关系触发词的句子中的实体对关系，提出并设计了新的模型，从而使关系抽取的整体效果有了显著性的提升，这极大地提高了从文本数据中获取有效信息的效率。本文的主要工作内容概括如下：

1）通过仔细分析国内外学者在关系抽取方面的研究，发现了已有方法存在的问题，基于关系抽取的研究现状，确立了本课题面临的重点问题，以及根据发现的问题明确了要解决的难点是什么，并概述了要研究的内容和研究的意义。

2）深入调研并分析了本课题开展的关系抽取研究所依托的重要理论基础，以及模型实现相关的关键技术。其中包括用于自动标注大规模语料的远程监督，进行序列学习的循环神经网络，学习模型深层次信息的残差网络，用于缓解噪声问题的注意力机制，以及增强泛化能力的对抗训练；此外，还有模型实现用到的StanfordNLP工具，及用于神经网络建模的Tensorflow框架。

3）对于现有方法无法有效抽取长而复杂或无直接对应关系触发词的句子中涉及的实体对关系的问题，本文深入剖析了造成这一困境的原因，并提出ETE嵌入表示与RPRN网络结构来解决所遇到的困境。基于所提出的模型，本文对整体架构进行了设计，并详细论述了模型各重要模块以及设计的核心算法。

4）基于所设计和实现的关系抽取模型，在NYT公开数据集上进行了实验，为了验证本模型的有效性，本文从各种指标上对模型进行了评估，并将提出的模型与基准模型进行了比较，包括定量的评估指标、P-R曲线图的比较，以及本文提出的模型与对应基模型抽取的结果进行定性分析的比较。评估结果表明，本文提出的模型较已有模型有显著的性能提升。

综上所述，本文深入剖析了国内外学者在关系抽取方面的研究，针对目前研究所遇到的困境，提出并设计了新的关系抽取模型，以解决现有模型在长而复杂或无直接关系触发词的句子中抽取失败的问题。基于设计与实现的模型，本文在NYT公开数据集上进行了实验与评估，实验结果显示，本文提出的关系抽取模型效果提升显著，这极大地促进了文本数据的信息化建设。

## 未来工作展望

本文对关系抽取任务进行了深入研究，提出了基于ETE与RPRN的模型，解决了现有方法遇到的困境，并显著提高了关系抽取的整体效果，达到了本课题研究的目的。到此为止，本文的研究暂告一段落，但与本课题相关的工作仍存在不少值得探索、优化及扩展的地方，现对未来工作做以下简要概述：

1. 其他NLP任务的探索

在NLP中，有许多的任务都是和本课题开展的关系抽取任务密切相关的，如关系抽取依赖于NER，且本文设计的RPRN网络结构具有学习深层次信息的优良特性，因此未来的一个重要工作就是将RPRN应用于NER中，并和关系抽取任务进行联合起来建模，以同时提升这两项任务的效果。此外，基于NLP任务之间的相互关系，还计划将RPRN结构用在SRL任务中。

1. 多解析树、多模态关系抽取方法

自然语言形式非常灵活，一个句子在不同的上下文中可能具有不同的含义，有研究表明，如果考虑一个句子的多个可能的句法解析树，关系抽取的效果会更好，因此，在后续对关系抽取模型进行优化时会将这一点考虑进来。目前的关系抽取研究都是在单语言文本上进行的，实际上人类的只是蕴藏在不同的模态中，在未来的关系抽取建模中，我们会探索如何利用多语言语料、音频、视频等信息来进行关系抽取。

1. 关系抽取模型的应用

本课题进行的是关系抽取的基础性研究，该研究有着非常广泛的应用。在未来的工作中，我们会将本文提出的模型应用在科研文档中，并构建一个语义检索系统，在该系统中用户可以搜索实体，然后系统会将得到的与实体相关的关系事实数据进行可视化显示，显示的形式为关系图谱、知识卡片等，而且还会将提出的模型用于构建一个复杂的问答系统，从而为科研人员提供便利。

# 致 谢

时间过得真快，三年的研究生生活转眼即逝。在这忙碌而充实的学习生活中，有过失败，有过抱怨，感觉压力挺大的，但更多的是收获与回忆。研究生期间，我不仅学到了很多理论知识和专业技能，而且增长了不少为人处世、社会事态等方面的见识。这一路走来，有老师和同学的热情帮助，也有亲朋好友的鼓励与支持，我非常感动。在本论文即将落成之际，我衷心地感谢他们。

我第一个要感谢的就是我的导师王玉明副教授，从选题开始一直到整个论文完成之际，他给予了我极大的帮助，多次提出论文中的不足之处及改进意见，并耐心地给我讲解和分析原因，要不是他，我的论文不会进展得这么迅速。在此，感谢王老师对我的谆谆教导，感谢他对我的鼓励与包容，也感谢他给我们提供了这么优质的实验室环境及先进的实验设备，让我的论文实验进展得那么顺利。同时，我还要感谢涂来老师，感谢他多次与我交流发表在国外会议上的论文，并给了我许多建设性的意见和指导，从而为本论文的撰写提供了重要的基础。

感谢这三年以来一直陪伴我成长，在学习上给予我很大帮助，在生活上给我带来很大乐趣的实验室同学，他们是张恒杰、钟远维、汪贝贝、陈虹雨、唐福龙、祝文君、司玮辰、高士杰、陈洪、龙彦博、潘玉婷，陈聪、肖崇中、张志等。在这三年的研究生学习与生活中，我们常常在一起交流学习心得，谈论社会事态，正因为他们，我才能进步得这么快，生活也才会如此的丰富有趣。

最后，要感谢一直以来默默支持和鼓励我的两位朋友赵克请和徐明亮，以及对我关怀备至的亲人。感谢我的爸爸妈妈、哥哥姐姐及其他的亲人，谢谢你们在我成长的道路上给予我最纯粹的爱和最长久的陪伴，你们的爱与陪伴是我不断成长的动力，衷心地祝愿你们身体健康，快快乐乐，万事顺心！

# 参考文献

1. Grishman, Ralph. “Information Extraction.” IEEE Intelligent Systems 30 (2015): 8-15.
2. Cui, Lei et al. “Neural Open Information Extraction.” ACL (2018).
3. Hass, Richard W.. “Semantic search during divergent thinking.” Cognition 166 (2017): 344-357.
4. Hoffmann, Michael et al. “Benchmarking question answering systems.” Semantic Web 10 (2019): 293-304.
5. Shi, Baoxu and Tim Weninger. “Open-World Knowledge Graph Completion.” AAAI (2018).
6. Chinchor, Nancy and Elaine Marsh. “Appendix D: MUC-7 Information Extraction Task Definition (version 5.1).” MUC (1998).
7. Doddington, George R. et al. “The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation.” LREC (2004).
8. Gábor, Kata et al. “SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers.” SemEval@NAACL-HLT (2018).
9. Verhagen, Marc et al. “SemEval-2007 Task 15: TempEval Temporal Relation Identification.” SemEval@ACL (2007).
10. Hendrickx, Iris et al. “SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals.” SemEval@ACL (2010).
11. Miller, Scott et al. “A Novel Use of Statistical Parsing to Extract Information from Text.” ANLP (2000).
12. Kambhatla, Nanda. “Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Information Extraction.” ACL (2004).
13. Moncecchi, Guillermo et al. “A survey of kernel methods for relation extraction.” (2010).
14. Zhao, Shubin and Ralph Grishman. “Extracting Relations with Integrated Information Using Kernel Methods.” ACL (2005).
15. Bunescu, Razvan C. and Raymond J. Mooney. “Subsequence Kernels for Relation Extraction.” NIPS (2005).
16. Culotta, Aron et al. “Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text.” HLT-NAACL (2006).
17. Mintz, Mike et al. “Distant supervision for relation extraction without labeled data.” ACL/IJCNLP (2009).
18. Zhang, Congle et al. “Ontological Smoothing for Relation Extraction with Minimal Supervision.” AAAI (2012).
19. Krause, Sebastian et al. “Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web.” International Semantic Web Conference (2012).
20. Färber, Michael et al. “Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO.” Semantic Web 9 (2018): 77-129.
21. Takamatsu, Shingo et al. “Reducing Wrong Labels in Distant Supervision for Relation Extraction.” ACL (2012).
22. Yao, Limin et al. “Collective Cross-Document Relation Extraction Without Labelled Data.” EMNLP (2010).
23. Riedel, Sebastian et al. “Modeling Relations and Their Mentions without Labeled Text.” ECML/PKDD (2010).
24. Hoffmann, Raphael et al. “Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations.” ACL (2011).
25. Surdeanu, Mihai et al. “Multi-instance Multi-label Learning for Relation Extraction.” EMNLP-CoNLL (2012).
26. Min, Bonan et al. “Distant Supervision for Relation Extraction with an Incomplete Knowledge Base.” HLT-NAACL (2013).
27. Angeli, Gabor et al. “Combining Distant and Partial Supervision for Relation Extraction.” EMNLP (2014).
28. Goldwater, Sharon. “Part of Speech Tagging.” Encyclopedia of Machine Learning and Data Mining (2017).
29. Tan, Zhixing et al. “Deep Semantic Role Labeling With Self-Attention.” AAAI (2018).
30. Collobert, Ronan et al. “Natural Language Processing (almost) from Scratch.” Journal of Machine Learning Research 12 (2011): 2493-2537.
31. LeCun, Yann et al. “Deep Learning.” Nature 521 (2015): 436-444.
32. Socher, Richard et al. “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.” EMNLP (2013).
33. Zeng, Daojian et al. “Relation Classification via Convolutional Deep Neural Network.” COLING (2014).
34. Nguyen, Thien Huu and Ralph Grishman. “Relation Extraction: Perspective from Convolutional Neural Networks.” VS@HLT-NAACL (2015).
35. Xu, Kun et al. “Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling.” EMNLP (2015).
36. Santos, Cícero Nogueira dos et al. “Classifying Relations by Ranking with Convolutional Neural Networks.” ACL (2015).
37. Zeng, Daojian et al. “Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks.” EMNLP (2015).
38. Miwa, Makoto and Mohit Bansal. “End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures.” CoRR abs/1601.00770 (2016): n. pag.
39. Yao, Kaisheng et al. “Depth-Gated LSTM.” CoRR abs/1508.03790 (2015): n. pag.
40. Peng, Nanyun et al. “Cross-Sentence N-ary Relation Extraction with Graph LSTMs.” Transactions of the Association for Computational Linguistics 5 (2017): 101-115.
41. Song, Linfeng et al. “N-ary Relation Extraction using Graph-State LSTM.” EMNLP (2018).
42. Lin, Yankai et al. “Neural Relation Extraction with Selective Attention over Instances.” ACL (2016).
43. Wang, Linlin et al. “Relation Classification via Multi-Level Attention CNNs.” ACL (2016).
44. Du, Jinhua et al. “Multi-Level Structured Self-Attentions for Distantly Supervised Relation Extraction.” EMNLP (2018).
45. Wu, Yi et al. “Adversarial Training for Relation Extraction.” EMNLP (2017).
46. Qin, Pengda et al. “DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction.” ACL (2018).
47. Vo, Duc-Thuan and Ebrahim Bagheri. “Self-training on refined clause patterns for relation extraction.” Inf. Process. Manage. 54 (2018): 686-706.
48. Su, Yu et al. “Global Relation Embedding for Relation Extraction.” NAACL-HLT (2018).
49. Zhang, Qianqian et al. “A Review on Entity Relation Extraction.” 2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE) (2017): 178-183.
50. Kumar, Shantanu. “A Survey of Deep Learning Methods for Relation Extraction.” CoRR abs/1705.03645 (2017): n. pag.
51. Mikolov, Tomas et al. “Distributed Representations of Words and Phrases and their Compositionality.” NIPS (2013).
52. Leimeister, Matthias and Benjamin J. Wilson. “Skip-gram word embeddings in hyperbolic space.” CoRR abs/1809.01498 (2018): n. pag.
53. Shi, Weiwei and Sheng Gao. “Relation Extraction via Position-Enhanced Convolutional Neural Network.” 2017 International Conference on Intelligent Environments (IE) (2017): 142-148.
54. Wang, Jianyong et al. “Recurrent Neural Networks With Auxiliary Memory Units.” IEEE Transactions on Neural Networks and Learning Systems 29 (2018): 1652-1661.
55. He, Kaiming et al. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 770-778.
56. Röder, Michael et al. “GERBIL - Benchmarking Named Entity Recognition and Linking consistently.” Semantic Web 9 (2018): 605-625.
57. “Software - The Stanford Natural Language Processing Group.” Stanford.Edu, 1 Jan. 2014, nlp.stanford.edu/software/.
58. “TensorFlow.” TensorFlow, 1 Jan. 2019, www.tensorflow.org/.
59. Ding, Boyang et al. “Leveraging Text and Knowledge Bases for Triple Scoring: An Ensemble Approach - The BOKCHOY Triple Scorer at WSDM Cup 2017.” CoRR abs/1712.08356 (2017): n. pag.
60. Wikipedia Contributors. “Knowledge Base.” Wikipedia, Wikimedia Foundation, 17 Feb. 2019, en.wikipedia.org/wiki/Knowledge\_base.
61. Zhang, Qi et al. “Adaptive Co-attention Network for Named Entity Recognition in Tweets.” AAAI (2018).
62. Sil, Avirup et al. “Neural Cross-Lingual Entity Linking.” AAAI (2018).
63. Goodfellow, Ian et al. “Deep Learning.” MITPress, 2016, www.deeplearningbook.org.
64. Ranganathan, Varun and S. Natarajan. “A New Backpropagation Algorithm without Gradient Descent.” CoRR abs/1802.00027 (2018): n. pag.
65. Gruslys, Audrunas et al. “Memory-Efficient Backpropagation Through Time.” NIPS (2016).
66. Pascanu, Razvan et al. “On the difficulty of training recurrent neural networks.” ICML (2013).
67. Dey, Rahul and Fathi M. Salem. “Gate-variants of Gated Recurrent Unit (GRU) neural networks.” 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS) (2017): 1597-1600.
68. Kanai, Sekitoshi et al. “Preventing Gradient Explosions in Gated Recurrent Units.” NIPS (2017).
69. Russakovsky, Olga et al. “ImageNet Large Scale Visual Recognition Challenge.” International Journal of Computer Vision 115 (2015): 211-252.
70. Krizhevsky, Alex et al. “ImageNet Classification with Deep Convolutional Neural Networks.” Commun. ACM 60 (2012): 84-90.
71. Agarap, Abien Fred. “Deep Learning using Rectified Linear Units (ReLU).” CoRR abs/1803.08375 (2018): n. pag.
72. Abadi, Martín et al. “TensorFlow: A System for Large-Scale Machine Learning.” OSDI (2016).
73. Alatawi, Abdulrahman et al. “Bayesian Unigram-Based Inference for Expanding Abbreviations in Source Code.” 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI) (2017): 543-550.
74. Rohekar, Raanan Y. Yehezkel et al. “Bayesian Structure Learning by Recursive Bootstrap.” NIPS (2018).
75. Lipton, Zachary Chase et al. “Optimal Thresholding of Classifiers to Maximize F1 Measure.” Machine learning and knowledge discovery in databases : European Conference, ECML PKDD ... : proceedings. ECML PKDD 8725 (2014): 225-239 .
76. Huo, Jing et al. “Cross-Modal Metric Learning for AUC Optimization.” IEEE Transactions on Neural Networks and Learning Systems 29 (2018): 4844-4856.

1. <http://image-net.org/challenges/LSVRC/2015/> and <http://cocodataset.org/#detections-challenge2015> [↑](#footnote-ref-1)
2. <https://www.w3.org/RDF/> [↑](#footnote-ref-2)