2019届研究生博士学位论文

分 类 号: _____ 学校代码: 10269

密 级: _____ 学 号: 52164500005

華東師範大學

基于深度学习的联合实体关系抽取

院 系: 计算机科学与技术学院

专 业: 计算机应用技术

研 究 方 向: 自然语言处理

指 导 老 师: 孙仕亮 教授

吴苑斌 副教授

论 文 作 者: 孙长志

2019 年 09 月

Dissertation for Doctorate Degree in 2019 University Code: 10269

Student ID: 52164500005

East China Normal University

Joint Entity Relation Extraction with Deep Learning

Department: Computer Science and Technology

Major: Computer Application Technology

Research Direction: Natural Language Processing

Supervisor: Professor Shiliang Sun

Associate Professor Yuanbin Wu

Candidate: Changzhi Sun

September, 2019

华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《基于深度学习的联合实体关系抽取》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名：_____ **日　期**：　　年　　月　　日

华东师范大学学位论文著作权使用声明

《基于深度学习的联合实体关系抽取》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的著作权归本人所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和学校指定的相关机构送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

(　　) 1. 经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文 *，于　　年　　月　　日解密，解密后适用上述授权。

(　　) 2. 不保密，适用上述授权。

导师签名：_____ **本人签名**：_____

年　　月　　日

* “涉密”学位论文应是已经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权）。

孙长志 博士学位论文答辩委员会成员名单

姓 名	职 称	单 位	备 注
周日贵	教授	上海海事大学	主席
赵 海	教授	上海交通大学	
王 喆	教授	华东理工大学	
何晓丰	研究员	华东师范大学	
王晓玲	教授	华东师范大学	

摘　要

随着互联网的兴起，每天都有不同形式的大量的文本数据产生：新闻、研究文献、博客、论坛文字以及社交媒体评论等。很多重要有用的信息隐藏在其中，如何从这些自由文本中自动抽取所需要的信息是一个关键并且重要的一步。信息抽取任务就是为此目标而诞生。本文主要研究信息抽取子任务之一的实体关系抽取任务。该任务旨在识别文本中出现的实体，并判断出实体之间存在的关系。

传统的有监督实体关系抽取通常采用基于流水线的方法，即实体模型和关系模型分开训练。在测试阶段，先用实体模型识别出实体，然后关系模型找出这些实体之间的关系。这种流水线的方法存在着错误传播的缺点，前一个任务的错误会累积到后一个任务。为了缓解这一问题，研究人员提出了联合模型。联合模型将两个子模型统一建模，可以进一步利用两个任务之间的潜在信息，以缓解错误传播的缺点。联合模型的难点是如何加强实体模型和关系模型之间的交互，比如实体模型和关系模型的输出之间存在着一定的约束，在建模的时候考虑到此类约束将有助于联合模型的性能。

另一方面，为了解决实体关系抽取数据集难以获得的问题，远程监督的方法也被提出来。其主要思想是利用知识库和大规模文本数据对齐，自动构建大规模的训练集。然而，远程监督方法的缺点是自动构建的训练集中存在着很多的噪音数据，这些噪音数据的存在对远程监督实体关系抽取有着很大的负面影响。此外，在有些应用场景中可能没有现成的知识库可以用来进行远程监督，如何解决类似的数据噪音和数据缺失问题也是一大挑战。

根据实体关系抽取方法的研究现状，本文从**数据**和**联合模型**两个角度探索了几种实体关系抽取联合模型，并且探究了所提出模型的优势和不足。具体来说，本文的主要贡献有

1. 为了缓解远程监督中的噪音样本问题，本文提出利用少量高质量异构的人工标注数据集帮助远程监督实体关系抽取任务。本文设计了一个基于多任务学习的融合框架，并且在融合过程中考虑到子模型之间的一致性约束，从而实现知识的迁移。本文提出的系统在标准远程监督数据集能够显著的提高联合抽取的性能（数据角度）。

2. 为了解决某些领域没有现成知识库无法进行远程监督的问题，本文提出利用语言学规则进行远程监督。首先应用领域无关的语言学规则自动构建训练集，然后使用分类器在得到的训练集上进行训练，最后利用分类器进一步抽取语言学规则无法覆盖的新的实体关系。本文提出的算法很快并且适用于大规模数据。在 Amazon 在

线评论数据集上的实验表明了本文提出的算法明显优于多个基准模型（数据角度）。

3. 为了加强实体模型和关系模型之间的交互，本文提出基于风险最小化训练方法的联合实体关系抽取模型，通过优化全局的损失函数以达到加强实体模型和关系模型之间联系的目的。在 ACE05 数据集上的实验证明了提出模型的有效性（联合模型角度）。

4. 为了同时考虑到实体类型和关系类型的信息，本文提出一个基于图卷积网络的联合模型用于实体关系抽取。我们构造了实体-关系二分图，并在图上运行图卷积网络，从而捕获多个实体和多个关系之间的信息。在 ACE05 数据集上的实验证明了提出模型的有效性（联合模型角度）。

关键词：联合实体关系抽取，信息抽取，远程监督，神经网络

ABSTRACT

With the fast development of Internet, a large amount of free texts are produced in different forms everyday. Information extraction, which is about how to automatically extract knowledge from these free texts, is a key and important task in natural language processing. The information extraction task is to tackle this problem. In this work, we focus on the entity relation extraction task which is a subtask in information extraction. It aims to identify entities that appear in the text, and semantic relations among entities.

One common framework for supervised entity relation extraction is a pipeline-based. Specifically, entity models and relation models are trained separately. An entity model is first used to identify entities in inputs, and then a relation model finds relations between these entities. The pipeline method suffers the error propagation problem. Errors from the previous task will accumulate to the next task. To alleviate this problem, many researchers study joint extraction models: extracting entity and relation in a unified model. The main difficulty of the joint extraction is how to handle the interaction between the entity model and the relation model.

On the other hand, since large-scale annotation data is often difficult to obtain, distant supervision methods have been applied to the entity relation extraction. The main idea is try to align knowledge bases and large-scale text data which can automatically obtain a large number of training data. However, there are a lot of noise in the obtained dataset. It limits the performance of the distantly supervised entity relation extraction. Furthermore, in some applications, knowledge bases may be unavailable for distant supervision, which enables the distant supervision more challenging.

This work explores several entity relation extraction methods from the perspective of **data** and **joint model**, and investigates the advantages and limitations of proposed methods. Specifically, the main contributions are as follows:

1. In order to alleviate the problem of noisy samples in distant supervision, we propose to use a small amount of high quality heterogeneous manual labeled dataset to help distantly supervised entity relation extraction task. We design an adaptation framework based on multi-task learning, and consider some consistency constraints in the adaptation process, so as to achieve knowledge transfer. Experiments on distantly supervised dataset demonstrate the effectiveness of the proposed framework (perspective of data).

2. In order to tackle the problem that there is no knowledge base for distant supervision in some domains, we propose to use linguistic rules to help distant supervision. Firstly, a

training set is constructed automatically by using domain-independent linguistic rules, and then a classifier is built based on the training data. Comparing with only rule-based model, the classifier can extract relations that linguistic rules cannot cover. The proposed algorithm is fast and scalable on large-scale dataset. Experiments on Amazon online review dataset demonstrate that the proposed model is able to achieve promising performances (perspective of data).

3. In order to handle the interaction between entity model and relation model, we propose a joint extraction model based on minimum risk training, which can strengthen the connection between entity model and relation model by optimizing the global loss function. Experiments on ACE05 dataset demonstrate the effectiveness of the proposed joint model (perspective of joint model).

4. In order to handle the joint type inference on entities and relations, we propose a joint model based on graph convolutional network for entity relation extraction. An entity-relation bipartite graph is constructed and a graph convolution network is run on the graph to capture information between multiple entities and relations. Experiments on ACE05 dataset demonstrate the effectiveness of the proposed model (perspective of joint model).

Keywords: Joint Entity Relation Extraction, Information Extraction, Distant Supervision, Neural Network

目 录

插图目录

表格目录

第一章　绪论

随着大数据时代的到来，我们的身边每天都涌现出大量的文本数据，比如新闻报道、博客、研究文献以及社交媒体评论等。在这样信息爆炸的现代社会，如何快速有效地利用这些海量文本数据从而为用户提供更好的服务已经成为亟需解决的挑战。信息抽取（Information Extraction，IE）就是为了实现此目标而诞生。而实体关系抽取（Entity and Relation Extraction）是信息抽取的关键任务之一，近些年来受到学术界和工业界的广泛关注。它可以为自动问答（Question Answering，QA）、信息检索（Information Retrieval，IR）、知识库填充（Knowledge Base Population，KBP）、知识推理（Knowledge Reasoning，KR）等下游任务提供支持。

实体关系抽取包含两个子任务：实体识别（Entity Recognition，ER）和关系抽取（Relation Extraction，RE）。并且这两个子任务通常存在先后顺序关系，即先识别出实体，然后抽取实体之间存在的关系。这样的任务在自然语言处理领域出现很多，比如词性标注（POS Tagging）和成分句法分析（Consitituent Parsing）。如何更好地处理这样的任务也逐渐成为近些年研究的热点。

本文主要研究联合实体关系抽取（Joint Entity Relation Extraction）任务，将实体识别和关系抽取联合建模。相比传统的流水线方法（Pipeline）建模，联合模型通常可以取得更好的性能。同时，此任务的人工标注数据极为稀少，进一步加大该任务的难度。因此，针对以上两个挑战，本文从“数据”和“联合模型”两个层面提供相应解决方法。

1.1　研究背景

为了使得机器更近一步理解人类语言文字，需要将自由文本转化为结构化的数据。这也是信息抽取技术的目标所在。图1.1 的例子解释了信息抽取如何整理非结构化数据 [1]。信息抽取作为自然语言处理领域的重要分支之一，旨在识别给定文本中的实体、实体之间的关系、事件以及它们的内在联系。同时，信息抽取将文本中的信息转换成一种更容易访问的形式，供其它自然语言处理下游任务使用。信息抽取涉及多个任务，比如实体识别（Entity Recognition，ER），实体链接（Entity Linking，EL），指代消解（Coreference Resolution，CR），关系抽取（Relation Extraction，RE），事件抽取（Event Extraction，EE）。针对不同的任务，需要设计

[1]图1.1改编于论文 [182] 图 6。

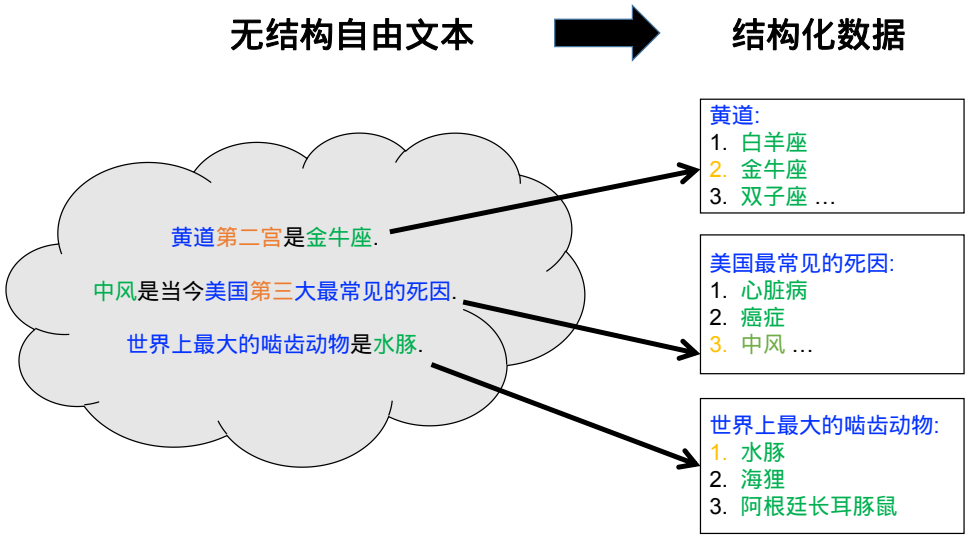

图 1.1 信息抽取样例

合适的模型去解决。

本文主要研究信息抽取中的实体关系抽取任务（Entity Relation Extraction），即给定一段自由文本（通常是一个句子），识别出句子中的实体并且抽取出命名实体之间的关系。所谓实体就是指文本中出现的时间（TIME）、地点（LOC）、组织（ORG）、人物（PER）、武器（WEAPON）等，也可以是抽象的概念。所谓关系就是指实体之间存在某种语义联系，比如：雇佣关系（EMP-ORG）和地理位置关系（PHYS）等。给定一个句子“任正非是华为董事长”，其中“任正非”和“华为”分别是人物（PER）和组织（ORG）实体，同时他们又具有一种雇佣关系（EMP-ORG），即“任正非”受雇于“华为”。通常来说实体的类型和关系的类型都是人为预先定义的。通过以上介绍，可以看出实体关系抽取就是将一段自由文本转化为结构化的数据，可以保存在数据库中以便于其他任务使用。一旦这种结构化数据被抽取，可以有很多用途。例如知识库的填充，向现有的知识库中添加新的关系三元组。此外，它还可以为自动问答提供支持。

实体关系抽取的历史可以追溯到 20 世纪 80 年代，美国国防高级研究计划局（Defense Advanced Research Projects Agency，DARPA）举办了 7 届消息理解会议（Message Understanding Conference，MUC），并且于 1988 年最后一次 MUC-7 会议上定义了模板关系（Template Relation，TR）。该会议极大推进了实体关系抽取的研究。继 MUC 会议之后，美国国家标准技术研究院（National Institue of Standard and Technology，NIST）举办了自动内容抽取（Automatic Content Extraction，ACE）的测评会议。ACE 主要针对新闻语料中的实体关系抽取，并且发布了一系列的人工标注语料。SemEval 会议在 2007 年第四 4 个任务 [53] 和 2010 年第 8 个任务 [62] 都是针对关系抽取，吸引了全球很多研究者的关注。此外，由于标注数据的稀少，

弱监督 [3, 8, 20, 50] 和远程监督 [127] 的方法被提出。弱监督学习通常面向开放领域的信息抽取（Open Information Extraction，OpenIE）。远程监督方法的兴起，出现了面向知识库构建过程的实体关系抽取，比如文本分析知识库填充会议（Text Analysis Conference Knowledge Base Population，TAC KBP）。最近清华大学自然语言处理实验室发布了大规模标注的关系抽取数据集 FewRel [58] 以及大规模文档级别的关系抽取数据集 DocRED [212]。

1.2 本文贡献

本文主要研究实体关系抽取任务，从“数据”和“联合模型”两个角度进行探索以便于更好地提升此任务的性能。

- **数据**：当前我们可以从互联网等其他源头获得大量的自由文本数据，人们应该如何才能更好地理解这些数据呢？一个自然地想法就是通过标注语义信息将无结构的文本转换为结构话的数据。针对这个问题，许多方法被研究者提出。然而这些方法存在一些缺陷，其中包括：

  1. 过去为了自动标注关系数据，一些有监督和无监督的方法被提出。一方面，有监督的方法可以取得不错的性能，但是依赖于人工标注数据，而通常很难得到大量的人工标注，并且训练得到的模型只适用特定的领域。另一方面，无监督方法的主要思想就是人工编写规则（包括词法规则和句法规则）自动抽取数据。通常来说，这些规则比较简单有效，并且适用于大量数据。但是，在实际应用中这些规则的健壮性存在很大问题，有许多可以改进之处。因此，我们需要一个算法可以更好地利用规则以减少人工标注。
  2. 另外一种可以自动生成训练数据的方法是远程监督 [127]。简单来说就是将自由文本和知识库里的三元组进行对齐。这种方法主要的缺点是得到的数据中混有大量噪音，对齐的关系在某个上下文中可能并不存在。比如知识库中（“奥巴马”，“美国”）具有关系“出生于”，但是并不是所有同时出现“奥巴马”和“美国”实体对的句子都表达了“出生于”的关系。之前的方法都致于减少噪音的影响或者识别出噪音数据，但是由于是否是噪音没有具体的标注使得这个问题难以解决。另一方面，存在少量的人工标注语料，然而这些语料标注的实体关系类型和远程监督数据实体关系类型可能不一致。之前的方法没有考虑到用这些异构数据减少远程监督中的噪音问题。因此，我们需要一个框架可以实现融合异质数据，从而提升远程监督方法的性能。

针对以上两个问题，本文的主要贡献在于：

1. 在给定大规模无标记语料时，本文提出了一个有效的基于语言学规则匹配和神经网络分类器的远程监督学习框架。规则可以自动得到训练数据，利用得到的数据可以训练一个神经网络分类器，它可以捕获各种词法和句法的特征。最终的算法适用于大规模数据并且取得不错的效果。
2. 为了缓解远程监督学习中的噪音问题，本文提出一个可以融合异质数据的框架，可以用少量人工标注语料提升远程监督学习的性能。并且设计的框架具有可解释性，使其健壮性和一致性得到保证。实验结果也同时验证了该框架的有效性。

- **联合模型**：实体关系抽取包含两个子任务，即实体识别和关系抽取。传统的解决方法将两个子任务独立对待，分别训练两个独立的模型。在实际应用中用流水线方式进行预测，即首先用实体模型检测出所有实体，然后对得到的实体对进行关系分类。此种方法比较灵活，比如实体模型和关系模型可以用不同的数据进行训练。然而也存在比较明显的缺点，就是实体模型和关系模型存在错误传播。为了缓解这一问题，近些年来很多研究者都聚焦在联合模型上，并且已经取得了不错的进展。但是之前的联合模型也存在一些不足，其中包括：

  1. 实体模型和关系模型简单地通过共享参数（共享输入特征或者内部隐层状态）实现联合，此种方法对子模型没有限制，但是由于使用独立的解码算法，导致实体模型和关系模型之间交互不强。为了加强实体模型和关系模型的交互，一些复杂的解码算法被提出来，比如整数线性规划等。在这种情况下，就需要对子模型特征的丰富性以及联合解码的精确性之间做权衡。换句话说，一方面如果设计精确的联合解码算法，往往需要对特征进行限制，例如用条件随机场建模，使用维特比解码算法可以得到全局最优解，但是往往需要限制特征的阶数。另一方面如果使用近似解码算法，比如集束搜索，在特征方面可以抽取任意阶的特征，但是解码得到的结果是不精确的。针对联合实体关系抽取任务，在解码的精确性和特征的丰富性之间做权衡通常很重要并且很难。所以，我们需要一个算法可以在不影响子模型特征丰富性的条件下加强子模型之间的交互。
  2. 之前的方法在判断实体类型时并没有直接用到关系的信息，然而这种信息是很重要的。比如如果知道“奥巴马”和“美国”具有“出生于”关系，这将有助于对“奥巴马”和“美国”实体类型的判断。同时，在一个句子中多种关系之间也是存在联系的，这种联系如果能够捕获，对整个联合

模型是有益的。所以，我们需要一个方法可以同时考虑一个句子中所有实体、实体与关系、所有关系之间的交互。

针对以上两个问题，本文的主要贡献在于：

1. 和之前工作不同的地方，本文提出了一个基于风险最小化训练的新颖的轻量的联合学习方法。具体来说，本文提出的算法优化了一个全局损失数据函数，同时很灵活并且有效地探索实体模型和关系模型之间的交互。此外，本文实现了一个简单并且强大的神经网络可以承载风险最小化训练。实验结果也证明了本文提出算法的有效性。
2. 针对联合实体关系抽取，本文提出了一个新的方法。首先识别出实体的边界，然后执行实体和关系的联合类型推理。为了解决联合类型推理任务，本文提出一个新颖的运行在实体-关系二分图上的图卷积神经网络。通过引入一个二元关系分类任务，能够以一种更有效和更有解释性的方法利用实体—关系二分图的结构。实验结果也说明了提出方法的有效性。

1.3 各章组织

本文剩余部分章节组织如下：

第二章介绍实体关系抽取技术的组成、特点、典型方法以及相关数据集。同时也介绍必要的深度学习相关知识。

第三章提出一个可以融合异构数据神经网络框架来处理实体关系抽取任务。通过多任务学习的方式训练整个网络参数。

第四章提出基于语言学规则的远程监督关系抽取。精心设计了一组语言学规则并且设计一个神经网络分类器。

第五章提出基于风险最小化训练方法的联合实体关系抽取。设计一个简单并且强大的神经网络，在最大似然方法预训练之后，再结合风险最小化训练方法进行微调。

第六章提出基于图卷积网络的联合实体关系抽取。定义了实体-关系二分图，并在其上运行图卷积网络，最终执行联合类型推理。

第七章总结全文，结合现状给出将来可能的研究方向。

第二章　实体关系抽取技术基础

为了使得计算机更好的理解和处理自然语言以及文本，通常会将自由的文本数据转换为有结构的数据存储。信息抽取技术即为此目标而生。本文研究联合实体关系抽取，是信息抽取技术中重要的一个方向，包含实体识别和关系抽取两个子任务。为了很好的理解本文提出的各种模型，本章将叙述实体关系抽取技术所需要的各种前提知识。首先会分别介绍实体识别（2.1 节）和关系抽取（2.2 节），然后再总结联合实体关系抽取的技术发展概况（2.3 节），最后简单叙述实体识别技术以及本文所需要的深度学习基础知识（2.4 节）。

2.1　实体识别

实体识别旨在识别出文本中诸如人名、地名、组织、时间等实体，是自然语言处理领域一项关键的任务[1]。很多任务之前都需要通过实体识别系统检测出实体，然后才能进行相应的处理，比如自动问答、信息检索、知识库填充等。针对此任务，各种各样的解决方法被研究者们提出来。特别是近些年来，深度学习的崛起使得实体识别技术得到长足的发展。在本节，会首先介绍任务的描述（2.1.1 节），再对已有的解决方法进行分类总结（2.1.2 节），最后简单介绍常用的数据集（2.1.3 节）。

2.1.1　任务描述

给定一段自然语言文本，实体识别任务是抽取特定的文本片段，对于每个文本片段，可以是单个词或者多个词，并且具有相应的类别，比如人名、地名、组织等。图 2.1 是一个实体识别的样例。其中“华为”是组织名称（ORG），“任正非”是人名（PER）。

华为 董事长是 任正非

组织 (ORG)　　人名 (PER)

图 2.1 实体识别样例

[1]本文所指实体皆为命名实体。

评价指标：该任务使用的通常评价指标是 F_β [2]，其中 $\beta = 1$ [161]。

$$F_\beta = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \tag{2.1}$$

精确率（$precision$）是指系统输出正确实体的百分比。召回率（$recall$）是指系统在数据集中找到正确实体的百分比。仅当预测实体与标注的实体完全匹配时（实体的边界与类型均要匹配），实体才算正确。

2.1.2 相关方法

本节主要叙述实体识别的相关方法，主要包括基于传统方法的实体识别和基于深度学习方法的实体识别。

基于传统方法的实体识别

传统的实体识别涉及到两个大类，即无监督的实体识别和有监督的实体识别。以下将具体介绍这两大类的传统实体识别方法。

无监督学习 无监督的实体识别主要是基于规则的，规则的设计需要领域相关的知识库、词典，甚至需要专家的精心设计。比较有名的基于规则的实体识别系统有 LaSIE-II [69]，NetOwl [90]，Facile [17]，SAR [4]，FASTUS [5] 和 LTG [125]。这些系统主要基于手动设计的语义和句法规则来识别实体，比如对句子进行词性标注，将满足某些限制的名词短语视为实体。当词典资源非常丰富时，通常可以取得不错的性能。文献 [38] 仅仅使用少量的种子标注数据和 7 个特征，包括拼写（比如大小写），实体上下文，实体本身等，进行实体识别。KnowItAll [50] 系统是无监督的，利用领域无关的规则模板，可以自动从网页上抽取大量的实体（和关系）。文献 [134] 提出一个无监督的系统，用于地名索引和实体消歧。该方法使用简单有效的启发式规则来实现实体识别和实体消歧。文献 [223] 将无监督实体识别方法应用在生物医学领域，主要通过术语，语料库统计（比如逆文档频率和上下文向量）和浅层句法知识（比如名词短语）来实现无监督抽取实体。文献 [34] 根据数据库领域的成功经验，设计了基于规则的信息抽取系统。无监督学习的实体识别的优势是不需要任何标注数据，可以借助于词典和人工设计规则得到大量的实体。然而，由于规则是特定领域的以及词典的不完整性，这些系统往往具有较高的精确率和较低的召回率，并且很难将此系统应用在其他的领域。

[2]在之后的章节中，F_1, F1, F 均表示相同含义。

有监督学习 当具有一定规模的标注数据时，可以用有监督的机器学习算法训练一个实体识别模型。通常来说，实体识别模型是一个序列标注问题。序列标注的输入是一个序列，输出也是一个等长的序列，适合于处理自然语言文本。通常可以采用 BIO 或者 BIOUL 标签计划来表示序列标注的输出。以条件随机场为例，这个是经典的序列标注模型，我们可以抽取每个位置 t 上的特征以及相邻输出标签之间的特征，假设有 K 种，表示为 $\{f_k(y_t, y_{t-1}, \mathbf{x}_t)\}_{k=1}^{K}$。根据图 2.2所示，模型的条件概率为

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\}$$

其中 $Z(\mathbf{x})$ 是归一化函数。

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \prod_{t=1}^{T} \exp\left\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right\}$$

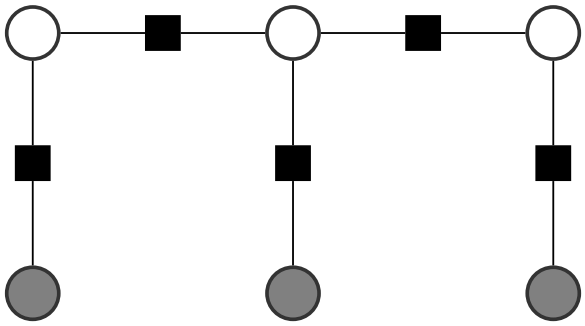

图 2.2 条件随机场的图模型表示

为了使用监督式的机器学习算法，需要对每个样本进行设计特征。良好的特征能够有效的预测出是不是实体。一旦训练完成，得到的模型一定程度上可以从未见过的文本中来识别相似的实体。在监督式的实体识别系统中，特征工程非常之关键。特征向量表示是对词的抽象表示，其中一个单词可以用一个或者多个布尔值，或者数值表示 [38, 99, 133]。词级别的特征（比如大小写，形态以及词性标注）[105, 165, 229]，查表式的特征（比如维基地名词典等）[64, 83, 124, 181]，文档和整个语料级别的特征（比如局部语法和共现信息）[73, 89, 154, 232]，这些特征也被广泛应用在各种监督式的实体识别系统中。具体地，简单的特征函数可以设计为

$$f_k(y_t, y_{t-1}, \mathbf{x}) = \begin{cases} 1 & \mathbf{x}_t.\text{pos} = \text{NN 并且 } y_{t-1} = \texttt{B} \text{ 并且 } y_t = \texttt{I} \\ 0 & \text{否则} \end{cases}$$

关于更多的特征设计可以参考文献 [26, 133, 168]。

基于以上的这些特征，许多机器学习算法被用于监督式的实体识别系统中，包括隐马尔可夫模型（Hidden Markov Models，HMM）[47]，决策树（Decision Trees，DT）[152]，最大熵模型（Maximum Entropy Models，MEM）[79]，支持向量机（Support Vector Machines，SVM）[61] 和条件随机场（Conditional Random Fields，CRF）[93] 等。已有的模型汇总见表 2.1，为了简单起见，特征部分列出比较新颖的特征。

表 2.1 特征工程的监督式实体识别系统汇总

文献	部分特征	机器学习算法
[14, 15]	-	HMM
[229]	互信息	HMM
[178]	-	DT（C4.5），AdaBoost
[11, 19, 42]	-	MEM
[33]	全局特征	MEM
[70, 104, 123]	拼写，标点符号	SVM
[122]	特征归纳	CRF
[89, 111, 114, 158–160, 165]	-	CRF

基于有监督学习的实体识别相比于无监督学习，通常可以取得更好的性能。缺点是需要一定数据的标签数据，并且需要人工设计许多复杂的特征，即特征工程。为了减轻人工干预，基于深度学习的方法在实体识别中逐渐流行起来。

基于深度学习方法的实体识别

当前深度学习主要聚焦于有监督学习，深度神经网络替代传统的特征工程，可以自动抽取有效的特征，避免了大量繁琐的特征工程，有效减少了很多的人工干预。近些年来，基于深度学习的各种实体识别模型如雨后春笋一样破土而出。在这里，首先定义常用的几种编码器和解码器。

1. 序列到序列编码器：将一个向量序列作为输入并返回相同长度的向量序列，输出后的每个向量融合了整个序列的信息。例如在执行序列标注任务时，使用此种编码器之后可以抽取到更加强大有效的特征。常见的几种序列到序列编码器如图 2.3 中的虚线部分。

2. 序列到向量编码器：将一个向量序列作为输入返回单个向量。这里通常只需要在序列到序列编码器之后加一个池化操作即可。得到的向量可以看成整个序列的表示。例如可以用这个向量直接做分类任务。常见的几种序列到向量编码器如图 2.3。

3. 序列到序列解码器：将一个向量序列作为输入返回序列的标签。例如在序列

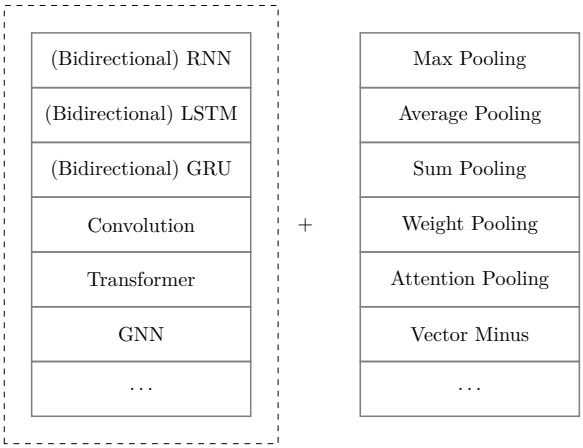

图 2.3 序列到向量编码器

标注模型中，当我们使用序列到序列编码器，得到每个词的表示后，为了预测对应的标签，需要将向量转化为离散的标签信号，这就是序列到序列解码器的主要作用。常见的几种序列到序列解码器如图 2.4。

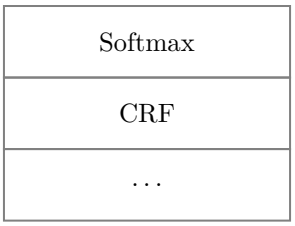

图 2.4 序列到序列解码器

4. 分类解码器：将一个向量作为输入返回对应的标签。常见的分类解码器即 Softmax。

与传统方法的实体识别一样，通常用序列标注框架实现实体识别。根据当前深度学习的主流实体识别模型，本文将序列标注模型的基本架构总结如图 2.5，自底向上依次包括词编码器，序列编码器以及序列解码器。

- 词编码器：将离散的单词转换为向量表示。通常可以包含三个级别的信息，词级别（比如词向量），字符级别（一个单词包含多个字符，可以对字符序列进行编码，从而得到字符级别单词的表示）以及其他特征（传统方法中的特征也可以融入到该单词的表示中）。
  - 词级别：将离散的单词信息转化为向量，比如传统的 one-hot 表示，以及现在比较流行的词向量。词向量相对于 one-hot 表示，具有更低的维度，有效避免了“维度灾难”。常见的词向量可以随机初始化，或者使用预训练好的词向量，例如 word2vec，glove 等。

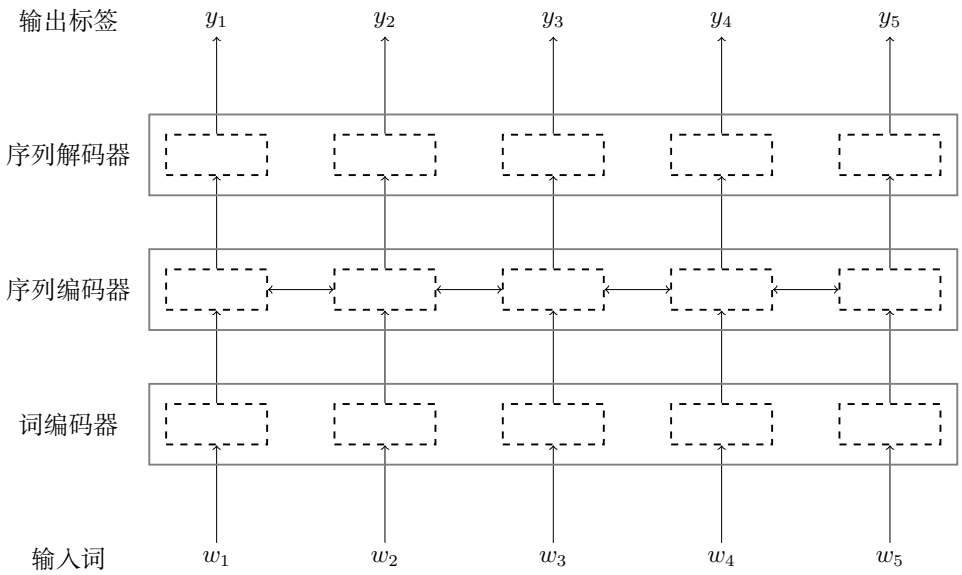

图 2.5 序列标注模型基本架构

  - 字符级别：一个单词通常由字符序列构成，所以一个单词的表示可以将字符序列的信息考虑进来。比如引入字符向量，在字符序列上使用任意的序列到向量编码器，即可得到字符级别的单词表示。
  - 其他特征：在传统的特征工程方法中，每个词的特征可以用文档级别或者语料级别的统计信息，同样在深度学习中这些特征也可以很容易的融入进来。常见的比如说词性标注，文档频率等。

- 序列编码器：使用词编码器后，对于序列中的每个单词都有一个向量表示，但是该表示只是基于单词的，并没有考虑的上下文的信息。所以需要序列编码器对该序列进行重新编码，得到新的单词表示，在新的表示中融入了上下文信息。可以使用任意的序列到序列编码器。

- 序列解码器：在序列标注模型中，对于每个单词都需要预测一个多元分类问题。在经过序列编码器之后，每个单词都有一个向量表示，为了预测每个单词对应的标签，需要序列解码器来完成从序列向量到对应预测标签的转换。可以使用任意的序列到序列解码器，见图 2.4。

根据词编码器，序列编码器和序列解码器使用不同的模块，可以将以往大部分实体识别模型总结如下，见表 2.2。

此外，还有许多的基于多语言，多任务，低资源的实体识别模型被研究 [201]。基于深度学习方法的实体识别比起传统有监督学习的实体识别通常可以取得更好的性能，并且不怎么需要人工设计复杂的特征，而是靠着神经网络自动抽取任务所需要的特征表示。然而，训练一个神经网络通常需要更多的数据，在实际的应

表 2.2 基于深度学习方法的实体识别系统汇总

文献	词编码器 词	词编码器 字符	词编码器 其他	序列编码器	序列解码器
[39]	-	-	Traditional Features	Convolution	Softmax
[40]	Word Embedding	-	-	Convolution	Softmax
[28, 57, 68, 195]	Word Embedding	-	-	biLSTM	CRF
[103]	-	-	Component enhanced	biLSTM	CRF
[167]	Word Embedding	-	-	biLSTM	CRF
[144]	-	CNN，biLSTM	-	biLSTM	CRF
[92]	-	Character Embedding	-	biLSTM	CRF
[119]	Word Embedding	CNN	-	biLSTM	CRF
[35]	Word Embedding	CNN	Lexicons Capitalization	biLSTM	CRF
[106]	Word Embedding	CNN	Feature Embedding	biLSTM	CRF
[162]	Word Embedding	CNN	-	Convolution	CRF
[13, 94]	Word Embedding	biLSTM	-	biLSTM	CRF
[209]	Word Embedding	biGRU	-	biGRU	CRF
[202]	Word Embedding	biLSTM	Affix	biLSTM	CRF
[233]	Word Embedding	Convolution + Attention Pooling	-	biGRU，Self Attention	CRF
[143]	-	Character Embedding	-	biLSTM	Softmax
[46]	Word Piece	-	Position Segment	Transformer	Softmax

用中，大量的标注数据通常也很难获得，这也是深度学习面临的一大问题。

2.1.3 常用数据集

高质量的标注数据对于模型的训练极为重要，本节将列举常用的实体识别数据集（英文）。一个实体标注语料库是包含一个或者多个实体类型注释的文档集合。表 2.3 [3] 列出了一些广泛使用的实体识别语料，以及相关的一些属性。

在不同的数据集中，可能所定义的实体类型并不一样，甚至实体的标注标准也存在差异。通常实体类型是该数据集合预先已经定义完成的。以 CoNLL03 数据集为例，该数据集包含 4 种实体类型，分别是 PER （persons），LOC （locations），ORG （organizations），MISC （miscellaneous）。而 OntoNotes 数据集标注了 18 种粗粒度，89 种细粒度的实体类型。网上还存在大量的实体识别的现在工具包，例如 StanfordCoreNLP，spaCy，NLTK，AllenNLP 等等。

[3] 该表格来源于文献 [99]。

表 2.3 英文实体识别语料列表

语料名称	年份	来源	实体类型数量	网址
MUC-6	1995	Wall Street Journal texts	7	https://catalog.ldc.upenn.edu/LDC2003T13
MUC-6 Plus	1995	Additional news to MUC-6	7	https://catalog.ldc.upenn.edu/LDC26T10
MUC-7	1997	New York Times news	7	https://catalog.ldc.upenn.edu/LDC2001T02
CoNLL03	2003	Reuters news	4	https://www.clips.uantwerpen.be/conll2003/ner/
ACE	2000 - 2008	Transcripts，news	7	https://www.ldc.upenn.edu/collaborations/past-projects/ace
OntoNotes	2007 - 2012	Magazine，news，conversation，web	89	https://catalog.ldc.upenn.edu/LDC2013T19
W-NUT	2015 - 2018	User-generated text	18	http://noisy-text.github.io
BBN	2005	Wall Street Journal texts	64	https://catalog.ldc.upenn.edu/ldc2005t33
NYT	2008	New York Times texts	5	https://catalog.ldc.upenn.edu/LDC2008T19
WiNER	2012	Wikipedia	4	http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner
WikiFiger	2012	Wikipedia	113	https://github.com/xiaoling/figer
N3	2014	News	3	http://aksw.org/Projects/N3NERNEDNIF.html
GENIA	2004	Biology and clinical texts	36	http://www.geniaproject.org/home
GENETAG	2005	MEDLINE	2	https://sourceforge.net/projects/bioc/files/
FSU-PRGE	2010	PubMed and MEDLINE	5	https://julielab.de/Resources/FSU_PRGE.html
NCBI-Disease	2014	PubMed	790	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
BC5CDR	2015	PubMed	3	http://bioc.sourceforge.net/
DFKI	2018	Business news and social media	7	https://dfki-lt-re-group.bitbucket.io/product-corpus/

2.2 关系抽取

在对文本进行实体识别之后，为了让机器利用更深的语义信息，需要将存在关系的实体对抽取出来，保存在数据库中，供其他任务使用，即关系抽取任务。针对此任务，研究者提出过了很多类方法，其中包括无监督学习算法、半监督学习算法、有监督学习算法以及远程监督学习算法。本节会首先介绍任务的描述（2.2.1 节），再总结其主要解决方法（2.2.2 节），最后列出常用的数据集（2.2.3 节）。

2.2.1 任务描述

“关系”的定义是指多个实体之间蕴含着某种语义联系。本文主要讨论二元关系，即只考虑两个实体之间存在的语义关系。对于“关系抽取”这个自然语言处理领域的术语，对应的两种不同的解释。一种是知识库级别的关系抽取，另一种是句子级别的关系抽取。本文将会覆盖这两种类型关系抽取的相关知识。事实上，一旦实例级别的关系抽取做完，可以很容易得到知识库级别的关系抽取结果。

- 知识库级别关系抽取：旨在产生一个关系的列表，对于每个关系实例，包含

两个实体以及其存在的语义关系类型。通常来说，该任务将大规模的文本语料作为输入，然后输出一个关系的列表。

- 句子级别关系抽取：通常需要给定文档或者句子，以及两个候选实体，然后确定该实对存在某种语义关系。如图 2.6，当给定句子和两个候选实体，关系抽取的目标是确定“华为”和“任正非”在该句子中具有雇佣关系。

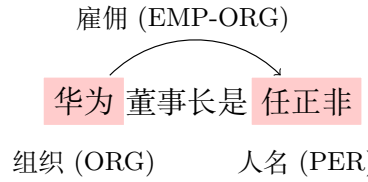

图 2.6 关系抽取样例

评价指标：不管是知识库级别的关系抽取还是句子级别的关系抽取，通常都是采用 F1 值作为评价指标。对于知识库级别的关系抽取，存在一个正确的关系列表，也有预测的关系列表，根据公式 2.1 很容易计算对应的 F1 值。同样的，对于句子级别的关系抽取，也可以计算对应的 F1 值。

2.2.2 相关方法

本节将关系抽取的相关方法大致分为五类，分别是基于无监督学习、半监督学习、传统有监督学习以及远程监督方法，同时也介绍近年来深度学习模型的关系抽取方法。以下将一一详细叙述。

基于无监督学习的关系抽取

当没有任何标注数据时，完全无监督的关系抽取主要有基于聚类的方法。这类方法可以简单归纳为：

1. 使用某种实体抽取系统将文本中的实体标注出来；
2. 记录共现的实体对以及其上下文；
3. 计算步骤 2 中的到实体对的上下文相似度；
4. 根据计算得到的相似读进行聚类；
5. 聚类的到的每一个集群都代表一个关系，因此会自动为每个集群分配一个标签，描述由它表示的关系类型。

算法 2.1 Yarowsky 算法

输入: 无标注数据 D 和少量标注数据 S，分类器 C
输出: 训练好的分类器 C

```
while 收敛条件未达到 do
    在 S 上训练分类器 C
    用分类器 C 对数据 D 打标签
    N = 分类器 C 输出置信度前 n 的样本
    S = S ∪ N
    D = D − N
end while
```

文献 [31, 43, 59, 93, 151, 170, 205, 211] 就是此类方法的代表。无监督的方法通常需要大规模的自由文本作为支撑，利用大规模自由文本的冗余性，挖掘可能的关系模式集，然后确定其关系名称。该方法的不足之处在于关系名称难以准确描述，并且对于低频关系的召回率特别低。

基于半监督学习的关系抽取

当具有少量的标注数据以及大量的未标注数据时，半监督学习算法成为自然语言处理领域的一个重要课题。对于半监督的关系抽取，同样获得很多研究者的关注。半监督的关系抽取大都基于文献 [18, 213] 提出的算法。文献 [213] 提出 Yarowsky 算法，文献 [18] 提出 Co-training 算法。文献 [2] 证明了 Yarowsky 算法是 Co-training 算法的一种特殊情况。Yarowsky 算法大致流程参照算法 2.1。其主要思想是利用弱的分类器的一部分置信度比较高的输出作为下一次迭代的训练数据。

早期比较经典的是 DIPRE（Dual Iterative Pattern Relation Expansion）算法 [20]。假设目标关系是（author, book），DIPRE 从很小的种子集合开始。种子集合中是少量的（author, book）实例，再假设当前的种子样本只有一个，即（Arthur Conan Doyle, The Adventures of Sherlock Holmes）。该系统会用搜索引擎检索出一些包含该种子的很多实例。为了从这些实例中学习到规则，需要定义规则模板，例如 [order, author, book, prefix, suffix, middle]，其中 order 表示关于作者的字符串是否出在书名之前，author 表示作者的字符串，book 表示书名的字符串，prefix 表示该关系左侧字符串，suffix 表示该关系右侧字符串，middle 表示作者和书名之间字符串，其他特征含义请参考原文献。根据规则模板可以对这些实例进行分组，每一组都有着如下的结构：

[longest-common-suffix of prefix strings, author, middle, book, longest-common-prefix of suffix strings],

然后对于规则进行简化，加入某些通配符。再通过搜索引擎进行搜索，从而的到更多的实例句子，以此方法不断的迭代。一般形式的 DIPRE 流程参照算法 2.2。

从这里可以看出，DIPRE 算法和 Yarowsky 算法非常相似。都是从很小的的种子样本出发，DIPRE 使用的分类器是规则匹配，通过从种子关系中提取规则进行迭代训练。给定一个字符串，如果它与某个规则匹配，则该字符串被分类为正例，并用于提取新的关系，否则为负例。在种子集合中加入新的关系，对分类器进行再训练。DIPRE 是 Yarowsky 算法在关系提取中的应用实例。基于此架构，Snowball [3] 被人提出，可以看成是对 DIPRE 的进一步改进。

与 DIPRE 和 Snowball 算法不一样，KnowItAll [50] 使用一小部分领域无关的规则来的到训练需要的种子样本。当提取特定的关系时，这些领域无关的规则可以进一步变换为基于特定关系的规则，从而可以达到抽取特定领域的关系。DIPRE，Snowball，KnowItall 都是基于特定关系的抽取系统。换句话说，目标关系必须由人提前指定。TextRunner [8] 就是为了解决此问题。该系统不需要在输入中指定关系，而是以自监督的方式在大量的文本中抽取关系。

半监督的关系抽取还有很多种其他的方法，例如主动学习（Active Learning），标签传播（Label Propagation）等 [138] 。

算法 2.2 DIPRE 算法

```
使用种子样本标注一些数据
while 收敛条件未达到 do
    从标注的数据中总结规则
    应用规则到自由文本，从而的到更多的样本
end while
```

基于有监督学习的关系抽取

有监督的关系抽取通常是句子级别的。这些方法需要具有一定量的标注数据，即每个实体对都用一种预定义的关系类型标记。如果某个实体对没有关系，则可以引入一个特殊的关系类型 `None` 表示。一般来说，有监督的关系抽取是一个多元分类问题，每一个类别对应一个不同的关系类型（包含无关系类型 `None`）。这些方法大致可以分成两类，基于特征的方法和基于核函数的方法。

给定一定数量的正负样本，可以从中提取句法和语义上的特征。这些提取的特征可以用来判断句子中两个实体具有某种关系的提示。常用的特征见表 2.4。所提取的句法和语义特征都以特征向量的形式呈现给分类器进行训练或者预测。文献 [78] 基于这些特征训练一个最大熵分类器进行关系抽取。文献 [56] 则更一步加入更多的特征，并采用支持向量机作为分类器，取得了更好的结果。一些特征可以很好的预测对应的关系，然而也有一些特征并没有效。如何选择这些相关性比较大的特征也是非常重要的一部分。文献 [74] 对此进行了系统的探究。总体来说，为了最大限度的提升性能，需要人为启发式设计特征。由于一般的自然语言处理

任务的输入都是结构化的数据，尤其是关系抽取任务，很难获得相关特征的最佳子集。为了解决这种问题，基于核函数的关系抽取被提出来。核函数提供了一种新的方式，在高维空间以隐式的方法探索输入的表示。

表 2.4 传统有监督关系抽取的常用特征

句法特征	语义特征
实体字符串的特征 两个实体类型的特征 两个实体中间的单词序列 两个实体中间单词的数量 解析树上包含两个实体的路径 …	实体的 WordNet 相关特征 …

在基于核函数的方法中，需要设计核函数来计算两个关系实例表示之间的相似性，然后采用支持向量机进行分类。大多数的基于核函数的方式都是根据它们之间共享的子表示（例如子序列或者子树）的数量来度量两个关系实例之间的相似性。本文将此基于核函数的方法大致划为以下三类：

- 基于序列的核函数：受基于字符串的核函数在文本分类应用启发 [116]，文献 [131] 将此应用到关系抽取任务，提出一种基于序列的核函数。该核函数可以计算两个序列在词级别上的相似度而不是字符级别的相似度 [116]。同时设计了三种核函数分别捕获实体对左边，实体对之间，实体对右边的上下文信息，最终简单的用求和的方式将三种核函数统一起来，再放入支持向量机分类器中进行训练和预测，在关系抽取任务中取得不错的提升。

- 基于树的核函数：文献 [217] 提出基于句子浅层解析树的核函数，与之前基于序列的核函数的区别是可以考虑到结构化的信息。类似地，文献 [41] 提出基于句子依存句法树的核函数。总之，文献 [41, 217] 使用树形式更丰富的信息，以便于在关系抽取任务中获得良好的性能。

- 基于依存路径的核函数：文献 [23] 发现一个非常有趣的观察，依存树中两个实体之间的最短路径编码了足够的信息来提取它们之间的关系。因此提出一个基于依存路径的核函数方法。与基于树的核函数方法相比，这种方法的计算效率更加高效，具有更加简化的特征空间，同时只需要线性时间复杂度，最终的性能也比基于树的核函数效果要好。

最后，表 2.5 对基于特征和基于核函数的方法做一个简单对比。

表 2.5 监督关系抽取的方法对比

	特征集合	计算复杂度
基于特征的方法	需要在文本分析以及实验后定义特征集合	相对低
基于核函数的方法	不需定义特征集合，在高维空间隐层计算相似度	相对高

基于远程监督的关系抽取

基于监督学习的关系抽取效果往往依赖于大量的训练数据，人工标注的数据往往很少，这就限制了监督学习关系抽取的发展。远程监督方法的兴起 [127]，为关系抽取领域打开一扇新的大门。本质上，远程监督方法是一种基于规则的方法，可以自动生成大量的训练数据，但是其中会有不少的噪音数据。

远程监督学习的主要做法是将文档与知识库对齐，然后基于一个假设：如果知识库中的实体对存在某种关系，那么则包含这个实体对的每个文档均表达了此关系。从这里可以看出，该假设是一个非常强的假设。实际上，可能很多包含这个实体对的文档并不表达此种关系。例如，给定知识库中存在关系（任正非，华为，创始人）和一个句子“任正非将华为推到了全世界”，该句子出现了“任正非”和“华为”这两个实体，但是并没有表达“创始人”这种关系。

表 2.6 基于卷积神经网络的远程监督关系抽取模型的特征总结

文献	多实例学习中 Bag 表示	最大池化方式
[219]	One sentence per bag	Piecewise in a sentence
[107]	Attention weighted sum over bag	Piecewise and Full
[76]	Max of each feature over bag	Cross sentence in a bag

为了缓解远程监督中的噪音数据问题，许多方法被提了出来。文献 [156] 通过将问题建模成一个多实例问题来松弛远程监督的假设。相似地，文献 [65, 76, 176] 采用多实例多标签方法对问题进行建模。随着神经网络的兴起，卷积神经网络 [218] 在远程监督的关系抽取上成为广泛使用的网络架构。典型的几种方法和特征总结如表 2.6 [4]。文献 [219] 将多实例学习的神经网络用在了远程监督，并提出 PCNN。文献 [107] 针对多实例问题，对包中的所有实例应用了注意力机制。基于对抗训练的模型 [191]，基于噪音的模型 [118]，基于软标签的模型 [113, 185] 也纷纷被提出。文献 [147, 148] 提出利用生成对抗网络和强化学习的方法直接识别出噪音数据，从而提升远程监督的性能。最近，图卷积神经网络 [183] 胶囊网络 [222] 也被应用到远程监督的关系抽取中。此外，语言学知识和语义知识也证明对此任务有帮助，但是这样的系统通常依赖于显式的特征，比如依存树，实体类型以及关系别名等 [183, 203]。

[4] 表格改编于文献 [91]。

基于深度学习的关系抽取

最近，基于深度学习的关系抽取获得了很多研究者的关注。本节将对当前主流的方法进行叙述，主要聚焦在有监督学习的句子级别的关系抽取。在此设置下，关系抽取可以看成多类分类任务。根据使用的编码器以及是否使用依存句法树，可以大致将相关系统划分为三种：基于卷积神经网络的关系抽取，基于循环神经网络的关系抽取和基于依存句法树的关系抽取。

表 2.7 基于卷积神经网络和循环神经网络的关系抽取方法

	词编码器	句子编码器	损失函数
[108]	word embedding lexical features synonym	CNN	Cross Entropy
[218]	word embedding position embedding lexical features	CNN	Cross Entropy
[136]	word embedding position embedding	CNN	Cross Entropy
[163]	word embedding position embedding	CNN	Pairwise Ranking
[67]	word embedding position embedding POS Tag	Attention-CNN	Pairwise Ranking
[187]	word embedding position embedding	CNN，Attention	Ranking
[96]	word embedding position embedding POS Tag，entity type	CNN	Cross Entropy
[220]	word embedding	biLSTM，Max Pooling	Cross Entropy
[224]	word embedding lexical features relative position	biLSTM，MLP	Cross Entropy
[231]	word embedding	biLSTM，Attention	Cross Entropy
[194]	word embedding	biLSTM，Attention Average Pooling，Max Pooling	Cross Entropy
[97]	word embedding relative position latent entity features	biLSTM，Attention	Cross Entropy

- 卷积神经网络：当给定两个实体时，对于句子中的每个词的表示，可以加入实体相关信息，以及其他额外的词法信息，比如词性标注。使用卷积神经网络时，为了考虑到序列信息，通常需要加入位置词向量（Position Embedding）。经过卷积之后，再通过池化操作即可得到当前实体相关的整个句子的表示。最终

选择合适的损失函数来指导模型的训练。主流的基于卷积神经网络的关系抽取方法见表 2.7。使用卷积神经网络的优势主要是可以有效提取局部的特征，并且容易并行，计算速度比较快。其缺点是难以捕获到长距离的信息。

- 循环神经网络：与卷积神经网络相比，循环神经网络天然更适合处理带有序列信息并且变长的文本。所以自然而然地许多基于循环神经网络的关系抽取方法被提出。实际使用中，简单的循环神经网络具有梯度消失等问题，为了缓解这些问题，通常使用循环神经网络的变体，比如长短时记忆网络（LSTM）等，进一步可以加入注意力（Attention）机制直接捕获长距离的依赖。主流的基于循环神经网络的关系抽取方法见表 2.7。相比较卷积神经网络，循环神经网络能够捕获长距离的依赖。然后由于循环的机制是的计算很难并行，所以计算效率不如卷积神经网络。

表 2.8 基于依存句法树的关系抽取方法（损失函数均为 Cross Entropy）

	词编码器	句子编码器
[173]	word embedding，word matrix **dependency tree**	RNN （Recursive）
[215]	word embedding，NER **dependency tree**，WordNet	FCM
[115]	word embedding，NER，WordNet **augmented dependency path**	RNN（Recursive） CNN
[204]	word embedding，NER，WordNet **shortest dependency path** grammatical relations	RNN（Recursive）
[196]	word embedding **shortest dependency path**	CNN
[200]	word embedding，POS Tag **shortest dependency path** grammatical relations，WordNet	RNN，Max Pooling
[25]	word embedding，POS Tag **shortest dependency path** WordNet	LSTM，Convolution，Pooling
[112]	word embedding，relative position **sub-tree parse**	biGRU，Attention

- 依存句法树：在传统方法的关系抽取中，依存句法树可以为关系抽取提供非常强大的特征。在深度神经网络的时代，如何有效的利用依存句法信息也是非常重要的方向。在实际应用中，可以对依存路径使用卷积或者循环神经网络。或者直接利用递归神经网络在依存树上的到基于树的表示。这些方法都可以将结构化的树的信息融合到神经网络中，结果也证明了这些方法的有效性。主流的基于依存句法树的关系抽取方法见表 2.8。

深度神经网络在关系抽取中的发展，最重要的是减少了人工设计的特征，让网络自动提取出需要的特征。并且，传统的特征也可以很简单地加入到深度学习模型中，不管以传统特征向量的形式还是类似于词向量的形式，均可进一步提升关系抽取的性能（比如依存句法树相关特征）。

2.2.3 常用数据集

本节介绍常用的关系抽取数据集。对于有监督的实体关系抽取（包括基于传统方法和深度学习方法）通常是句子级别的，并且这种数据需要大量的人工标注，这意味着数据都是高质量的，几乎没有噪音。但是由于人工标注费时费力，因此这样的数据集通常很小。下面提到的数据集都是已经标注好实体和关系的文档。

- ACE：该系列可以用来进行关系抽取任务的数据集有 ACE 2003，ACE 2004，ACE 2005。该系列标注的实体类型和关系类型稍有差别。比如 ACE 2004 具有 7 种关系类型（PHYS，PER-SOC，EMP-ORG，ART，OTHER-AFF，GPE-AFF，DISC），而 ACE 2005 的关系类型有 6 种（PHYS，PER-SOC，ORG-AFF，ART，GPE-AFF，PART-WHOLE）。
- SemEval-2010 Task 8：该数据集是 SemEval 2010 年任务 8 使用的 [62]。总共包含 10717 个样本，其中 8000 个作为训练，2717 个作为测试。一共具有 9 种关系类型，当考虑到方向和其他关系时，总共有 $2*9+1$ 种关系类型。
- TACRED：由文献 [225] 提出的大规模关系抽取数据集。通过众包标注每年的 TAC KBP 评价（2009-2015）。
- FewRel：该数据集是由清华大学自然语言处理实验室公布 [58]。共有 70000 个句子包含 100 种关系类型，来源于维基百科，并通过众包进行标注。
- DocRED：该数据集也是由清华大学自然语言处理实验室公布 [212]，是目前最大的文档级别的关系抽取数据集，同时标注了实体和关系。

以上数据集及其更多数据集的统计信息见表 2.9 [5]。

另一方面，远程监督的关系抽取使用的数据集是通过将 Freebase 中的关系和 New York Times 语料库对齐产生。使用了斯坦福实体标注工具标注其中的实体，并与 Freebase 中的实体进行匹配。总共具有 53 种关系类型（包括一中特殊的关系 None 表示实体间不存在关系）。训练数据中包含 522611 个句子，281270 个实体对和 18252 关系事实，测试数据包含 172448 个句子，96678 个实体对和 1950 个关系事实。

[5] 来源于文献 [212]。

表 2.9 关系抽取数据集统计

数据集	文档数	单词数	句子数	实体数	关系类型数	关系实例数
SemEval-2010 Task 8	-	205k	10717	21434	9	8853
ACE 2003-2005	-	297k	12783	46108	24	16771
TACRED	-	1823k	53791	152527	41	21773
RewRel	-	1397k	56109	72124	100	70000
BC5CDR	1500	282k	11089	29271	1	3116
DocRED	5053	1002k	40276	132375	96	63427

2.3 联合实体关系抽取

实体关系抽取任务需要识别自由文本中的实体，并且抽取出存在关系的实体对。显然地，实体关系抽取包含两个子任务，首先需要执行实体识别，然后对抽取的实体对做关系抽取。该任务是信息抽取领域重要的子任务之一，近些年来，也获得了学术界和工业界的广泛关注。针对此任务，简单的有流水线式的解决方案，接着又有许多学者提出联合模型解决此问题。本节首先介绍任务的描述（2.3.1 节），再对主流的相关方法进行总结（2.3.2 节），最后会介绍常用的数据集（2.3.3 节），尤其是本文后需要用的数据集。

2.3.1 任务描述

给定一段自由文本，实体关系抽取的目标是识别文本中的实体，并找出存在关系的实体对。其实就是实体识别（2.1.1 节）和关系抽取（2.2.1 节）两个任务合在一起。如图 2.6，实体关系抽取任务可以识别出句子中的两个实体，“华为”的实体类型是组织（ORG），“任正非”的实体类型是人名（PER）。两个实体在这个句子中表达了雇佣（EMP-ORG）关系。

评价指标: 实体识别结果和关系抽取结果均采用 F1 值的方法。详细计算方法请参照 2.1.1 节和 2.2.1 节评价指标部分。

2.3.2 相关方法

对于实体关系抽取任务，出现了很多解决方法，从一开始的流水线解决方案到近些年来的联合解决方案，都有效的提升了实体关系抽取的性能，为其他任务做好了铺垫。本文大致将实体关系抽取相关方法归为两大类，分别是流水线方法和联合方法，以下将分别叙述各自的主流模型和相应的利弊。

背景介绍

解决实体关系抽取任务最简单是的方法就是流水线的方法，将实体识别和关系抽取看成两个独立分开的任务 [29, 130]。该方法简单有效，非常灵活。实体模型和关系模型分别训练，对于新来的句子，只需要先用实体模型进行实体识别，然后再用关系模型进行判断实体对中存在的语义关系。此外，在训练模型时，实体模型和关系模型可以用任意的各自的数据集，并不需要同时标注了关系和实体的数据集。这样对于单独的模型，可以很方便地进行加数据重新训练而不影响另外的模型。但是，这种流水线式的建模方法导致了错误传播问题。即如果实体模型识别实体出现错误，会直接影响后面关系模型的效果。另外，由于实体模型和关系模型分别建模，一些实体模型和关系模型之间的交互没有考虑到，而这些信息对于实体关系抽取尤为重要。为了缓解这些问题，许多学者提出联合实体模型和关系模型的建模方法。

联合方法

随着神经网络的普及，多个任务可以很方便的共享某部分参数，实现联合训练，以达到多个任务互补的目标。因此，对实体模型和关系模型联合建模，比较直接的方式就是共享参数。在此之上，为了更好地建模实体模型和关系模型的交互，许多联合解码算法也被提出。以下将详细介绍这两类方法在联合实体关系抽取中的应用。

共享参数 文献 [128] 的实体模型采用了句子级别的循环神经网络，而关系模型设计了一个基于依存句法树的循环神经网络，关系模型的特征输入来源于实体模型中循环神经网络的隐层状态，整个模型联合训练。文献 [81] 具有相似的基于循环神经网络的实体模型，不同的是使用注意力机制实现关系模型。这些提出的联合模型都是通过共享参数实现的。实体模型和关系模型之间只有隐式的参数共享，并没有显式地刻画两个模型之间的交互。

联合解码 为了加强两个子模型之间的交互，一些联合解码算法被提出。文献 [207] 提出使用整数线性规划（ILP）对实体模型和关系模型的预测结果进行强制约束。文献 [80] 利用条件随机场（CRF）同时建模实体和关系模型，并通过维特比解码算法得到实体和关系的输出结果。文献 [228] 将实体识别和关系抽取统一成一个序列标注问题，采用简单的双向循环神经网络进行求解。文献 [102] 将实体关系抽取看为一个结构化预测问题，采用结构化感知机算法，设计了全局特征，并使用集

束搜索进行近似联合解码。文献 [221] 提出使用全局归一化（Global Normalization）解码算法。文献 [188] 针对实体关系抽取设计了一套转移系统（Transition System），从而实现联合实体关系抽取。对于当前主流的联合实体关系抽取模型，根据是否共享参数和联合解码的方式总结见表 2.10。

表 2.10 联合实体关系抽取方法总结

文献	词编码器	句子编码器（实体）	关系编码器	共享参数	联合解码
[128]	word embedding POS Tag dependency tree	biLSTM	tree biLSTM	是	无
[81]	word embedding	biLSTM	Attention	是	无
[207]	traditional features	-	-	否	整数线性规划
[102]	traditional features	-	-	否	结构化感知机 集束搜索
[80]	word embedding	biLSTM	biLSTM	是	条件随机场
[228]	word embedding	biLSTM	biLSTM	是	序列标注标签扩展
[221]	word embedding character embedding POS Tag dependency tree	biLSTM	biLSTM-Minus MLP	是	全局归一化
[188]	word embedding	biLSTM	biLSTM	是	转移系统

表 2.11 ACE05 数据集实体类型和关系类型列表

实体类型	关系类型
PER （Person） ORG （Organization） GPE （Geographical Entities） LOC （Location） FAC （Facility） WEA （Weapon） VEH （Vehicle）	ART （Agent-Artifact） PART-WHOLE （Part-Whole） PER-SOC （Person-Social） PHYS （Physical） GPE-AFF （GPE-Affiliation） ORG-AFF （Organization-Affiliation）

2.3.3 常用数据集

通常关系抽取的数据集都可用来做实体关系抽取任务。在本节将会详细介绍本文使用到的实体关系抽取数据集，着重于 ACE05 和 NYT 数据集。

- ACE05：该数据集是常用的实体关系抽取数据集，是完全由人工标注。在本文的使用中，共有 7 种实体类型和 6 种关系类型。数据集的划分参照于文献 [102]，其中 351 个文档作为训练，80 个文档作为验证，80 个文档作为测试。更为详细的数据集介绍见表 2.11。
- NYT：该数据集包含大量的数据，共有 3 种实体类型（PER，ORG，LOC）和 24 种关系类型。训练数据包含超过 353k 个关系元组，并且训练数据并非人工标注，而是通过远程监督学习得到 [156]。测试数据包含 3880 个关系元组，该

部分数据是人工标注。和之前的工作类似 [155, 228]，本文排除标注为 None 的关系，并且从中随机选取 10% 的标注数据作为验证，余下的作为测试。

2.4 深度学习基础

在过去几年里，神经网络重新成为强大的机器学习模型，并在计算机视觉（Computer Vision，CV）和语音处理（Speech Processing，SP）等领域产生了很多最好的结果。近几年来，神经网络模型也开始应用在自然语言处理领域，也取得了不错的结果。深度学习是一种特殊的统计机器学习技术。为了更好地理解深度学习，本节会先介绍必要的统计机器学习基础，然后介绍结构化学习基础，最后叙述常见的网络结构以及特点。

2.4.1 统计机器学习基础

统计机器学习算法使计算机能够从数据中学习。以经典的二分类问题为例，给定一个**训练数据**集合：$\mathcal{D} = \{x_i, y_i\}_{i=1}^{|\mathcal{D}|}$，其中 $\mathcal{X}$ 是输入样本点，$y_i \in \mathcal{Y} = \{0, 1\}$ 是对应的正确标签。统计机器学习算法的目标是通过训练数据 $\mathcal{D}$，学习一个好的近似函数 $f : \mathcal{X} \mapsto \mathcal{Y}$。当新的数据到来时，$f$ 可以给出好的预测结果。假设存在概率空间 $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}, \mathbb{P})$，其中 $\mathcal{F}$ 是 $\mathcal{X} \times \mathcal{Y}$ 上的 σ 代数，$\mathbb{P}$ 是 $(\mathcal{X} \times \mathcal{Y}, \mathcal{F})$ 上的概率测度，即 $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$。机器学习模型 f 的设计目标为最小化泛化错误：

$$\mathbb{P}(f(x) \neq y) \tag{2.2}$$

定义损失函数 $\mathcal{L}(f(x), y)$：

$$\mathcal{L} : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R} \tag{2.3}$$

$\mathcal{L}$ 度量了预测错误的程度，损失函数值越小说明模型越好。因此可选 f 最小化**泛化损失函数**，即 $\mathcal{L}$ 在概率 $\mathbb{P}$ 上的期望：

$$\int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}(f(x), y)\, \mathrm{d}\mathbb{P} = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}(f(x), y)\mathbb{P}(x, y)\, \mathrm{d}x\, \mathrm{d}y \tag{2.4}$$

由于联合分布 $\mathbb{P}$ 未知，泛化损失函数无法直接计算。但是我们具有训练数据 $\mathcal{D}$，函数 f 关于 $\mathcal{D}$ 的平均损失称为**经验损失函数**，计算公式为：

$$\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathcal{L}(f(x_i), y_i) \tag{2.5}$$

根据大数定律，当训练集合 $\mathcal{D}$ 规模趋于无穷大时，经验损失函数趋于泛化损失函数。之后我们需要选择最优的 f 使得经验损失最小，此时的问题归结为最优化问题。如果该最优化问题有闭式解，则可直接输出最优的 f。通常来说闭式解不存在，需要用数值计算的方法进行求解，比较流行的有基于梯度的优化方法，比如随机梯度下降（Stochastic Gradient Descent，SGD）等。

2.4.2 结构化学习基础

有别于传统的统计机器学习，在自然语言处理领域，许多任务的标签集合 $\mathcal{Y}$ 存在“结构”。以词性标注任务为例，给定一个单词序列长度为 5 的句子“This is a tagged sentence”，对应的输出也是长度为 5 的序列“DT VBZ DT JJ NN”。在自然语言处理领域，有很多这样输出为结构化对象的任务。实际上，结构化任务可视为具有很多个类别的分类任务。由于类别无法一一枚举，导致无法应用传统的机器学习分类算法解决。为了解决这样的结构化问题，本文定义**联合评分函数** $s : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$。$s(x, y)$ 表示输入为 x 输出为 y 时的得分。当给定评分函数时和输入 x，对应的输出 $f(x)$ 即为：

$$f(x) = \underset{y \in \text{GEN}(x)}{\arg\max}\, s(x, y) \tag{2.6}$$

其中 GEN 是生成候选输出的函数并且有 $\text{GEN}(x) \subset \mathcal{Y}$。为了简单起见，通常情况下联合评分函数设计为线性函数，即：

$$s = w^T \phi(x, y) \tag{2.7}$$

其中 $\phi : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^n$ 为**联合特征函数**，n 是特征的维度，并且是固定的，不随着输入输出的改变而改变。$w \in \mathbb{R}^n$ 是需要训练的参数。最终的模型函数 f 为：

$$f(x) = \underset{y \in \text{GEN}(x)}{\arg\max}\, w^T \phi(x, y) \tag{2.8}$$

当给定的训练数据 $\mathcal{D}$ 有标签时，除了需要根据特定任务设计联合特征函数 $\phi(x, y)$ 外，还需要选个合适的损失函数，比如常见的 Hinge Loss，形式如下：

$$\mathcal{L}(x, y) = [1 - s(x, y) + \max_{\hat{y} \in \text{GEN}(x)} s(x, \hat{y})]_+ \tag{2.9}$$

其中 $[a]_+ = \max(a, 0)$。根据损失函数确定参数 w 的过程称之为**学习**（Learning）或**训练**（Training），通常使用基于梯度的优化方法求解。当参数 w 固定时，求解等式 2.8 的过程称之为 **推理**（Inference）或 **解码**（Decoding）。通常情况结构化问题并不

能枚举所有的标签 $\mathcal{Y}$，所以可能需要设计特定的**解码算法**有效地得到精确的或者近似的解。常用的精确解码算法一般是基于动态规划，比如维特比（Viterbi）。常用的近似解码算法有集束搜索（Beam Search）。在实际应用中，通常需要在 $\phi(x,y)$ 联合特征函数的丰富性和解码算法的精确性上做权衡。换句话说，当 $\phi(x,y)$ 比较强大时（比如使用任意阶特征），解码算法通常都是近似的。当使用精确的解码算法时，$\phi(x,y)$ 通常是受到限制的（比如使用一阶或者二阶特征）。在不同的任务中，使用合适的解码算法是非常重要的，关系到最终模型的效率和精度。

2.4.3 常见网络介绍

现代深度学习为监督式学习（Supervised Learning）提供了一个强有力的框架。通过叠加更多的层数以及在每层使用更多的隐层单元，一个深度网络可以表示更大复杂度的函数。当给定足够大的模型和足够多的标注数据，深度学习模型通常可以取得不错的性能。接下来介绍在自然语言处理领域常用的几个网络。

前馈神经网络

前馈神经网络（Feedforward Neural Network，FNN），又名多层感知机（Multiple Layer Perceptron，MLP），是最简单常见的一种神经网络结构。顾名思义，神经网络受人工大脑计算机制启发，这种机制依赖于称之为神经元的计算单元。神经元是一个具有标量输入和标量输出的计算单元。对于每一个输入都关联一个权重，神经元将每个输入乘以其权重，然后求和，再对结果施加非线性变换，并将其传递给输出。神经元之间相互连接形成一个网络：一个神经元的输出可以馈入一个或者多个神经元的输入。这种网络已经被证明具有非常强大的计算能力。如果权重设置合适，一个具有足够多神经元和足够多非线性变换的神经网络可以近似非常广泛的数学函数。

一个经典的前馈神经网络如图 2.7所示 [6]，每一个圆圈代表一个神经元，传入箭头是神经元的输入，传出箭头是神经元的输出。每个箭头都关联一个权重，表示其重要性（图中未显示）。神经元按层排列，反映了信息的流动。最底层没有传入箭头，是网络的输入。最顶层没有传出箭头，是网络的输出。其它的层可以认为是“隐藏”的。中间层神经元中的 sigmoid 形状代表非线性变换（通常为 $1/(1+\exp(-x))$，在将其传递给输出之前应用于神经元的值。在图中，每个神经元都连接到下一层的所有神经元，所以这也称为全连接层（full-connected layer）或仿射层（affine layer）。

接下来，我们用数学符号解释前馈神经网络。网络中每一行神经元的值可以

[6]图 2.7改编于论文 [54] 图 2

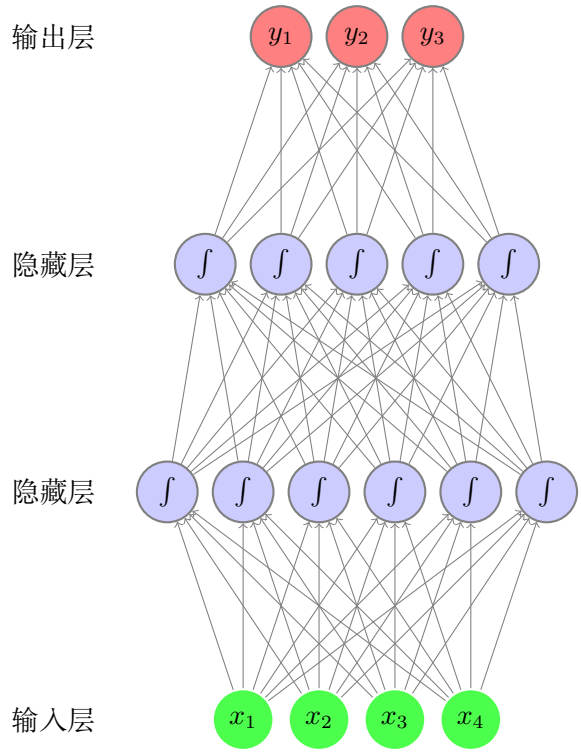

图 2.7 前馈神经网络（隐藏层数为 2）

看作是一个向量。以图 2.7为例，输入层是一个 4 维的向量 $\mathbf{x}$，其上的层是一个 6 维的向量 $\mathbf{h}^1$。全连接层可以看成是从 4 维到 6 维的线性变换。一个全连接层等价于一次矩阵-向量乘法，$\mathbf{Wx}$，其中从输入行第 j 个神经元到输出行第 i 个神经元的连接权重为 W_{ij}。然后，$\mathbf{h}$ 再经过非线性变换 g。从输入到输出的整个过程可以写成：$g(\mathbf{Wx})$。根据以上所述，图 2.7用数学公式表达出来，如下：

$$\mathrm{FFN}(\mathbf{x}) = \mathbf{W}^3(g^2(\mathbf{W}^2 g^1(\mathbf{W}^1\mathbf{x}))) \tag{2.10}$$
$$\mathbf{x} \in \mathbb{R}^4, \mathbf{W}^1 \in \mathbb{R}^{6\times4}, \mathbf{W}^2 \in \mathbb{R}^{5\times6}, \mathbf{W}^3 \in \mathbb{R}^{3\times5}$$

也可以使用中间变量使得表达更为清晰：

$$\mathbf{h}^1 = g^1(\mathbf{W}^1\mathbf{x}) \tag{2.11}$$
$$\mathbf{h}^2 = g^2(\mathbf{W}^2\mathbf{h}^1) \tag{2.12}$$
$$\mathbf{y} = \mathbf{W}^3\mathbf{h}^2 \tag{2.13}$$
$$\mathrm{FFN}(\mathbf{x}) = \mathbf{y} \tag{2.14}$$

循环神经网络

在自然语言处理领域，处理序列的情况时常存在，例如单词（字母序列），句子（单词序列）。如果使用之前的前馈神经网络，输入必须是定长的，比如输入向量可以是词袋模型（Bag of Word），长度等于词表大小。但是，词袋模型存在很严重的缺陷，就是忽略了顺序信息，而在自然语言里，顺序信息极为重要，不同的顺序可能有着截然不同的语义。循环神经网络（Recurrent Neural Networks，RNNs）[49] 可以将任意长度的句子表示为一个固定的向量，同时保留了句子中的顺序信息，即当顺序不一样时得到的句子表示也不一样。接下来从数学公式的角度介绍循环神经网络。

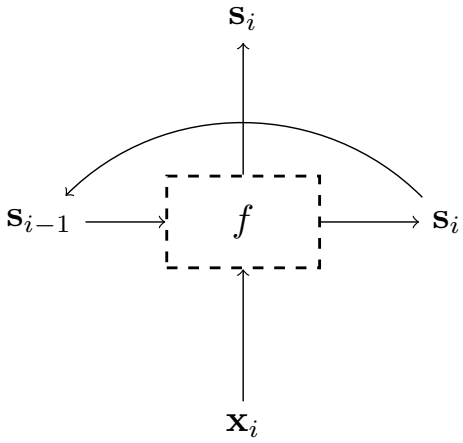

图 2.8 循环神经网络（递归形式）

通常用 $\mathbf{x}_{i:j}$ 表示单词向量的序列 $\mathbf{x}_i, \ldots, \mathbf{x}_j$，每个 $\mathbf{x}_i$ 即为第 i 个单词的向量表示，可以使用最简单的词的 One-Hot 表示，或者比较流行的词向量（Word Vector），或者叫做词嵌入（Word Embeding）。循环神经网络的输入是一组有序的向量列表 $\mathbf{x}_1, \ldots, \mathbf{x}_n$（$n$ 表示句子中单词的总个数）以及一个初始**状态向量**（state vector）$\mathbf{s}_0$，返回一组对应的有序状态向量列表 $\mathbf{s}_1, \ldots, \mathbf{s}_n$。输入向量 $\mathbf{x}_i$ 以顺序方法呈现给循环神经网络，状态向量 $\mathbf{s}_i$ 表示在观察了输入序列 $\mathbf{x}_1, \ldots, \mathbf{x}_i$ 之后得到的循环神经网络状态信息，也就是说 s_i 可以看作融合了序列信息 $\mathbf{x}_1, \ldots, \mathbf{x}_i$ 之后第 i 个单词的新的表示。很自然地，$\mathbf{s}_n$ 即为整个句子的表示。我们可以用公式表示整个过程：

$$\mathrm{RNN}(\mathbf{s}_0, \mathbf{x}_{1:n}) = \mathbf{s}_{1:n} \tag{2.15}$$

具体到每一个时刻，有：

$$\mathbf{s}_i = f(\mathbf{s}_{i-1}, \mathbf{x}_i) \tag{2.16}$$

在这个序列过程中共享一个函数 f。循环神经网络递归形式的图形化表示如图 2.8。此递归形式适用于任意长的序列。当给定一个有限的输入序列，我们可以将此递归展开，如图 2.9。

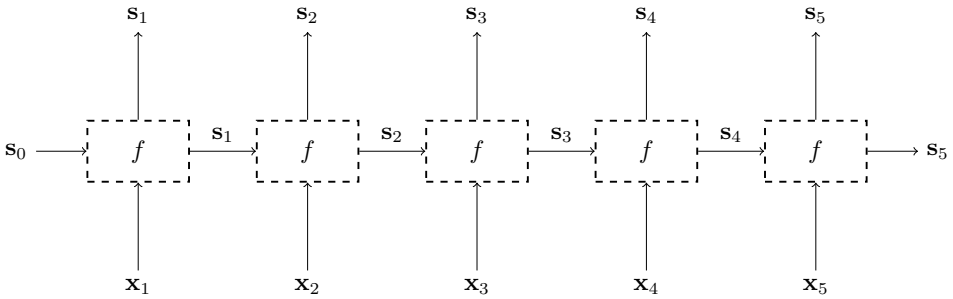

图 2.9 循环神经网络（展开形式）

循环神经网络可以叠加很多层，形成一个类似网格的结构 [48]，称为多层循环神经网络（Multi-layer RNN or Stacked RNN）。考虑一个 L 层的循环神经网络 $\text{RNN}_1, \ldots, \text{RNN}_L$，第 j 个循环神经网络 RNN_j 的隐层状态为 $\mathbf{s}_{1:n}^j$。RNN_1 的输入为 $\mathbf{x}_{1:n}$，同时第 j 个循环神经网络 $\text{RNN}_j (2 \leq j \leq L)$ 的输入为前一个循环神经网络的隐藏状态 $\mathbf{s}_{1:n}^{j-1}$。一个三层循环神经网络的图形化表示见图 2.10。

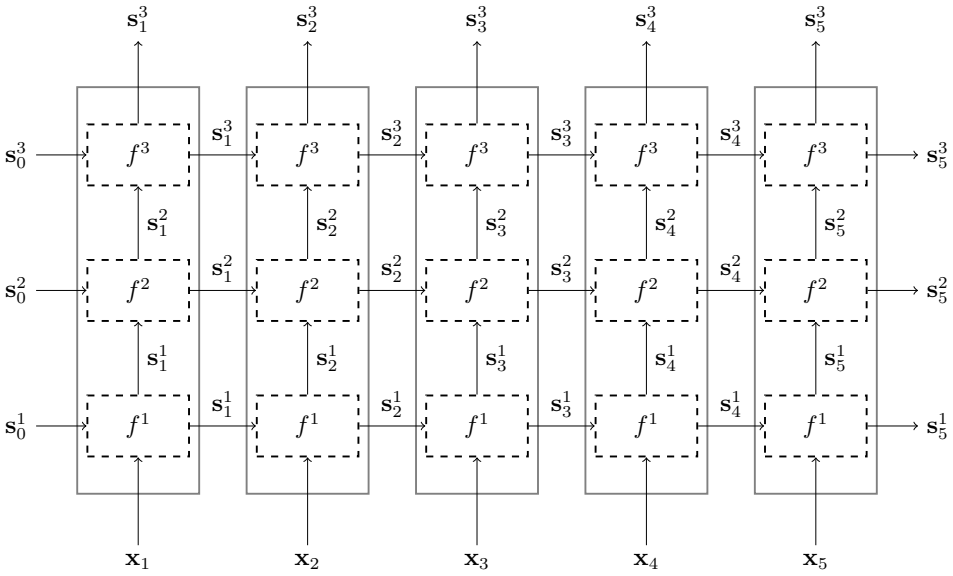

图 2.10 多层循环神经网络（展开形式，$L = 3$）

另外一个循环神经网络非常实用的扩展便是双向循环神经网络（Bidiectional RNN，Bi-RNN）[82, 164]。给定输入序列 $\mathbf{x}_{1:n}$，对于每个一个位置 i，双向循环神经网络计算保存了两个独立的隐层状态：前向隐层状态 $\overrightarrow{\mathbf{s}}_i$ 和后向隐层状态 $\overleftarrow{\mathbf{s}}_i$。$\overrightarrow{\mathbf{s}}_i$ 基于序列 $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_i$，而 $\overleftarrow{\mathbf{s}}_i$ 基于序列 $\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots \mathbf{x}_i$，前向和后向状态由两个不同的循环神经网络产生。对于每个位置 i 来说，最终的隐层状态 $\mathbf{s}_i$ 由前向和后向两

种状态组成，即 $\mathbf{s}_i = \overrightarrow{\mathbf{s}}_i \oplus \overleftarrow{\mathbf{s}}_i$，其中 $\oplus$ 表示向量拼接。对应的图形化表示见图 2.11。

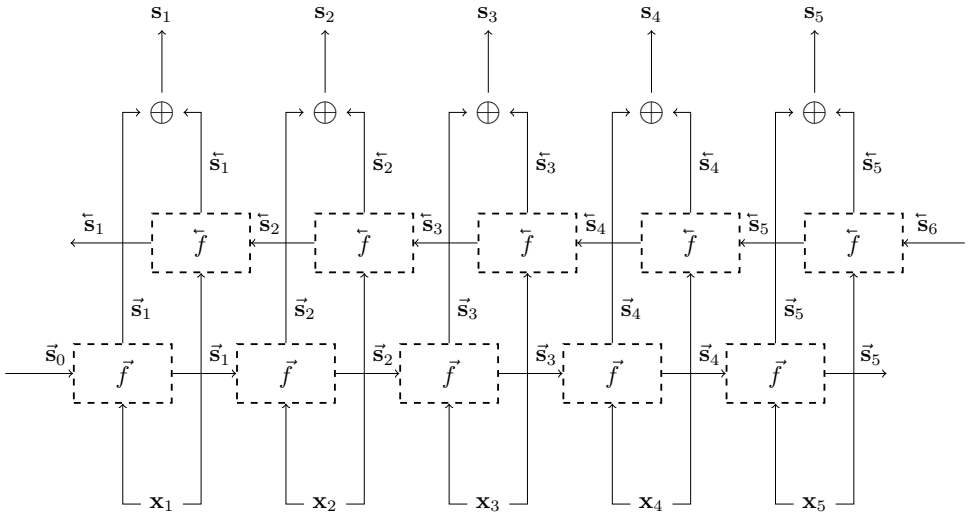

图 2.11 双向循环神经网络（展开形式，$n = 5$）

当使用不同的函数 f，就可以得到不同的循环神经网络架构。在这里简单介绍一下常用的三种不同的实例化循环神经网络：简单循环神经网络（Simple RNN），长短时记忆网络（Long Short-Term Memory，LSTM），门控循环单元网络（Gated Recurrent Unit，GRU）。接下来，我们用 d_x, d_s 表示向量的维度。

定义 2.1 (简单循环神经网络) 这是最简单的一种循环神经网络，又被称为 Elman 网络 [49]。在自然语言处理领域构建语言模型被使用过 [126]。具体的数学表达式如下：

$$\begin{aligned}\mathbf{s}_i &= f(\mathbf{s}_{i-1}, \mathbf{x}_i) \\ &= g(\mathbf{W}^x \mathbf{x}_i + \mathbf{W}^s \mathbf{s}_{i-1} + \mathbf{b})\end{aligned}$$

$$\mathbf{x}_i \in \mathbb{R}^{d_x}, \mathbf{W}^x \in \mathbb{R}^{d_s \times d_x}, \mathbf{W}^s \in \mathbb{R}^{d_s \times d_s}, \mathbf{b} \in \mathbb{R}^{d_s}, \mathbf{s}_i \in \mathbb{R}^{d_s}$$

其中函数 g 是非线性激活函数（可以是 sigmoid，tanh 或者 ReLU）。

定义 2.2 (长短时记忆网络) 在简单循环神经网络的基础上，引入门机制，可以缓解“灾难性遗忘”和“梯度爆炸与消失”问题，这被称为长短时记忆网络 [63]，在对句子建模时通常可以捕获一定的长距离依赖关系。具体的数学表达式如下：

$$\begin{aligned}\mathbf{s}_i &= \{\mathbf{c}_i, \mathbf{h}_i\} \\ \mathbf{s}_i &= f(\mathbf{s}_{i-1}, \mathbf{x}_i) \\ \mathbf{c}_i &= \mathbf{c}_{i-1} \odot \mathbf{f} + \mathbf{g} \odot \mathbf{i}\end{aligned}$$

$$
\begin{aligned}
\mathbf{h}_i &= \tanh(\mathbf{c}_i) \odot \mathbf{o} \\
\mathbf{i} &= \sigma(\mathbf{W}^{ix}\mathbf{x}_i + \mathbf{W}^{ih}\mathbf{h}_{i-1}) \\
\mathbf{f} &= \sigma(\mathbf{W}^{fx}\mathbf{x}_i + \mathbf{W}^{fh}\mathbf{h}_{i-1}) \\
\mathbf{o} &= \sigma(\mathbf{W}^{ox}\mathbf{x}_i + \mathbf{W}^{oh}\mathbf{h}_{i-1}) \\
\mathbf{g} &= \tanh(\mathbf{W}^{gx}\mathbf{x}_i + \mathbf{W}^{gh}\mathbf{h}_{i-1})
\end{aligned}
$$

$$
\mathbf{x}_i \in \mathbb{R}^{d_x}, \mathbf{s}_i \in \mathbb{R}^{2\cdot d_s}, \mathbf{c}_i, \mathbf{h}_i, \mathbf{i}, \mathbf{f}, \mathbf{o}, \mathbf{g} \in \mathbb{R}^{d_s}, \mathbf{W}^{*x} \in \mathbb{R}^{d_s \times d_x}, \mathbf{W}^{*h} \in \mathbb{R}^{d_s \times d_s}
$$

定义 2.3 (门控循环单元网络) 门控循环单元网络 [36] 是对长短时记忆网络的改进。相比于长短时记忆网络，使用更少的门，因此在计算效率上要优于长短时记忆网络。具体的数学表达式如下：

$$
\begin{aligned}
\mathbf{s}_i &= f(\mathbf{s}_{i-1}, \mathbf{x}_i) \\
\mathbf{s}_i &= (\mathbf{1} - \mathbf{z}) \odot \mathbf{s}_{i-1} + \mathbf{z} \odot \mathbf{h} \\
\mathbf{z} &= \sigma(\mathbf{W}^{zx}\mathbf{x}_i + \mathbf{W}^{zh}\mathbf{h}_{i-1}) \\
\mathbf{r} &= \sigma(\mathbf{W}^{rx}\mathbf{x}_i + \mathbf{W}^{rh}\mathbf{h}_{i-1}) \\
\mathbf{h} &= \tanh(\mathbf{W}^{hx}\mathbf{x}_i + \mathbf{W}^{hh}(\mathbf{h}_{i-1} \odot \mathbf{r}))
\end{aligned}
$$

$$
\mathbf{x}_i \in \mathbb{R}^{d_x}, \mathbf{s}_i \in \mathbb{R}^{d_s}, \mathbf{z}, \mathbf{r}, \mathbf{h} \in \mathbb{R}^{d_s}, \mathbf{W}^{*x} \in \mathbb{R}^{d_s \times d_x}, \mathbf{W}^{*h} \in \mathbb{R}^{d_s \times d_s}
$$

卷积神经网络

卷积神经网络（Convolutional Neural Netowrks，CNNs）[95] 最开始应用于计算机视觉领域，尤其是在图像分类上表现出很强的能力。后来被引入自然语言处理领域，在很多任务上也取得了不错的性能，比如语义角色标注（Semantic Role Labelling，SRL）[40] 和句子分类（Sentence Classification）[84]。在对文本用卷积神经网络进行建模时，通常不需要多次卷积和池化（Pooling），标准做法是使用一次卷积和一次池化即可得到句子的表示。对于文本任务，卷积背后主要的思想是在句子中的 k-词滑动窗口的每个实例上应用非线性（可学习的）函数。此函数（也被叫做**卷积核**）将 k 个单词的窗口转换为一个 d 维的向量。该向量可以捕获窗口中单词重要属性（每个维度有时也被称为一个**通道**）。然后，使用池化操作，将不同窗口产生的向量组合成单个 d 维向量，通常是取不同窗口每个通道中的最大值或者平均值，也叫做最大化池化（Max Pooling）和均值池化（Average Pooling）。池

化的目地主要是聚焦于句子中最重要的特征，不管位置如何。直觉上，当滑动窗口在序列上运行时，卷积核学习到的是识别 k-gram 信息。

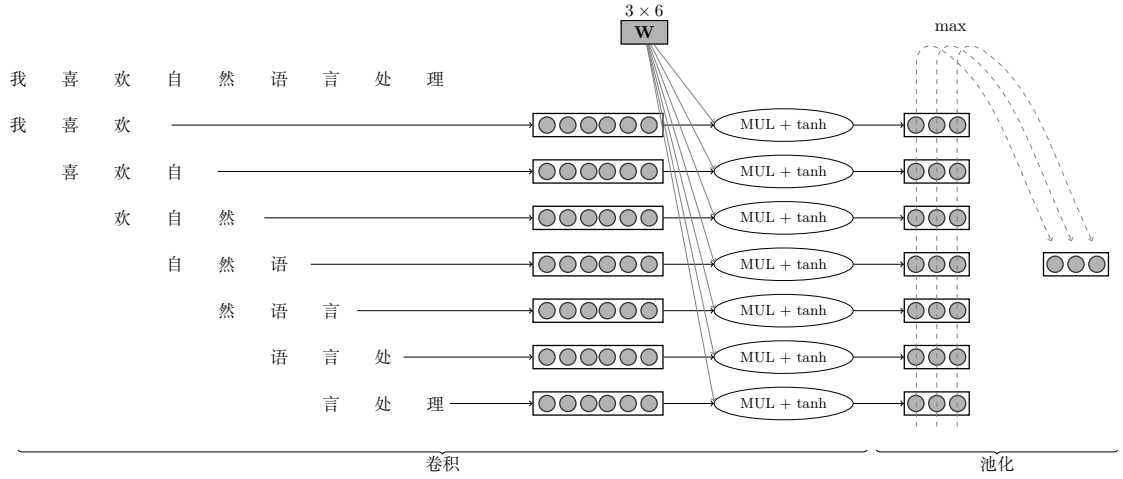

图 2.12 卷积神经网络（句子长度为 9，窄卷积，窗口大小为 3）

考虑一个单词序列 $\mathbf{x}_{1:n}$，$\mathbf{x} \in \mathbb{R}^{d_x}$ 表示词向量。宽度 k 的卷积层通过在句子上移动大小为 k 的滑动窗口来工作，并且在序列中的每个窗口 $\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+k-1}$ 应用相同的卷积核函数。卷积核函数通常是线性变换，后接一个非线性激活函数。假设第 i 个窗口的拼接向量为 $\mathbf{w}_i = \mathbf{x}_i \oplus \mathbf{x}_{i+1} \cdots \oplus \mathbf{x}_{i+k-1}$，则 $\mathbf{w}_i \in \mathbb{R}^{k \cdot d_x}$。根据是否在句子两边补充 $k-1$ 个单词，可以得到 $m = n-k+1$（窄卷积）以及 $m = n+k+1$（宽卷积）个窗口。卷积层的结果是 m 个向量 $\mathbf{c}_1, \ldots, \mathbf{c}_m$，其中 $\mathbf{c}_i \in \mathbb{R}^{d_c}$，

$$\mathbf{c}_i = g(\mathbf{W}\mathbf{w}_i + \mathbf{b}) \tag{2.17}$$

函数 g 是非线性激活函数，$\mathbf{W} \in {}^{d_c \times k \cdot d_x}$ 以及 $\mathbf{b} \in \mathbb{R}^{d_c}$ 是网络中待训练的参数。每一个 $\mathbf{c}_i$ 都是对 $\mathbf{w}_i$ 的进一步编码。理想情况下，每个维度捕获不同类型的指示信息。然后使用最大池化层得到一个 d_c 维向量 $\mathbf{c}$。

$$\mathbf{c} = \begin{pmatrix} c[1] \\ c[2] \\ \vdots \\ c[d_c] \end{pmatrix} = \max_{1<i\leq m} \mathbf{c}_i = \begin{pmatrix} \max_{1<i\leq m} \mathbf{c}_i[1] \\ \max_{1<i\leq m} \mathbf{c}_i[2] \\ \vdots \\ \max_{1<i\leq m} \mathbf{c}_i[d_c] \end{pmatrix} \tag{2.18}$$

$*[j]$ 表示向量 $*$ 的第 j 个分量，是一个标量值。最大池化操作的效果是捕获窗口位置最显著的信息。理想情况下，每个维度是针对型特定类型的预测特征，最大操作将选择每种类型最重要的预测特征。图 2.12 提供了该过程的一个示例。

由此产生的向量 $\mathbf{c}$ 是句子的一种表示，其中每个维度反映了与某些预测任务

相关的最显著的信息。然后将 **c** 输入下游网络层，最终形成用于预测的输出层。网络的训练过程计算任务的损失函数，梯度可以通过池化层、卷积层和输入层反向传播。

第三章　融合异构数据的实体关系抽取

本章探究利用知识库进行远程监督实体关系抽取的任务。利用知识库进行远程监督，其中非常重要的一个难点是噪音数据问题。为了解决这个问题，本文提出一个使用人工标注数据来帮助远程监督的实体关系抽取。通过设计两个共享任务（实体边界识别和二元关系检测），然后使用多任务学习框架来实现异构数据集的融合，同时对于共享任务和原始任务之间施加一致性约束。本文提出的方法能够以更鲁棒和一致的方式迁移高质量的人工标注数据的知识。在标准远程监督数据集 NYT 上的实验结果表明在联合实体关系抽取任务上本文的方法超过目前最好的系统。

3.1　引言

信息抽取的一个基本问题是从自由文本中识别实体和关系。给定一个句子，该任务旨在找出代表不同对象的字符串（比如人名（“Jobs”，“Obama”），组织机构（“Labour Party”））和实体之间的语义关系（比如人名和组织机构之间的 affiliation 关系）。由于实体关系抽取是将非结构化的自由文本转换为结构化知识的第一步，所以该任务在自然语言处理的研究和应用中有着重要的地位。

给定人工标注的数据集，完全有监督的模型（尤其是基于神经网络的模型）在抽取任务上取得了非常惊人的进步 [81, 128, 221, 228]。然而，限制这些方法的应用的主要因素是获得这些实体关系的高质量人工标注的成本。

为了获得开放领域的更多训练数据，文献 [127] 开始利用知识库的远程监督。具体地说，我们不需要手动标注实体和关系，可以通过将知识库中的三元组和自由文本进行对齐，从而自动生成训练数据。远程监督数据集的主要问题是噪声样本：对齐得到的关系在上下文中并不总是正确的。例如，（“Obama”，“United States”）在知识库中存在“born in”关系，但并不一定意味着所有出现这个实体对的句子都表达了这个“born in”关系。

在本章工作中，我们探究通过高质量人工标注的数据集来减轻远程监督实体关系抽取模型中的噪声训练样本的影响。将高质量的人工标注数据集的某些知识迁移到远程监督实体关系抽取中，从而使的远程监督模型更加准确和鲁棒。

这里的主要挑战是两个数据集通常是异构的：它们可能标注的实体类型和关系类型有差异，甚至标注的准则也不太一样。例如，图 3.1 列出了两个数据集

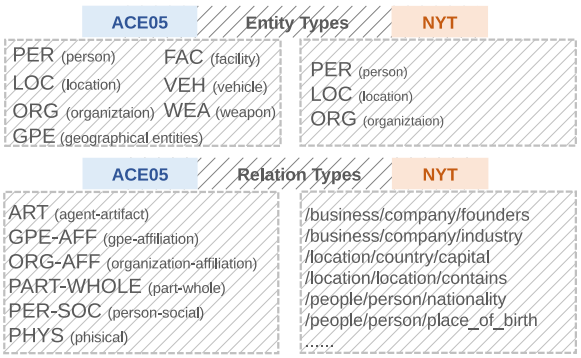

图 3.1 ACE05 和 NYT 数据集实体关系类型标注比较

(ACE05 和 NYT) 的实体关系类型标注。ACE05 是一个人工标注的数据集，包含 7 种实体类型和 6 种关系类型。NYT 是一个远程监督数据集（来自 Freebase），有 3 种实体类型和 24 种关系类型。解决这一挑战的直接方法是简单地合并两个数据集。但是，它忽略了两种标注策略之间的固有差异，并且很难为远程监督的模型带来性能提升。另一种方法是尝试手动建立两个模式之间的（近似）映射 [150]。但是，这种方式很不灵活，很难扩展到更多的标注策略。因此，如何平衡可扩展性和特异性，以及如何仔细控制融合过程是解决该挑战的关键因素。

本文提出了一种新颖的细粒度框架，将人工标注的数据整合到远程监督的学习过程中。具体来说，本文引入了两个共享任务来弥合两个异构标注类型之间的差异。

- 对于抽取实体，共享任务是定位实体的边界，即找出实体的开始和结束位置，忽略具体的实体类型。
- 对于抽取关系，共享任务是确定两个实体是否形成某种有效关系。

本文假设两个共享任务对不同的数据集不太敏感，并由此从人工标注的数据集中学习到关于实体关系的某些依赖于上下文的知识。为了进一步控制融合过程，本文还研究了共享任务与原始任务之间的一致性约束，利用约束关系设计了新的目标函数。

本文在 NYT 数据集上进行了大量实验，结果说明了本文提出的融合模型相对于最先进的系统实现了 3.7% 的提高。综上所述，本章的主要贡献是

- 本文首先探索了用人工标注的数据集来帮助远程监督实体关系抽取任务。
- 本文设计了一个新的融合框架，通过新的共享任务传递异构知识。
- 在 NYT 数据集上，本文提出的模型明显优于目前最先进的方法。

3.2 相关工作

从自由文本中抽取实体和关系得到了研究人员的广泛关注。目前最先进的系统通常采用的是有监督联合建模方法，这样可以缓解错误传播的问题并且加强了实体模型和关系模型之间的交互。基于特征工程的联合模型使用人工设计特征来同时执行实体和关系的抽取 [102, 129]。这些方法依赖于人工设计特征，这会带来额外的开销。为了克服这种缺点，一些基于神经网络模型的联合方法被提出，比如基于依存树的 biLSTM 模型 [128]，基于注意力机制的模型 [81]，序列标注模型 [228] 和全局归一化模型 [221]。除此之外，文献 [155] 通过联合建模实体关系交互，提出一个领域无关联合模型框架。文献 [188] 提出一个基于转移系统的模型来解决联合实体关系抽取任务。

另外一个相关工作的主线是多任务学习。多任务学习已被证明在许多自然语言处理任务中有效 [39, 71, 141]。基本方法是硬参数共享 [27]。文献 [174] 对于低层任务只共享低层网络的参数。文献 [32, 110] 引入对抗的共享-私有空间。但是，这些方法不会建模标签之间的关系。与此同时，文献 [6] 提出一个在不同标签空间上的多任务框架，可以学习到不同标签之间的转换。文献 [30, 140] 在相同任务上训练不同的标签任务。相比之下，本文的方法是运行上多个数据集上的多个任务。除此之外，文献 [75, 150] 在异构数据集上研究联合中文分词和词性标注任务。与他们的方法相比，本文学习到的共享表示更具有解释性。

3.3 融合异构数据的实体关系抽取框架

3.3.1 任务定义

给定一个输入句子 $s = w_1, \ldots, w_{|s|}$（w_i 是一个单词），本文研究的目标是从句子 s 中抽取一组实体集合 $\mathcal{E}$ 和一组关系集合 $\mathcal{R}$。一个实体 $e \in \mathcal{E}$，是具有某种实体类型的多个单词序列（比如人名（PER），组织机构（ORG））。用 $\mathcal{T}_e$ 表示可能的实体类型集合。一个关系是一个三元组 (e_1, e_2, l)，其中 e_1 和 e_2 是两个实体，l 是关系类型，描述了这两个实体之间的语义关系（比如组织从属关系（ORG-AFF））。用 $\mathcal{T}_r$ 表示可能的关系类型集合。

在本章工作中，本文研究使用高质量的异构数据集提高远程监督的信息抽取。具体来说，给定两个标注了实体和关系的训练数据集 D^a, D^b，假设 D^a 是大规模自动生成的数据，包含了大量的噪音样本，而 D^b 是少量的准确的人工标注的数据集。更进一步，D^a 和 D^b 具有不同的实体类型集合 $(\mathcal{T}_e^a, \mathcal{T}_e^b)$ 和关系类型集合 $(\mathcal{T}_r^a, \mathcal{T}_r^b)$，它们可能遵循不同的标注策略（异构）。本文的目标是为了提高在有噪音的数据 D^a

上训练的模型的性能。

3.3.2 基础模型

在描述本文提出的模型之前，我们首先介绍目前信息抽取系统中常用的两个模块：一个基于 biLSTM 的序列标注模型抽取实体 ($\mathcal{M}_{\text{seq}}$) 和一个基于 CNN 的多元关系分类模型抽取关系 ($\mathcal{M}_{\text{rel}}$)。

序列模型 $\mathcal{M}_{\text{seq}}$

为了抽取句子中的实体，本文采用 BILOU 标签方案，B，I，L 和 O 分别表示目标实体的开始、中间、最后和外面，U 表示单个词的片段。例如，为了抽取实体类型在 $\mathcal{T}_e$ 中的实体，对于每个单词 w_i，我们分配一个标签 t_i，其中 $t_i \in \{\texttt{B},\texttt{I},\texttt{L},\texttt{O},\texttt{U}\} \times \mathcal{T}_e$ [1] 编码了实体的片段信息和类型信息（比如（B，PER) 表示 PER 实体的开始）。

给定一个输入句子 s，序列标注模型使用 biLSTM（参数为 $\boldsymbol{\theta}$）尝试预测真实标签序列 $\mathbf{t} = t_1, t_2, \ldots, t_{|s|}$：

$$\mathbf{h}_i = \text{biLSTM}(\mathbf{x}_i; \boldsymbol{\theta}), \tag{3.1}$$

其中 $\mathbf{h}_i$ 是 biLSTM 的隐层状态向量（每个位置的前向 LSTM 和后向 LSTM 两个隐层状态向量拼接而成），$\mathbf{x}_i$ 是单词 w_i 的向量表示，是由预训练的词向量和基于字符级别的 CNN 输出向量拼接而成。然后，预测标签 $\hat{t}_i$ 的后验概率依照以下公式计算：

$$P_{\text{seq}}(\hat{t}_i|s) = \text{Softmax}(\mathbf{W}_e \mathbf{h}_i),$$

其中 $\mathbf{W}_e$ 是模型参数。序列模型的目标函数是最小化以下损失函数。

$$\mathcal{L}_{\text{seq}} = -\frac{1}{|s|}\sum_{i=1}^{|s|} \log P(\hat{t}_i = t_i|s). \tag{3.2}$$

关系模型 $\mathcal{M}_{\text{rel}}$

给定一个由序列标注模型 $\mathcal{M}_{\text{seq}}$ 抽取的实体对，本文使用多类分类器来确定它们形成某种类型的关系。例如，为了识别 $\mathcal{T}_r$ 中的关系，分类器对于该实体对输出标签 $l \in \{\texttt{NONE}\} \cup \mathcal{T}_r$，其中 NONE 表示不存在关系。

[1] 我们合并所有的（O, *）为一个标签 O，其中 $* \in \mathcal{T}_e$。

多元分类器首先利用多个 CNN 抽取实体对 (e_1, e_2) 的特征向量：

$$\mathbf{f}_{e_1,e_2} = \text{CNNs}(e_1, e_2, s; \boldsymbol{\omega}), \tag{3.3}$$

其中不同的 CNN 分别用来抽取两个实体的表示以及他们的上下文表示。具体来说，$\mathbf{f}_{e_1,e_2}$ 由 6 个向量拼接而成，$\mathbf{f}_{e_1}, \mathbf{f}_{e_2}, \mathbf{f}_{\text{middle}}, \mathbf{f}_{\text{left}}, \mathbf{f}_{\text{right}}$ 和 $\mathbf{f}_{\text{dist}}$。$\mathbf{f}_{e_1}$ 和 $\mathbf{f}_{e_2}$ 分别是两个运行在 e_1 和 e_2 上的 CNN 的输出向量。相似地，$\mathbf{f}_{\text{middle}}$ 是另外一个运行在 e_1 和 e_2 之间的 CNN 的输出向量。本文用 “LSTM-Minus” 方法 [221] 计算左侧特征向量 $\mathbf{f}_{\text{left}}$ 和右侧特征向量 $\mathbf{f}_{\text{right}}$。对于 $\mathbf{f}_{\text{dist}}$，本文使用 one-hot 特征向量来表示句子中 e_1 和 e_2 之间的距离特征。使用 CNN 时，每个单词的表示包含两个部分，一个是 biLSTM 的隐藏状态向量 $\mathbf{h}_i$ （与序列模型 $\mathcal{M}_{\text{seq}}$ 共享参数），另一个是 one-hot 向量表示的的实体标签。然后用单个隐层的多层感知机得到关系类型标签 $\hat{l}$ 的后验概率。

$$P_{\text{rel}}(\hat{l}|e_1, e_2, s) = \text{Softmax}(\mathbf{W}_{r_2}\text{ReLU}(\mathbf{W}_{r_1}\mathbf{f}_{e_1,e_2})), \tag{3.4}$$

最终关系模型的训练目标是最小化下面的损失函数。

$$\mathcal{L}_{\text{rel}} = -\sum_{(e_1,e_2)} \frac{\log P(\hat{l} = l|e_1, e_2, s; \boldsymbol{\omega})}{\#\text{ candidate pairs } (e_1, e_2)}, \tag{3.5}$$

其中正确标签 l 可以从标注中得到。$\mathbf{W}_{r_1}, \mathbf{W}_{r_2}$ 和 $\boldsymbol{\omega}$ 是模型的参数。

D^a 上的联合抽取模型

在两个模块的帮助下，本文能够在训练集 D^a 上构建一个简单的联合抽取系统 [2]。首先，本文用序列模型 $\mathcal{M}^a_{\text{seq}}$ 抽取实体（实体类型集合为 $\mathcal{T}^a_e$）。然后，对于每个抽取的实体对，本文用关系分类器 $\mathcal{M}^a_{\text{rel}}$ 识别关系（关系类型集合为 $\mathcal{T}^a_r$）。最后，为了联合训练模型，本文简单最小化两个模块损失函数的加和 $\mathcal{L}^a_{\text{seq}} + \mathcal{L}^a_{\text{rel}}$。

如果直接使用 D^a 样本进行训练，最终联合模型的效果比较差，因为训练数据中包含许多噪音数据。接下来，本文将尝试利用另外一种高质量但是异构的数据来提升联合模型的性能。我们期望少量的高质量数据的某些知识可以迁移到该联合模型中。

[2] 为了使符号清晰，本文将始终使用上标来表示不同的模型和参数。例如，$\mathcal{M}^a_{\text{seq}}, \mathcal{M}^b_{\text{seq}}$ 分别是在数据集 D^a 和 D^b 上训练的序列模型。$\mathbf{h}^a_i, \mathbf{h}^b_i$ 是它们的隐层状态。P^a, P^b 是它们的后验分布。

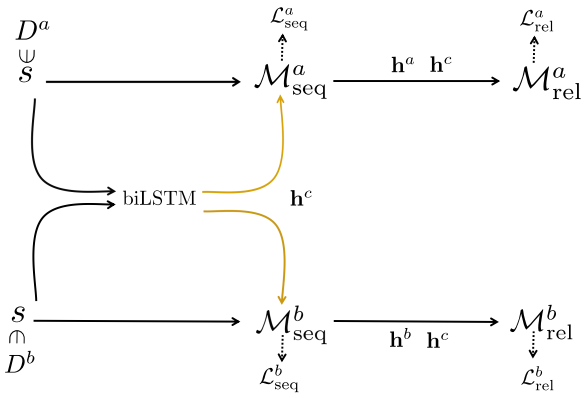

图 3.2 通过共享表示 $\mathbf{h}^c$ 融合

3.3.3 融合模型

虽然远程监督是获取大量数据的有效方法，但是自动生成 D^a 的缺点也很明显。例如，由于知识库三元组的启发式标注，一些标注关系是不正确的（false positive）。并且由于知识库的不完整性，一些真实关系可能没有标注（false negative）。

另一方面，存在许多人工标注的数据集，它们被用来构建各个领域的实体关系抽取系统。它们可能不遵循与 D^a 相同的标注准则，但是它们中的高质量标注可以提供一些领域无关的有用的知识。因此，研究是否可以通过调整这些异构数据集来提高 D^a 的模型性能是很有价值的。

基于合并数据集融合方式

本文第一个简单的尝试是合并两个数据集 D^a 和 D^b，然后用合并后的数据集训练联合模型 $(\mathcal{M}_{\text{seq}}, \mathcal{M}_{\text{rel}})$。除了实体类型集合 $(\mathcal{T}_e^a \cup \mathcal{T}_e^b)$ 和关系类型集合 $\mathcal{T}_r^a \cup \mathcal{T}_r^b$ 可能变大，其它的步骤与单独在 D^a 上训练的联合模型一样。

但是，根据我们的实验结果（在 D^a 上测试集），这个简单的解决方案无法提高性能。事实上，由于远程监督数据集 D^a 通常远大于人工标注数据集 D^b，直接混合两个训练集对于探索 D^b 中的样本是低效的。因此，我们可能需要设计更加精细的方法来控制融合过程。

基于共享表示的融合方式

我们可以不合并数据集，而是在 D^a 和 D^b 上保留两个模型的同时捕获两个模型之间的相互作用。具体来说，我们将融合两个数据集 D^a 和 D^b 分成两步。首先，我们尝试别两个数据集之间的共享信息，这些信息将作为抽取实体和关系的共有知识。其次，我们将共享和私有（数据集相关）信息融合在每个模型的预测中（图

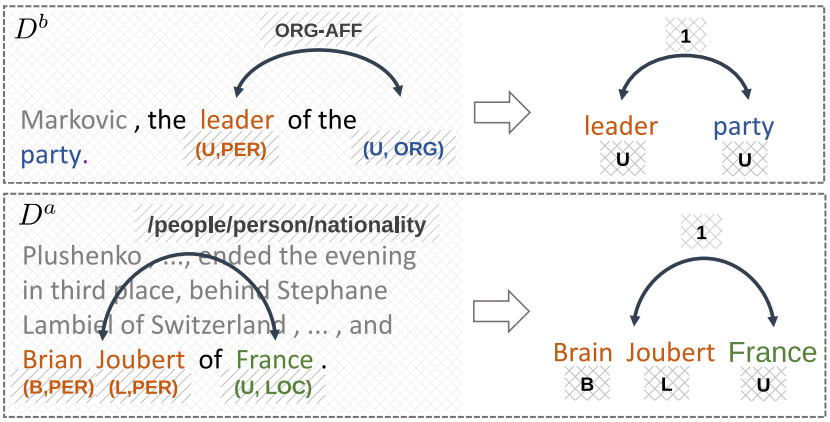

图 3.3 共享任务的不同人工标注以及转换版本样例

3.2)。

对于 D^a 中的句子 s，本文将它输入到联合抽取模型（序列模型 $\mathcal{M}^a_{\text{seq}}$ 和关系模型 $\mathcal{M}^a_{\text{rel}}$）。同样地，对于每个句子 $s \in D^b$，本文应用另外的联合模型 $\mathcal{M}^b_{\text{seq}}$ 和 $\mathcal{M}^b_{\text{rel}}$。对应的特征编码将会抽取在该数据集上的特征。

为了编码共享信息，本文引入一个新的 biLSTM，它可以接收来自 D^a 和 D^b 的句子。这个新的 biLSTM 的隐层状态输出记为 $\mathbf{h}^c_i$。在执行各自的数据集的预测之前，会将共享的隐层表示 $\mathbf{h}^c_i$ 与私有的隐层表示 $\mathbf{h}^a_i$ 或 $\mathbf{h}^b_i$ 做拼接，再做相应的预测。

$$P^a_{\text{seq}}(\hat{t}_i|s) = \text{Softmax}(\mathbf{W}^a_e(\mathbf{h}^a_i \oplus \mathbf{h}^c_i)), \tag{3.6}$$

$$P^b_{\text{seq}}(\hat{t}_i|s) = \text{Softmax}(\mathbf{W}^b_e(\mathbf{h}^b_i \oplus \mathbf{h}^c_i)). \tag{3.7}$$

关系模型 $\mathcal{M}^a_{\text{rel}}, \mathcal{M}^b_{\text{rel}}$ 够着的特征向量 $\mathbf{f}^a_{e1,e2}, \mathbf{f}^b_{e1,e2}$ 也使用了共享表示 $\mathbf{h}^c_i$。

基于共享表示的融合方式一个主要的假设是如果表示 $\mathbf{h}^c$ 同时有助于数据集 D^a 和 D^b，它可能会捕获它们之间的一些共享信息。因此，从 D^a 的角度看，本文不仅仅使用其私有特征表示，而且还通过共享表示来融合了一些来自 D^b 的知识。这种共享表示的主要问题是共享信息的含义不清晰，我们并不确定学习到了什么。另一方面，除了最终的模型性能之外，没有准则可以衡量学习到共享表示的好坏。因此，$\mathbf{h}^c$ 可能保留了不必要的信息以过度拟合训练集。

基于共享任务的融合方式

受共享和私有表示的分离的启发，我们可以在融合过程中增加更多控制，使得学习到的表示具有多的解释意义。尽管 D^a 和 D^b 具有不同的标注准则，但我们可以将原始任务分解，其中某些子任务是重叠的。例如，对于实体识别，我们可以先

找到它们的起始位置和结束位置，然后确定它们的实体类型。即使实体类型不同，有可能在两个数据集中开始和结束位置的分布更加接近。例如在图 3.3 中，“party”和“France”在 D^a 和 D^b 中被标注了不同实体类型，但是它们的起始位置都在介词“of”的右边。因此，假设它们共享实体边界的任务是合理的。

类似地，为了抽取关系，我们可以首先预测实体对是否形成有效关系，然后确定它们的关系类型。在图 3.3 中，实体(“leader”,“party”)和(“Brian Joubert”,“France”)在 D^a 和 D^b 中被标注了不同的关系类型，但它们都具有名词修饰的语法关系。这表明我们还可以添加一个共享任务来预测关系的存在。

本文引入了两个共享任务来帮助融合：实体边界检测和二元关系检测（图 3.4）。

- **实体边界检测**任务的目标是定位实体的边界。对于普通实体边界检测任务，本文采用序列标注模型 $\mathcal{M}^c_{\text{seq}}$。对于每个单词 w_i，$\mathcal{M}^c_{\text{seq}}$ 预测一个属于 $\{\texttt{B},\texttt{I},\texttt{L},\texttt{O},\texttt{U}\}$ 的标签，用来标记实体的边界（忽略特定的实体类型）。它同样共享 $\mathcal{M}^a_{\text{seq}}$ 和 $\mathcal{M}^b_{\text{seq}}$ 中的隐层向量 $\mathbf{h}^c$（公式 3.6 和 3.7）。更重要的是 $\mathcal{M}^c_{\text{seq}}$ 具有目标函数 $\mathcal{L}^c_{\text{seq}}$，它基于 D^a 和 D^b 中的实体边界标注。本文使用公式 3.2 通过将真实标注 $\mathbf{t}$ 转换为 BILOU 格式来计算 $\mathcal{L}^c_{\text{seq}}$。

- **二元关系检测**任务是预测实体对之间是否存在某种关系（忽略特定关系类型）。实体对可以通过 $\mathcal{M}^c_{\text{seq}}$ 模型的输出产生。同时可以利用隐层状态 $\mathbf{h}^c$ 作为关系模型的输入。本文使用关系模型 $\mathcal{M}^c_{\text{rel}}$ 抽取到候选实体对的表示 $\mathbf{f}^c_{e_1,e_2}$。它将为每个实体对分配 $\{0,1\}$ 中的标签，以表示是否存在关系，0 表示没有关系，1 表示存在关系。本文在 $\mathcal{M}^c_{\text{rel}}$ 的输出上添加一个损失函数 $\mathcal{L}^c_{\text{rel}}$ 来监督训练过程。$\mathcal{L}^c_{\text{rel}}$ 的计算遵循等式 3.5，其中二元真实标签可以从原始关系标签中转换得到。除此以外，我们还可以在私有关系模型中利用实体对 $\mathbf{f}^c_{e_1,e_2}$ 的表示。具体来说，为了预测最终的关系类型，可以将 $\mathbf{f}^a_{e_1,e_2}$ 和 $\mathbf{f}^c_{e_1,e_2}$ 向量表示进行拼接，这样在关系的预测中也可以用到共享的表示。

与共享表示相比，这里的主要区别是损失函数 $\mathcal{L}^c_{\text{seq}}$ 和 $\mathcal{L}^c_{\text{rel}}$ 依赖于共享任务。实际上，由于共享损失函数是在合并数据集 $D^a \cup D^b$ 上计算的，所以共享的表示可以在不同数据集之间迁移，同时基于此共享表示本文提出了两个共享任务，使得学习到的共享表示更具有解释性。

共享任务和原始任务的一致性约束

在共享任务和原始任务之间存在着某种一致性，本文设计目标函数来进行约束，进一步控制共享和私有模型之间的关系。

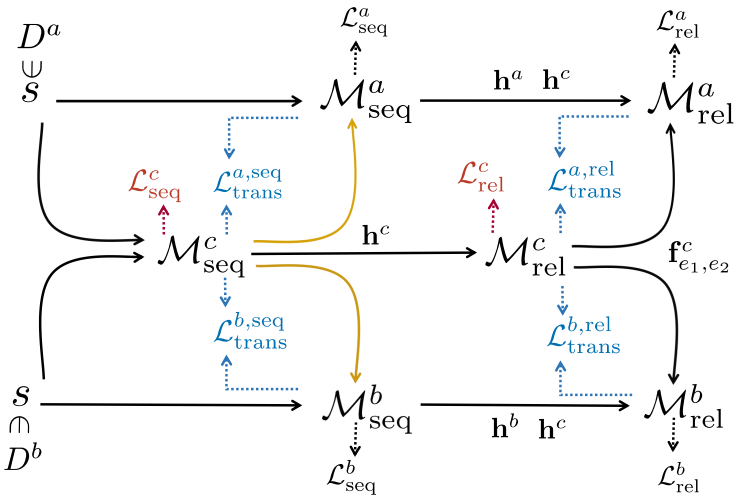

图 3.4 通过共享任务融合

对于实体模型，$\mathcal{M}^a_{\text{seq}}$ 输出一个在标签集合 $\{\texttt{B},\texttt{I},\texttt{L},\texttt{O},\texttt{U}\} \times \mathcal{T}^a_e$ 上的后验概率 $P^a_{\text{seq}}(\hat{t}_i|s)$，$\mathcal{M}^c_{\text{seq}}$ 输出一个在标签集合 $\{\texttt{B},\texttt{I},\texttt{L},\texttt{O},\texttt{U}\}$ 上的后验概率 $P^c_{\text{seq}}(\hat{t}_i|s)$。$P^a_{\text{seq}}$ 和 P^c_{seq} 一个很自然的约束就是边际约束。

$$P^c_{\text{seq}}(\star|s) = \sum_{*\in\mathcal{T}^a_e} P^a_{\text{seq}}((\star,*)|s), \quad \star \in \{\texttt{B},\texttt{I},\texttt{L},\texttt{O},\texttt{U}\}. \tag{3.8}$$

换句话说，给定句子 $s \in D^a$ 和一个位置 i，实体开始的标签概率（B）应该等于所有实体类型开始的标签概率（B, *）之和。P^b_{seq} 和 P^c_{seq} 之间也同样存在这样的一致性问题。

在这里本文使用软约束而不是使用硬约束来刻画一致性，具体来说，我们将最小化两个**转移损失函数** $\mathcal{L}^{a,\text{seq}}_{\text{trans}}, \mathcal{L}^{b,\text{seq}}_{\text{trans}}$。

$$\mathcal{L}^{\nabla,\text{seq}}_{\text{trans}} = \|P^c_{\text{seq}} - M^\nabla_{\text{seq}} P^\nabla_{\text{seq}}\|_2, \quad \nabla \in \{a,b\}. \tag{3.9}$$

其中 $M^a_{\text{seq}}, M^b_{\text{seq}}$ 是转移矩阵，目的是将概率 $P^a_{\text{seq}}, P^b_{\text{seq}}$ 转换为和概率 P^c_{seq} 相同的维度并且期望转换后的概率分布与 P^c_{seq} 分布一致。我们还可以为关系模型添加类似的转移损失函数。

$$\mathcal{L}^{\nabla,\text{rel}}_{\text{trans}} = \|P^c_{\text{rel}} - M^\nabla_{\text{rel}} P^\nabla_{\text{rel}}\|_2, \quad \nabla \in \{a,b\}. \tag{3.10}$$

融合模型最小化损失函数为

$$\sum_{\triangle\in\{a,b,c\}} \left(\mathcal{L}^\triangle_{\text{seq}} + \mathcal{L}^\triangle_{\text{rel}}\right) + \sum_{\nabla\in\{a,b\}} \left(\mathcal{L}^{\nabla,\text{seq}}_{\text{trans}} + \mathcal{L}^{\nabla,\text{rel}}_{\text{trans}}\right) \tag{3.11}$$

算法 3.1 训练过程

输入: D^a, D^b,
输出: 训练后的模型

随机初始化所有参数
while 模型未收敛 **do**
  for $\nabla \in \{a, b\}$ **do**
    从数据集 D^{∇} 中随机选择一批数据 B
    根据公式 3.2计算在批数据 B 上的损失函数 $\mathcal{L}_{\text{seq}}^{\nabla}$ 和 $\mathcal{L}_{\text{seq}}^{c}$
    根据公式 3.9 计算 $\mathcal{L}_{\text{trans}}^{\nabla,\text{seq}}$
    生成所有候选关系
    根据公式 3.5 计算 $\mathcal{L}_{\text{rel}}^{\nabla}$ 和 $\mathcal{L}_{\text{rel}}^{c}$
    根据公式 3.10 计算 $\mathcal{L}_{\text{trans}}^{\nabla,\text{rel}}$
    根据公式 3.12 计算最终的损失函数 $\mathcal{L}$
    通过最小化损失函数 $\mathcal{L}$ 更新参数
  end for
end while

总而言之，通过共享任务进行融合是一种比通过共享表示进行融合更精细的模型。它使得共享模型有清晰的解释，并且可以明确刻画共享和私有任务之间的联系。这些关于模型设计的先验知识可以使学习的模型更加稳定和鲁棒。

模型训练

为了训练联合模型，本文用交替的方式在两个数据集上优化目标函数（公式 3.11）。具体地，我们交替地从两个数据集 D^a 和 D^b 中随机选择批样本 B，然后批样本 B 上公式 3.11 可化简为

$$\sum_{\triangle \in \{\nabla, c\}} \left(\mathcal{L}_{\text{seq}}^{\triangle} + \mathcal{L}_{\text{rel}}^{\triangle}\right) + \left(\mathcal{L}_{\text{trans}}^{\nabla,\text{seq}} + \mathcal{L}_{\text{trans}}^{\nabla,\text{rel}}\right), \nabla \in \{a, b\}. \tag{3.12}$$

训练过程参考算法 3.1。本文在实体模型中采用调度采样策略 [12, 128] 本文使用 Adadelta [216] 优化模型，并采用梯度截断。整个网络使用 dropout 进行正则化。在一定数量的批样本迭代次数中，本文根据开发集合上的最佳关系性能选择模型。本文使用的预训练词向量是 100 维的 glove 向量 [142]。隐藏单元的维数为 128。对于本文网络中的所有 CNN，卷积核窗口大小为 2 和 3，输出通道数为 25。

表 3.1 NYT 数据集结果

Model	Relation		
	P	R	F
Gormley （2015）	55.3	15.4	24.0
Mintz （2009）	25.8	39.3	31.1
Tang （2015）	33.5	32.9	33.2
Hoffman （2011）	33.8	32.7	33.3
L&J （2014）	57.4	25.6	35.4
Ren （2017）	42.3	**51.1**	46.3
Zheng （2017）	61.5	41.4	49.5
Wang （2018）	64.3	42.1	50.9
Sun （2018）	67.4	42.0	51.7
Our Model	**70.4**	45.6	**55.4**
Sun （2018）（exactly match）	65.2	40.6	50.0
Our Model（exactly match）	68.3	44.2	53.7

3.4 实验

3.4.1 设置

本文在 NYT 远程监督数据集上评估我们提出的框架。本文使用 ACE05 数据集作为人工标注的数据集。有关 NYT 和 ACE05 数据集的详细情况请参考 2.3.3节。

本文评价模型结果用精确率（P），召回率（R）和 F 分数。具体地，如果输出实体 e 的类型和头部区域是正确的，则输出实体 e 是正确的，如果输出关系 r 的 e_1, e_2, l 是正确的，则输出关系 r 是正确的（精确匹配）。在以前的工作中，在计算关系的 F 分数时关系时不考虑实体类型 [155, 188, 228]。为了公平对比，本文也汇报此类结果。

在本章工作中，默认设置是具有实体转移损失的融合模型（表 3.2 中的第 6 行），它在 NYT 数据上取得了最优关系性能。

3.4.2 实验结果

首先，我们将方法与以前的工作进行比较（表 3.1）。第一部分是基于流水线方法，第二部分是基于联合抽取模型的方法，最后一部分是使用“完全匹配”评价方式的的联合抽取模型。总的来说，本文提出的方法在 F 分数方面比所有其他模型都有显著提高。特别地，比联合序列标注方法实现了 5.9% 的提高 [228]，并且与基于转移系统的联合模型相比，表现优于 4.5%。与现有最好的方法相比 [175]，取得了 3.7% 的提高。这表明本文的方法可以使用手动标记的高质量数据集提高远程监督的性能。

接下来，我们分析本文提出方法的各个组成部分的贡献和影响（表 3.2）。其中

- “only D^a” 是在 D^a 上训练的基础模型；
- “$D^a \cup D^b$” 表示通过合并数据集的融合模型；
- “only $\mathbf{h}^c$” 表示通过共享表示的融合模型；
- ‘$\mathbf{h}^c + \mathcal{L}^c$” 表示通过共享任务的融合模型；
- “$+ \mathcal{L}_{\text{trans}}$” 是具有实体关系转移损失的 “$\mathbf{h}^c + \mathcal{L}^c$” 模型；
- “$+ \mathcal{L}^{\text{seq}}_{\text{trans}}$” 和 “$+ \mathcal{L}^{\text{rel}}_{\text{trans}}$” 分别是只有实体和关系转移损失。

我们对此结果有一些结论如下：

1. “only D^a”（第一行）与当前当好的联合解码模型性能相当 [175, 188]。这说明基础模型 $\mathcal{M}_{\text{seq}}, \mathcal{M}_{\text{rel}}$ 是有效的。
2. “$D^a \cup D^b$”（第二行）和 “only D^a” 相比，关系抽取的性能比较差。我们认为由于 D^a 数据集远大于 D^b 数据集，直接混合两个数据集的效率很低。
3. 在通过共享表示（第 3 行）利用人工标注的 ACE05 数据集之后，实体识别和关系抽取的性能得到稍微的提升。这些观察结果表明，共享表示可以提高性能，是一种简单的融合模型方法。
4. 添加两个共享任务（第 4 行）后，实体识别和关系抽取的性能都有很大的改进（实体为 1.2%，关系为 2.7%）。它表明该模型可以通过合并数据集进行提高。与基于共享表示的融合方法相比，基于共享任务的融合方式是一种更加有效的方式。
5. 在实体模型和关系模型（第 5 行）上施加转移损失后，关系性能略有改善（0.1%），但实体性能下降。一个有趣的发现是，当我们只保留实体模型的转移损失时（第 6 行），它在很大程度上提升了关系性能。与此同时，当我们只保留关系模型的转移损失时（第 7 行），它无法改善关系绩效但实现了最优的实体性能。这些观察表明转移损失施加到实体模型或关系模型中可能会改变实体或关系的表现。一个可能的原因是实体模型和关系模型彼此密切相关，并且对一方施加限制将影响另一方。然而，它们如何在这种联合设置中相互影响仍然是个开放的问题。

第三，本文研究了人工标注数据集的数量的影响（表 3.3）。本文随机选择 ACE05 数据集的 25%，50%，75%。我们注意到随着数量的增加，表现并没有稳

表 3.2 不同设置下 NYT 数据集的结果

Model	Entity			Relation		
	P	R	F	P	R	F
only D^a	82.6	91.2	86.7	61.8	43.3	50.9
$D^a \cup D^b$	84.5	91.9	88.1	60.7	42.3	49.8
only $\mathbf{h}^c$	83.5	93.1	88.0	65.6	42.0	51.2
$\mathbf{h}^c + \mathcal{L}^c$	86.2	92.5	89.2	66.5	45.4	53.9
$+ \mathcal{L}_{\text{trans}}$	82.8	89.6	86.1	64.8	46.2	54.0
$+ \mathcal{L}^{\text{seq}}_{\text{trans}}$	86.6	92.9	89.6	70.4	**45.6**	**55.4**
$+ \mathcal{L}^{\text{rel}}_{\text{trans}}$	**87.2**	**93.9**	**90.4**	**72.9**	40.9	52.4

表 3.3 随着 ACE05 数据集数量不同 NYT 数据集的结果

Percentage of D^b	Entity			Relation		
	P	R	F	P	R	F
100%	**86.6**	**92.9**	**89.6**	**70.4**	45.6	**55.4**
75%	83.6	91.1	87.2	69.0	42.0	52.2
50%	85.9	92.7	89.2	62.8	**47.6**	54.2
25%	84.5	91.4	87.8	67.3	44.5	53.6

步增加。我们认为这是由随机抽样引起的。换句话说，ACE05 数据集的每个样本对性能的影响是不同的。例如，50%ACE05 数据集的结果具有高召回率，但 100%ACE05 数据集的结果具有高精度。如何有效地选择样本可能是一项有趣的未来工作。

第四，本文对转移矩阵进行可视化（用 M^a_{seq} 作为例子），如图 3.5所示。深色表示较大的权值。我们可以看到对角线颜色偏深，这意味这近似满足公式 3.8。例如，标签 B 的概率主要来自标签（B，PER）、（B，LOC）和（B，ORG）的转移。此外，与自动学习的转移矩阵相比，本文也尝试使用了固定值的转移矩阵，转移矩阵的值根据原始任务和共享任务之间的标签映射关系决定。具体来说，在图 3.5 中，对

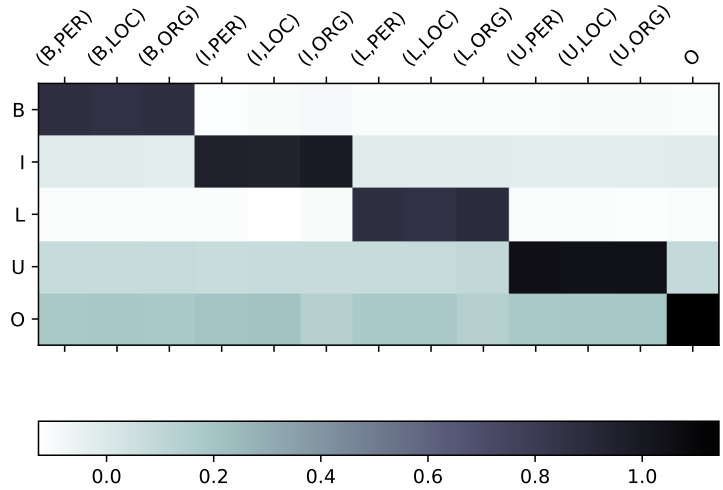

图 3.5 转移矩阵 M^a_{seq} 的可视化

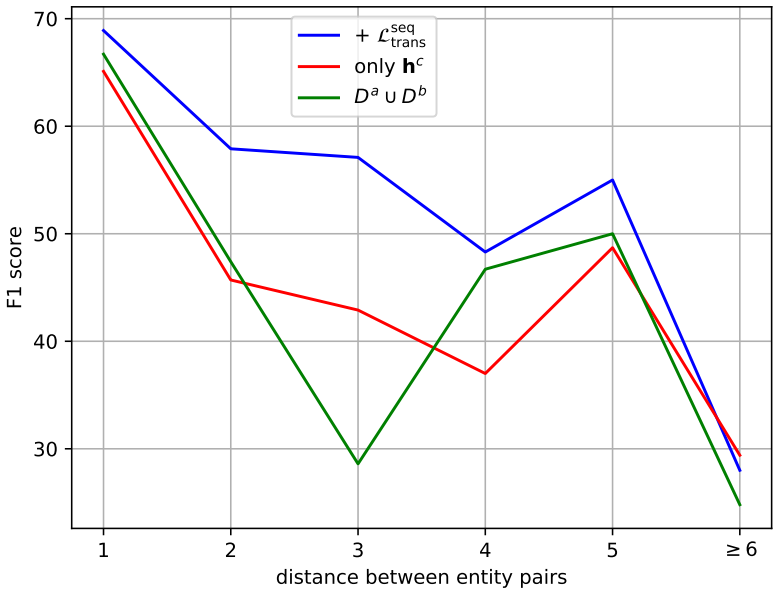

图 3.6 随着实体对距离不同 NYT 数据集的结果

角线的值设置为 1.0，而其他则设为 0.0。在这种情况下，获得 88.9% 实体 F 分数和 52.8% 关系 F 分数。这证明软约束对建模一致性的有效性。

第五，本文探究根据实体对之间的距离同模型的性能变化情况（图 3.6）。总体来说，当距离小于 6 时，本文的模型显着优于两个基准模型。此外，当距离大于 6 时，我们观察到所有模型的性能都非常低。因此，可以捕获长距离依赖性的联合解码算法可能是该联合抽取任务中的有希望的方向。

最后，在本章工作中，尽管目的是使用高质量的异构数据集来提高远程监督的关系抽取，本文也报告在 ACE05 数据集上的结果。本文使用和之前工作相同的数据集划分方式 [102, 128, 175]。本文在 ACE05 训练集上训练的基础模型达到 57.8% 的关系 F 分数。与这个结果相比，本文的系统（加上 NYT 数据）取得相近的性能。例如，本文最好的模型 “+ $\mathcal{L}_{\text{trans}}^{\text{seq}}$” 在测试集上达到 56.6 关系 F 分数。我们认为自动生成的数据会降低 ACE05 数据集的质量，因为它包含许多噪声样本并且覆盖率低。

3.4.3 案例分析

我们在一些具体的例子中将 “+ $\mathcal{L}_{\text{trans}}^{\text{seq}}$” 和 “only $\mathbf{h}^c$” 进行比较，结果见表格 3.4。♡ 表示正确标注。♣，♠ 分别是模型 “+ $\mathcal{L}_{\text{trans}}^{\text{seq}}$”，“only $\mathbf{h}^c$” 的输出。对于 S1，本文的模型识别了 “[chicago]$^{\text{LOC}}$” 和 “[sears tower]$^{\text{LOC}}$” 之间的 `contains` 关系。然而模型 “only $\mathbf{h}^c$” 没有找出这个关系，尽管实体均识别正确。对于 S2，模型 “+ $\mathcal{L}_{\text{trans}}^{\text{seq}}$” 没有识别出 `company` 关系。然而模型 “only $\mathbf{h}^c$” 准确找出来了。这两种情况说明本文

表 3.4 模型输出结果样例

S1	after the authorities described suspects talking about blowing up the [sears tower]$^{\texttt{ORG:}\heartsuit\clubsuit\spadesuit}_{\texttt{contains-2:}\heartsuit\clubsuit}$ in [chicago]$^{\texttt{LOC:}\heartsuit\clubsuit\spadesuit}_{\texttt{contains-1:}\heartsuit\clubsuit}$ and the f.b.i. 's [miami]$^{\texttt{LOC:}\heartsuit\clubsuit\spadesuit}$ headquarters .
S2	said [dennis rice]$^{\texttt{PER:}\heartsuit\clubsuit\spadesuit}_{\texttt{company-1:}\heartsuit\spadesuit}$, senior vice president of marketing for [disney]$^{\texttt{ORG:}\heartsuit\clubsuit\spadesuit}_{\texttt{company-2:}\heartsuit\spadesuit}$'s buena vista [pictures]$^{\texttt{ORG:}\clubsuit}$ unit .
S3	in [california]$^{\texttt{LOC:}\heartsuit\clubsuit\spadesuit}_{\texttt{contains-1:}\heartsuit\clubsuit\spadesuit\texttt{|contains-3:}\clubsuit}$, where parents first started educational foundations in response to a statewide law capping property taxes , the combined district of [santa monica]$^{\texttt{LOC:}\heartsuit\clubsuit\spadesuit}_{\texttt{contains-4:}\clubsuit\texttt{|contains-2:}\spadesuit}$ and [malibu]$^{\texttt{LOC:}\heartsuit\clubsuit\spadesuit}_{\texttt{contains-2:}\heartsuit\clubsuit}$ requires …

的模型可能比较擅长处理远距离的关系。对于 S3，模型 “+ $\mathcal{L}^{\text{seq}}_{\text{trans}}$” 错误地识别了 “[california]$^{\texttt{LOC}}$” 和 “[santa monica]$^{\texttt{LOC}}$” 的关系，尽管 “[california]$^{\texttt{LOC}}$” 和 “[malibu]$^{\texttt{LOC}}$” 已经被正确识别。我们认为在这种情况下，直接建立多个关系之间的依赖可能会进一步帮助提高模型性能。

3.5 总结

本文提出了一种新的融合框架，使用高质量的异构数据集来帮助提升远程监督的联合实体关系抽取。通过在共享抽取任务和原始抽取任务之间引入共享抽取任务和施加一致性约束，本文的框架以更加谨慎和可解释的方式控制融合过程。具体的，本文设计了两组共享任务，其目标函数可以作为共享表示的直接监督，从而使得学习到的表示更具有解释性。同时，对于原始任务和共享任务之间，本文使用边际约束刻画了两种任务之间的一致性。在 NYT 数据集的实验表明了所提方法的有效性。

第四章　基于语言学规则的远程监督实体关系抽取

前一章研究了利用少量高质量人工标注数据集来帮助远程监督实体关系抽取任务，提出了一种基于多任务学习的框架进行异质数据集之间的知识迁移。本章以倾向性关系抽取为例研究开放领域的实体关系抽取任务。给定大规模的无监督文本，本文提出一个高效的基于语言学规则和神经网络分类器的远程监督框架。首先人为设计一组领域无关的语言学规则，然后基于此规则自动生成训练集，并且使用神经网络来捕获词法和句法特征。本文设计的框架可以很快并且适用于大规模数据。在 Amazon 在线评论数据上测试了提出的框架，结果表明提出的模型在没有人工标注的前提下能够取得令人满意的结果。

4.1　引言

倾向性关系抽取系统旨在从文本中检测并抽取和情感相关的信息。在对用户生成的大规模自由文本使用某些自然语言处理算法后，研究者们可以更深入地了解人们如何表达对各种对象和主题的看法。该结果对于很多应用（比如推荐系统和检索）和语言学研究都很重要。

本章研究对象是倾向性关系抽取任务。该任务尝试识别评价词（表达情感的词语，情绪和评论，可以看成某种实体），评价对象（观点的目标，可以看成某种实体）和它们之间的关系（对某个目标的持有的观点，可以看成实体之间存在的关系）。以下是两个例子。

1. The unit is [well designed] and [perfect reception].
2. The Passion of The Christ will [touch your heart].

划横线的部分表示是一个评价对象，方括号的部分表示是一个评价词。这两句话中存在的倾向性关系是（“well designed”，“unit”），（“perfect reception”，“unit”），和（“touch your heart”，“The Passion of The Christ”）。抽取倾向性关系通常是对文本中情感进行细粒度分析的第一步，在其他与情感倾向性分析的应用（比如情感摘要）中起着关键作用。本章的目的是从开放领域的大规模文本中抽取倾向性关系。

以往的细粒度情感分析在很多方面取得了显著的成功：文献 [145] 在各种领域上做了尝试；文献 [51, 193] 在不同的关系类型上做了研究；许多有监督和无监督

（基于规则匹配）算法也都被提出来。但是我们在尝试使用现有的方法时发现了一些困难。

- 基于规则匹配的方法（包括词法规则和句法规则）在大规模数据中简单，快速并且易于扩展。然而在实际应用中，规则的健壮性通常是值得商榷的。例如句法规则对解析树中的错误很敏感，这种错误在用户生成的内容中很常见。词法规则的覆盖范围有限（比如使用固定的规则集合 [157]）以及很难控制噪音（比如基于 bootstrapping 的方法 [149]）。
- 当具有人工标注的数据时，有监督模型通常比基于规则的方法能获得更好的性能，但是通常很难获得大量的关系抽取标注，而且训练得到的模型往往受限于特定的领域。

因此，我们需要一种算法可以更好地结合词法和句法的特征，同时减少人工标注。由于人工标注的语料库成本高且在开放领域难以得到，本文更倾向于远程监督的分类器。另一方面，规则匹配可以产生一组倾向性关系而没有用到任何人工标注。尽管得到的这些关系并不完全正确，但它们易于收集并且很容易得到大规模数量。因此，我们可以把规则匹配得到的倾向性关系作为远程监督，希望可以得到更好的覆盖率并且胜过噪音带来的负面影响。

另一个问题是，许多现有的系统依赖于通用的情感词典来选择候选关系。如果一个评价词不能被词典所识别，那么系统是无法抽取相关关系。例如，许多现有的通用情感词典往往只是一个单词，缺乏对包含多个单词的评价词的支持。（例如，“more than what I expected”，“honest to the book”和“adrenaline pumping”），这些多个单词的表达在实际的文本中很常见。另外的一个例子是，由于词性标注器和句法解析器的错误，一些真正的评价词被忽略，比如把“a+”错误地标注为冠词。因此，为了扩大倾向性关系抽取结果的多样性，我们需要一些方法来更好地抽取评价词。

针对以上问题，本章的主要贡献在于以下几个方面：

1. 本文提出了一个基于远程监督的开放领域的倾向性关系抽取算法。该算法首先利用领域无关的规则得到一组倾向性关系，然后用得到的倾向性关系训练一个分类器。结果表明，尽管基于规则的倾向性关系不如人工标注的精确，但是基于远程监督训练的分类器仍然提高了性能。相比于当前无监督倾向性关系抽取系统双向传播算法 [149]，本文提出的算法明显优于它。
2. 本文设计了一个神经网络模型来抽取词法和句法的上下文信息。该模型利用 biLSTM 来抽取全局句子界别特征，利用卷积神经网络来获取局部低维特征

表示。与其它神经网络模型在关系抽取上的比较，本文受传统分类器的特征所启发，对于不同的上下文特征，使用不同的神经网络来抽取其表示，比如实体对的左边和中间本文使用两套卷积神经网络来抽取其特征。实验结果表明，本文提出的模型优于该任务中最先进的基于特征工程的逻辑回归模型分类器。

3. 本文探索了一个远程监督分类器来检测多词的评价词。给定一个候选词，分类器会根据相邻的单词然后预测它是否是一个评价词。新的分类器有助于我们发现不在通用情感词典中的评价词，从而有利于倾向性关系抽取。

本文的目标是使得算法简单，快速并且易扩展到大规模语料库。本文使用 Amazon 评论数据测试我们提出的系统。该数据集包含来自 15 个不同的领域的 3300 万条评论。本文提出系统输出的倾向性关系数据库包含 7250 万条倾向性关系。本文对该算法的各个方面进行了大量的实验验证，并且提出的远程监督模型的性能可以和之前的有监督模型相提并论。

4.2 相关工作

倾向性关系抽取是细粒度情感分析的重要任务。如果存在一定量的人工标注数据（比如 MPQA 语料库 [45]），我们可以将此任务作为有监督的关系抽取问题 [77, 87]。之前的工作可以分为两类：基于流水线的方法 [98, 192, 206] 和联合模型方法 [207, 208]。前者是首先抽取候选的评价词和评价对象，然后确定其存在的倾向性关系。后者是用一个统一的方式同时对评价词，评价对象，倾向性关系建模。使用有监督的方法的一个考虑因素是对领域和人工标注的依赖。

半监督和无监督的模型也被应用到倾向性关系抽取中。包括基于规则的 Bootstrapping 方法 [149]，图传播算法 [21, 109, 197]，整数规划算法 [117] 和概率主题模型 [132, 180]。

本文的模型受到之前远程监督算法的启发 [127, 172]。他们使用 WordNet 或知识库中的关系作为远程监督。由于没有类似可以用于倾向性关系抽取的资源，本文使用基于语言学规则匹配的方法来产生倾向性关系。另一方面，神经网络分类器在有监督的关系抽取中有着很好的性能，包括使用循环神经网络、卷积神经网络 [184, 196, 199, 219, 220]，序列模型和树模型 [100, 163]。文献 [128] 中的神经网络结构和我们的比较相似，他们使用两个 biLSTM 模型联合抽取实体和关系。另外一个最近的工作是文献 [72]，他们使用循环神经网络和卷积神经网络来解决倾向性关系分类任务。与他们模型不同的是，本文明确地学习不同词法和句法的特征表示。受到传统特征工程的启发，本文使用不同的神经网络分别抽取不同上下

文的特征。

另外一个联系比较密切的任务是基于角度（aspect）的观点挖掘 [186, 214, 227]。基于角度的观点挖掘并不找出评价词，而是直接分析对评价对象的情感极性。在 SemEval2015 和 SemEval2016 的比赛上也有此任务，并且针对此任务提出了各种解决方案 [145, 146]。与此任务相比，本文对评价对象并没有做任何限制，而且会抽取每个评价对象对应的评价词，所以本文的方法更适用于处理开放领域的倾向性关系抽取。

4.3 基于语言学规则的倾向性关系抽取框架

4.3.1 任务定义

给定一个输入句子 $s = w_1, w_2, \ldots, w_n$，其中 w_i 是一个单词，倾向性关系抽取任务是输出 (O, T) 对的集合，其中 $O = w_i, w_{i+1}, \ldots, w_j$ 是评价词，$T = w_k, w_{k+1}, \ldots, w_l$ 是评价对象，(O, T) 对是一个倾向性关系，说明评价词 O 直接指向目标 T [1]。O 和 T 都可能是包含多个单词。

4.3.2 语言学规则

根据以往的工作 [149] 可以知道句法规则对关系抽取任务是有效的。它们具有快速的优点，并且在跨领域上也通用，是开放领域大规模关系抽取任务的理想选择。尽管它们存在很多的优点，但是也有两个不足之处：一是由于文本中的噪音和语法错误，生成的解析树可能不可靠。二是规则的覆盖范围有限。为了解决第一个问题，本文致力于使用很严格的规则以保证输出的质量。针对第二个问题，本文使用一个远程监督分类器（4.3.3 节）和一个评价词分类器（4.3.4 节）来扩大覆盖范围。

表格 4.1 列出了本文提出系统中使用的语言学规则。当输入句子符合这些规则时，本文抽取相应的倾向性关系作为弱标注数据。表格 4.1 定义了几种词性标注集合：名词（NOUN）={NN，NNS}、动词（VERB）= {VB，VBD，VBN，VBP，VBZ}、形容词（ADJ）= {JJ，JJR，JJS}　和副词（ADV）= {RB，RBR，RBS}。w_i.pos 是指 w_i 的词性标注。w_i.np（w_i.vp, w_i.adjp）是指包含 w_i 的最小名词（动词，形容词）短语（如果没有短语则返回 w_i）。T, T' 是评价对象，O, O' 是评价词。“$(O', T') \rightarrow$”和“$\leftarrow (O', T')$”分别表示在 O' 和 T' 上的依存关系。如文献 [149] 中所述，利用输入句子的依存树制定相应的语言学规则，基本上捕获了（形容词-名词），（动词-补

[1] 本文假设 O, T 不会重叠，并且它们之间的距离 s 不超过某个阈值（在实验中取的是 10）。

表 4.1 语言学规则

Name	Pattern	Output	Example
P1	$w_1 \xrightarrow{\text{amod}}_{\text{dep}} w_2$ w_1.pos ∈ NOUN，w_2.pos ∈ ADJ	$T = w_1$.np $O = w_2$	It is a [cool] case. case $\xrightarrow{\text{amod}}$ cool
P2	$w_1 \xrightarrow{\text{acomp}}_{\text{xcomp}} w_2$ w_1.pos ∈ VERB，w_2.pos ∈ ADJ	$T = w_1$.vp $O = w_2$	The case looks [great]. looks $\xrightarrow{\text{xcomp}}$ great
P3	$w_1 \xrightarrow{\text{advmod}} w_2$ w_1.pos ∈ VERB，w_2.pos ∈ADV	$T = w_1$ $O = w_2$	The cover matches [perfectly]. matches $\xrightarrow{\text{advmod}}$ perfectly
P4	$w_1 \xrightarrow{\text{nsubj}} w_2$ w_1.pos ∈ NOUN， w_2.pos ∈ NOUN or VERB or ADJ has a coplua verb between w_1 and w_2	$T = w_1$.np ($O = w_2$.np $O = w_2$.vp $O = w_2$.adjp)	This case is [an excellent choice]. case $\xleftarrow{\text{nsubj}}$ choice
C1	$w_1 \xleftarrow{\text{conj}} (O', T')$ O'.pos. ∈ ADJ or ADV w_1.pos ∈ ADJ or ADV	$T = T'$ $O = w_1$	The case looks [great] and very [cute]. cute $\xleftarrow{\text{conj}}$ （great，looks）
C2	$(O', T') \xrightarrow{\text{nsubj}} w_1$ w_1.pos ∈ NOUN，T'.pos ∈ VERB O'.pos ∈ ADV，T' and O' are adjacent	$T = w_1$ $O = T'O'$	The case [fits perfectly]. （perfectly，fits）$\xrightarrow{\text{nsubj}}$ the case

语）和（副词-动词）关系。$w_1 \xrightarrow{l} w_2$ 表示单词 w_1 和单词 w_2 之间存在依存关系，且依存关系类型为 l。例如，如果在依存树中 w_1 是 w_2 的父亲，并且依赖类型是 amod 或者 dep，则表 4.1 中规则 P1 被激活。

为了克服依存树中的噪音和错误，本文对规则进行了限制，主要是从词性标注集合和通用情感词典 L 来实现。例如，规则 P1 中 w_1 只能是名词，w_2 只能是形容词，并且形容词必须出现在通用情感词典中。

本文还设计了规则能够处理评价词和评价对象有可能包含多个单词的情况（人工标注中大约存在 30% 的情况包含多个单词，而以往的工作通常忽略多词的情况）。具体而言：

- 当两个单词匹配一个规则时，我们将其扩展为包含这两个单词的最小短语。这有助于收集一些倾向性关系的局部上下文信息。例如规则 P4 中，单词 “case” 和单词 “choice” 分别扩展为 “the case” 和 “an excellent choice”。

- 可以将已有的倾向性关系组合成为一个新的评价词，这个新的评价词可能与其它的评价对象存在倾向性关系。例如在规则 C2 中，(“perfectly”, “fit”) 是一个倾向性关系，它可以构成新的评价词 “fit perfectly”，并且 (“fit perfectly”, “the case”) 又构成了新的一组倾向性关系，可以带来更多信息丰富的倾向性关系。

作为简单使用规则匹配的替代方案，本文还研究了基于 bootstrapping 的抽取算法 [149]。在此设置下，允许算法将新单词添加到通用情感词典，并使用更新的

情感词典进行接下来的规则匹配。虽然 bootstrapping 的方法可以发现原来不在通用情感词典中的评价词，但是我们发现由新添加的评价词引起的噪音关系难以控制，并且随着语料库的变大，bootstrapping 的优势反而被噪音所压制。在实验部分表明，相比于直接的语言学规则匹配，bootstrapping 方法精确率下降 30%。

4.3.3 远程监督

尽管基于规则匹配的方法精度很高，但是它的一个重要的缺点是覆盖率低。考虑以下的一个例子，

Ordered the k9ballistics Crate Pad and I am [so pleased].

表 4.1 中的规则没有适用于倾向性关系（“so pleased”，“Ordered the k9ballistics Crate Pad”），虽然这个关系可以从上下文中推断出来。事实上，这两个表达式距离很近，“Ordered the k9ballistics Crate Pad”是句子中唯一可能“please”的对象。为了进一步探索这些关系，本文设计了远程监督分类器用来整合各种词法和句法的上下文特征。实验结果表明，相比于基于规则的方法，分类器有助于提高 20% 的覆盖率。

具体来说，本文定义倾向性关系抽取任务：给定句子 s 中的一个候选关系 $x = (O, T)$，分类器输出概率 $p(y|x), y \in \{1, -1\}$ 表明 x 是一个正确倾向性关系的概率。

在训练阶段，我们将语言学规则匹配得到的倾向性关系作为正样本。对于每个正样本 (O, T)，根据 $T' \neq T$ （T' 是 NP，VP 或 ADJP）构造负样本 (O, T')。同时对于 (O', T) (其中 $O' \neq O$) 的所有实例也加入负样本。在测试阶段，我们考虑句子 s 中所有的 VP 和 ADJP 作为候选评价词，并且要求候选评价词中必须包括某个出现在通用情感词典中的单词。我们考虑所有的 NP 和 VP 作为候选评价对象。最后所有的候选评价词和候选评价对象的两两组合对作为候选关系。

本文设计了一个基于神经网络的远程监督分类器。与大多数先前的深度学习模型不同，我们受到传统特征工程的启发，设计多个神经网络分别捕获特定的上下文特征。我们从实验中观察到，以前的特征工程工作中的知识可以帮助神经网络模型获得更好的性能。在具体描述之前，本文先定义一些符号表示。对于句子 s 中的 $x = (O, T)$，其中 $O = w_i, \ldots, w_j, T = w_k, \ldots, w_l$。其它部分表示为 $L = w_1, \ldots, w_{i-1}, B = w_{i+1}, \ldots, w_{k-1}, R = w_{l+1}, \ldots, w_n$。$L$, R 分别是 x 的左右部分上下文，B 是 O 和 T 之间的上下文[2]。$D = w_{p1}, \ldots, w_{pm}$ 表示 O 和 T 在依存树上的路径。

[2]为了简单起见，假设 O 出现在 T 之前。在代码实现中，一个指示维度被用来表示 O 是否在 T 之前。

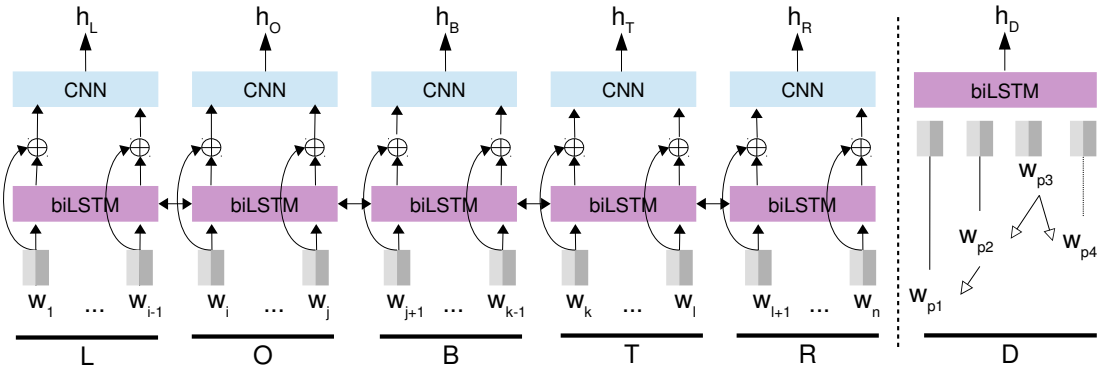

图 4.1 词法和句法上下文的表示学习过程

本文使用 biLSTM 对句子进行编码，从而使得每个词的表示融入了整个句子信息。因为 biLSTM 的循环结构和记忆门机制可以传播并共享句子 s 的长距离特征。为了进一步捕获 x 的词法上下文信息，本文使用 5 个 CNN 分别学习 L, O, B, T, R 的表示。以 B 为例，对输入进行卷积加池化后得到输出 $\mathbf{h}_B \in \mathbf{R}^d$。CNN 的输入包括 $w_{i+1}, \ldots, w_{k-1}$ 的词向量、词性标注向量以及 biLSTM 的输出隐层向量。所有表示 $\mathbf{h}_L, \mathbf{h}_O, \mathbf{h}_B, \mathbf{h}_T, \mathbf{h}_R$ 在我们的实验中使用相同的卷积神经网络结构。最后，为了对 y 进行预测，我们使用 softmax 函数 $p(y|x) = \frac{1}{Z}\exp\{\theta^\intercal \Phi(x, y)\}$ 在加权特征向量 $\Phi(x, y)$ 上，

$$\Phi(x, y) = a_L \mathbf{h}_L + a_O \mathbf{h}_O + a_B \mathbf{h}_B + a_T \mathbf{h}_T + a_R \mathbf{h}_R,$$

其中 $a_L, a_O, a_B, a_T, a_R \in \mathbf{R}, \theta \in \mathbf{R}^d$ 是模型参数。

为了更好地融入依存树信息，本文还尝试用文献 [196, 199] 的方式将依存路径 D 合并到模型中。具体地，我们在依存路径上的单词序列使用另外一个 liLSTM 进行编码，从而得到融合依存路径信息的表示 $\mathbf{h}_D \in \mathbf{R}^d$。然而，实验表明 $\mathbf{h}_D$ 并没有更进一步提升性能。我们猜测可能是依存解析错误限制了这个特征的贡献。

4.3.4 评价词分类器

在上述的语言学规则匹配和远程监督算法中，如果候选评价词中包含出现在通用情感词典 L 中的单词，则抽取该候选评价词。同时也提出两种简单的规则可以处理包含多个单词的评价词，但是也存在很多其他的多词评价词被忽略，比如有的多词评价词中并没有出现 L 中有的单词，或者词性标注结果出现错误从而不满足制定的语言学规则。在本节中，为了处理更多的多个词的评价词，本文引入远程监督的评价词分类器，它可以根据上下文信息预测候选短语是否是评价词。

对于句子 s 中的一个候选表达式 $O = w_i, \ldots, w_j$，本文用其上下文词 $w_{i_c}, \ldots, w_{i-1}$

和 $w_{j+1}, \ldots, w_{j+c}$ 作为一个 CNN 的输入（c 是上下文窗口大小，在实验中我们设为 5）。在经过卷积和池化层之后接 softmax 函数（与 4.3.3 节使用 CNN 相似），分类器输出一个概率 $p(z|O), z \in \{-1, 1\}$ 表示 O 是一个有效评价词的可能性。为了获取训练集，我们依赖于通用情感词典 L。给定无标签的句子并且 $w \in L$，我们认为句子中出现 w 的是正样本，其它随机选择的单词作为负样本。

为了在倾向性关系分类器中应用评价词分类器，本文对句子中所有的候选短语使用该评价词分类器，如果候选短语的概率 $p(z|O)$ 大于某个阈值 γ，那么我们就把该短语作为候选评价词。

4.4 实验

4.4.1 设置

本文在文献 [121] 提供的 Amazon 产品在线评论语料库中抽取倾向性关系，其中包含 15 个不同的领域 3300 万条评论。抽取得到的倾向性关系统计数据见表 4.2，这些关系是通过规则匹配（表 4.1）的方法得到的。所有的数字都是以 10^6 作为单位。

对于定量评价，本文选择四个领域（Cell Phones，Movie and TV，Food，Pet Supplies）进行详细分析。我们手动标注了 1000 个句子的所有倾向性关系，并选择 200 个作为验证，其余 800 作为测试 [3]。此外，为了和之前的监督方法进行比较，我们同样在 USAGE 语料库 [86] 进行实验，该语料库标注了 8 种产品总共有 4481 个倾向性关系。

本文使用 NLTK [16] 进行分词和分句，使用 Standard 解析器 [4] 获取词性标注，短语和依存树，使用 scikit-learn 工具包 [139] 和 TensorFlow [5] 用于机器学习算法。通用情感词典来源于文献 [190]。

4.4.2 主要结果

表 4.4 列出了 4 个领域的结果。为了评价本文提出的模型，我们采用的几个基准系统分别是

- **Adjacent** 是文献 [66] 中提出的一个简单的基准系统。它首先识别通用情感词典中的单词，然后将最近的名词或动词短语作为其评价对象。

[3] https://github.com/AntNLP/OpinionRelationCorpus
[4] http://nlp.stanford.edu/software/lex-parser.shtml
[5] https://www.tensorflow.org/

表 4.2 倾向性关系数据库统计情况

Domain	#Reviews	#Sents	#Relations
Cell Phones	3.4	19.1	6.7
Movie and TV	4.6	47.6	15.3
Food	1.3	7.6	2.5
Pet Supplies	1.2	8.0	2.4
Automotive	1.4	9.5	2.6
Digital Music	0.8	7.3	2.4
Beauty	2.2	6.4	3.8
Toys and Games	2.3	11.7	3.8
Instruments	0.5	13.5	1.4
Office Products	1.2	4.1	2.6
Patio	1.0	8.3	2.0
Baby	0.9	6.5	1.8
Clothing	5.7	29.1	10.2
Sports	3.3	20.2	7.1
Kindle	3.2	25.0	7.9
All	33	223.9	72.5

表 4.3 逻辑回归中使用的特征列表

Lexical Features
① POS tag sequences of O and T.
② The length of O and T.
③ The distance between O and T.
④ The word sequence between O and T in s.
⑤ POS tags of words between O and T in s.
⑥ Words, POS tags of w_{i-1}, w_{i-2}, w_{j+1}, w_{j+2}.
⑦ Words, POS tags of w_{k-1}, w_{k-2}, w_{l+1}, w_{l+2}.
⑧ Combined POS tags of O and T.
Syntactic Features
⑨ Does a dependency relation exist between O and T.
⑩ The dependency path between O and T.
⑪ The length of the dependency path.

- **Bootstrapping** 重现了双向传播算法 [149]，这是倾向性关系抽取目前最先进的无监督算法。它同样使用一组语言学规则，但是将规则发现的新的评价词添加到通用情感词典中，然后更新后的词典将用于下一轮的迭代中。

- **Pattern** 是指语言学规则匹配的方法（4.3.2 节）。

- **LR** 是逻辑回归模型，在与4.3.3 节相同的远程监督条件下进行训练。这些特征来源于最先进的远程监督关系分类器 [127]。表 4.3 列出了逻辑回归模型中使用的特征。

- **NN** 是本文提出的神经网络模型（4.3.3 节）。输出维度 d 设为 128（比如 $\mathbf{h}_B$），biLSTM 的输出维度同样是 128，词向量和词性标注向量的维度是 300。使用 3

种不同窗口的卷积核，窗口大小分别是 (1, 2, 3)。使用预训练好的 word2vec [6] 初始化词向量。默认情况下，我们在句子级别 biLSTM 上使用 5 个 CNN，而不包括依存路径和评价词分类器 ("NN" 相当于表 4.4 中的 "biLSTM + LOBTR")。为了得到训练集，本文在 6×10^4 的无监督句子上进行规则匹配。

- **NN+Pattern** 是结合 "NN" 和 "Pattern" 的结果（取并集）。

表 4.4 提出系统和多个基准系统的结果对比

System	Phone			Movie			Food			Pet		
	P	R	F	P	R	F	P	R	F	P	R	F
Adjacent	38.6	65.7	48.6	30.0	58.8	39.7	31.4	46.5	37.5	28.4	62.2	39.0
Bootstrapping	44.0	62.9	51.8	26.9	49.3	34.8	43.6	54.0	48.2	33.6	57.7	42.5
Pattern	**69.4***	64.4	61.1	**62.2***	42.4	50.4	**76.0***	41.9	54.1	**59.9***	51.3	55.3*
LR	60.1	64.7	62.4	55.6	57.0	55.3	65.5	49.2*	56.2	47.6	**59.3***	52.8
NN	63.4	67.9*	65.6*	56.8	58.2*	57.5*	70.5	47.7	56.9*	51.4	58.0	54.5
NN+Pattern	64.4	**70.5**	**67.3**	58.2	**59.9**	**59.1**	68.4	**50.8**	**58.3**	54.9	58.2	**56.5**

根据表格 4.4，我们可以得到以下几个结论。

1. "Adjacent" 的方法效果比较差，意味着对于这个任务需要利用一些先进的语言学特征。

2. "Bootstrapping" 在 4 个领域的性能都低于 "Pattern"。我们对 "Bootstrapping" 的结果进行了检查，发现新添加的评价词会给通用情感词典带来很多噪音，这些噪音影响了该方法的性能。

3. 虽然 "Pattern" 的方法具有很高的精确率，但是远程监督分类器（"LR" 和 "NN"）有助于提高召回率并且得到更好的 F1 分数（除了 Pet 领域）。因此，基于对规则的远程监督方法训练得到的分类器可以涵盖更多正确的倾向性关系。对于 Pet 领域，"Pattern" 的方法精确率很低，因此 "LR" 和 "NN" 的训练集中的错误样本比较多，而且可能无法学习到可靠的模型。

4. 神经网络模型 "NN" 在所有领域都优于传统的分类器 "LR"，这表明神经网络自动学习特征表示比特征工程设计特征上更具一些优势。

5. 简单的合并 "Pattern" 和 "NN" 的结果可以提高最后的 F1 分数。

接下来，本文测试使用不同特征神经网络模型的结果（表 4.5），可以得到以下几个结论。

1. 从第一行到第三行，我们比较了不同设置下 CNN 的模型性能。"biLSTM+B" 的 CNN 只用 B 作为输入（使用 O 和 T 之间的单词）；"biLSTM+OBT" 使

[6] https://code.google.com/archive/p/word2vec/

表 4.5 不同设置下的模型性能

System	Phone			Movie			Food			Pet		
	P	R	F	P	R	F	P	R	F	P	R	F
biLSTM +B	59.8	**69.5**	64.3	54.7	**59.1**	56.8	64.8	**48.5**	55.5	47.4	57.2	51.8
biLSTM +OBT	63.4	66.7	65.3	59.8	58.2	58.9	68.6	46.8	55.6	56.2	56.9	56.5
biLSTM +LOBTR	63.4	67.9	65.6	56.8	58.2	57.5	66.5	47.7	55.5	51.4	**58.0**	54.5
biLSTM +LOBTR+D	65.5	61.3	63.4	59.5	55.8	57.6	69.1	46.8	55.8	54.9	57.7	56.3
LOBTR	64.0	65.3	64.7	57.5	56.7	57.1	**70.6**	43.1	53.5	54.3	57.7	55.9
$\gamma = 0.5$	64.2	64.7	64.5	56.2	57.9	57.0	65.5	45.5	53.7	**61.6**	54.3	**57.7**
$\gamma = 0.8$	**65.6**	66.1	**65.9**	**61.1**	58.2	**59.6**	67.1	48.2	**56.1**	58.6	50.8	54.4

用 3 个 CNN 分别将 B, O, T 作为输入；“biLSTM+LOBTR”使用 5 个 CNN（等价与表 4.4 中的“NN”）。我们可以看到随着使用更多的 CNN，性能是不断提升（特别是召回率）。但是，依存路径上的特征并没有更多的提高性能（“biLSTM+LOBTR+D”）。

2. 在第 5 行，我们去掉句子级别的 biLSTM 只使用 5 个 CNN，发现与“biLSTM+LOBTR”相比损失了一些性能。因此，biLSTM 捕获长距离依赖的信息对该任务是有益的。

3. 我们在最后两行测试了评价词分类器的影响。γ 是分类器输出概率的阈值。当 $\gamma = 0.8$ 时，分类器添加的新的评价词可以提高性能，但是当 $\gamma = 0.5$ 时，噪音的影响已经大过其收益。我们进一步展示 $\gamma = 0.8$ 时，“NN”和“LR”方法的 precision-recall 曲线，见图 4.2。评价词分类器可以找到一些非常有趣的评价词，比如“not even enough”，“became extremely hot”，“*just* enough”，“arrived damaged”，“looks cool n cute”，表明了该分类器既发现通用情感词典中没有的评价词，也可以对噪音有一定的容忍度。

4.4.3 USAGE 语料库结果

为了和有监督的方法进行比较，本文也在 USAGE 语料库上评价了提出的模型，结果见表 4.6。为了构建远程监督模型，本文使用 USAGE 的 8 个产品在 Amazon 数据集中的无标注评论。比较的基准系统是来自文献 [72, 86]，是目前在该数据集上表现最好的系统。

表 4.6结果说明，在具有相同的设置时（“gold”）[7]，本文远程监督的模型与之间有监督的模型相比，可以获得差不多的精确率。通过查看具体的输出样例，我们发现造成性能差距的一个原因可能是 USAGE 和我们的数据集之间的标注准则上的差异。例如，我们不会把代词标注为评价对象，然而 USAGE 会将带此标注为

[7]这两个系统均没有汇报端到端的结果。在文献 [72] 中，评价词的 F1 分数是 50%，评价对象的 F1 分数是 67%。

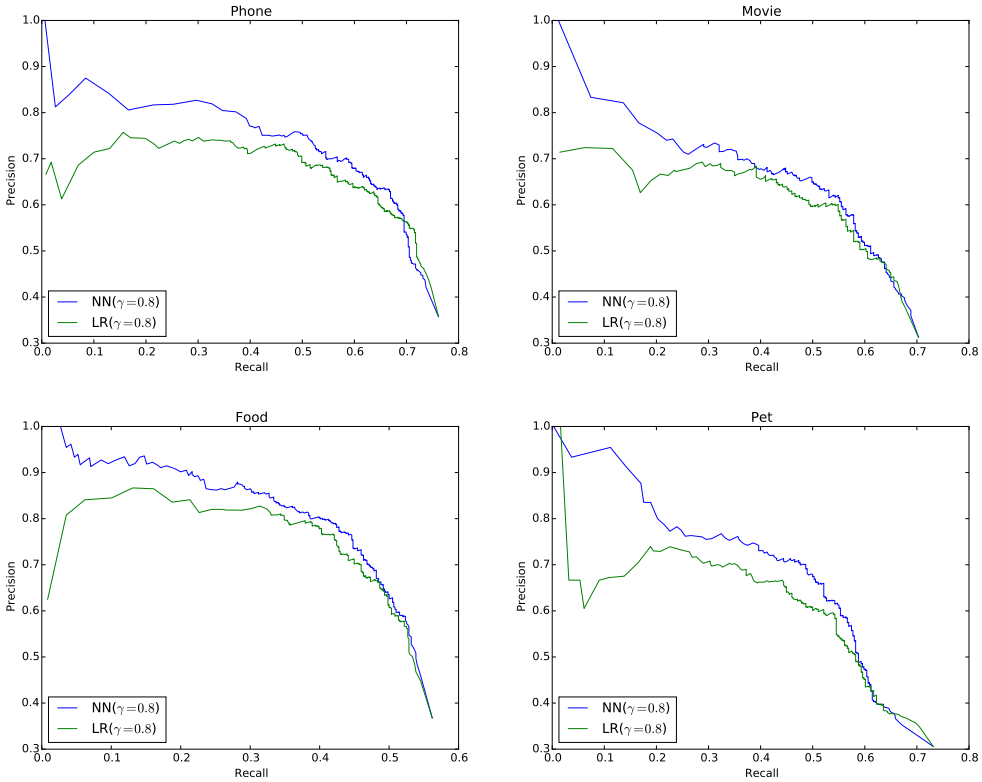

图 4.2 Precision-recall 曲线（$\gamma = 0.8$）.

表 4.6 USAGE 语料库上的结果

Systems	P	R	F
Klinger et al. [86]	-	-	65.0
Jebbara et al. [72]	87.0	75.0	81.0
Pattern	51.4	20.7	29.5
LR （end-to-end）	49.5	27.8	35.6
NN （end-to-end）	43.3	40.1	41.6
LR （gold）	89.1	47.9	62.3
NN （gold）	81.4	62.8	70.9

评价对象。比如 USAGE 中存在（“love”，“it”）这样的标注。我们还可以观察到，远程监督分类器要明显优于基于语言学规则的系统。

4.4.4 错误分析

最后，本文对抽取的倾向性关系进行一些错误分析。以电影领域为例，对于电影评论，我们发现评论者的评论与电影的情节和人物混在一起，这使得我们很难将背景主题与评论者的评论区分开来。例如，“Also said repeatedly how Tojo was [loyal to Emperor Hirohito]”，“loyal”这个词没有表示倾向性关系，因为它是对情节的描述。评论者的真实意图是“repeatedly”这个词，这里很难用本文定义的倾向性关系来表达。我们计划在未来的工作中引入更多背景知识和更完善的关系类

型来解决此种情况。

4.5 总结

本章研究了大规模开放领域的倾向性关系抽取任务。本文提出的算法首先基于人工制定领域无关的语言学规则来获得一组倾向性关系，然后基于这些得到的倾向性关系训练一个神经网络分类器。为了找出通常情感词典中没有的评价词，本文还提出一个评价词分类器用来发现更多的评价词。与远程监督实体关系抽取相比，本文主要探究了当没有现成的知识库时，利用语言学规则进行远程监督，以减少大量的人工标注成本。在 Amazon 在线评论数据集上的实验结果证明了所提方法的有效性。

第五章　基于风险最小化训练方法的联合实体关系抽取

前两章主要从“数据”角度对实体关系抽取任务进行探究，而本章工作从“联合模型”的角度对实体关系抽取任务进行探究。不同于前人的工作，本文提出一种新的轻量级的基于最小风险训练方法的联合学习方式。本文提出的算法优化了一种全局损失函数，并且可以灵活有效地探索实体模型和关系模型之间的联系。本文实现了一个简单并且强大的神经网络，并在其基础上使用风险最小化训练方法。ACE05 和 NYT 数据集上的实验结果表明，在联合抽取任务上，本文提出的模型可以取得最好的结果。

5.1　引言

检测实体和关系通常是从纯文本中抽取结构化知识的第一步。其目标是识别有类型对象（实体）的文本片段以及这些文本片段之间的语义关系（关系）。例如，在下面的句子中，

[Associated Press]$_{\texttt{ORG}}$ [writer]$_{\texttt{PER}}$ [Patrick McDowell]$_{\texttt{PER}}$ in [Kuwait City]$_{\texttt{GPE}}$.

“Associate Press”是一个组织实体（ORG），“writer”是一个人名实体（PER），这两个实体之间具有从属关系（ORG-AFF）。

解决实体关系抽取的方法主要有两种类型：基于流水线的模型和联合模型。在基于流水线的设置中，任务被分解为独立的子模型（实体模型和关系模型）。这种方法具有灵活性，比如对于实体模型来说，可以使用任意的实体数据来进行训练，同样对于关系模型也是。但此种方法忽略了两个模型之间的交互。例如，实体模型利用不到对于识别实体有用的关系标注，例如，如果两个实体之间存在 ORG-AFF 关系，则第一个实体必定是 ORG 类型，第二个实体必定是 AFF 类型，而基于流水线的方法很难捕获此种信息。相比于流水线模型，联合抽取模型在统一的框架下抽取实体和关系，可以利用共享信息并减轻模型之间的错误传播。在本章工作中，本文将关注联合模型的方法解决实体关系抽取任务。

一种简单的联合学习方法是共享参数 [81, 128]。通常，实体和关系模块不是训练两个独立的模型，而是可以共享一些输入特征或者中间隐层状态。这种方法的优点是两个子模型可以独立的抽取特征，不需要对特征的类型做任何的限制。

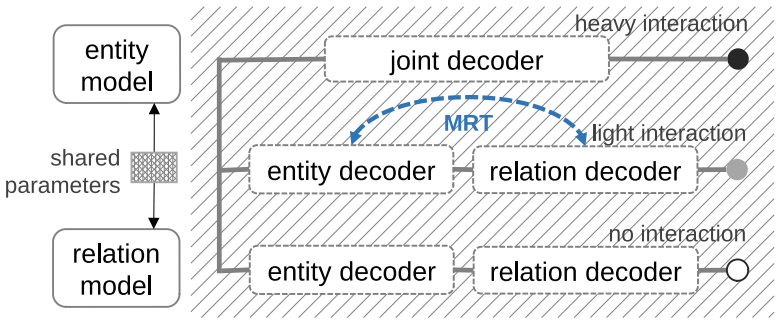

图 5.1 联合实体关系抽取的方式

基于共享参数的联合模型通常使用各自的子模型解码，并没有使用联合解码，子模型的输出结果之间某些联系不能够被完全利用。例如，实体模型要从关系标注中获取某些信息，需要等待关系模型更新共享的参数。为了进一步增强子模型解码器之间的交互，一些复杂的联合解码算法也被应用在联合实体关系抽取任务中[80, 102, 221, 228]。在这种情况下，需要在联合解码算法的精确性和各个子模型的容量之间取得适当的平衡。这种平衡通常会很关键，会直接影响最终的模型性能。

在本章工作中，本文提出了实体关系抽取的联合风险最小化训练（Minimum Risk Training，MRT）[137, 171] 方法。它提供了一种轻量的方案来加强实体模型和关系模型之间的交互，同时并不影响实体模型和关系模型特征的丰富性。给定输入 $\mathbf{x}$ 和损失函数 $\Delta(\hat{\mathbf{y}}, \mathbf{y})$，该损失函数衡量了模型输出 $\hat{\mathbf{y}}$ 与真实标注 $\mathbf{y}$ 之间的差异。MRT 目标是找出一个后验概率分布 $P(\hat{\mathbf{y}}|\mathbf{x})$，满足期望损失 $\mathbf{E}_{\hat{\mathbf{y}}\sim P(\hat{\mathbf{y}}|\mathbf{x})}\Delta(\hat{\mathbf{y}}, \mathbf{y})$ 最小。与之前的联合解码算法相比，基于 MRT 的算法很容易应用在实体关系抽取模型上，无需改变原始子模型和其解码器（图 5.1）。

本文提出的联合 MRT 的优点如下：

- 基于 MRT 的方法的一个优点是它可以明确地优化全局句子级别损失函数，例如 F 分数。通常的最大似然估计都是优化局部的损失函数，比如在序列标注中词级别的交叉熵损失函数。MRT 可以在训练过程中捕获更多的句子级信息，并在测试阶段更好地匹配评估指标。
- 除了手工指定的损失函数之外，本文还尝试在联合 MRT 过程中直接从数据中学习损失函数。自动学习的损失函数将有助于 MRT 探索特定数据集的优化目标，因为人为制定的全局损失并不一定是最优的损失函数，我们期望自动学习的损失可以在一定程度上对人为制定的损失函数进行微调，从而得到更好的性能。

之前的 MRT 通常使用于单任务，比如机器翻译 [60, 169]。在联合抽取任务中，实体模型和关系模型有着各自的损失函数，如何利用关系模型的损失函数来提高

实体模型的性能，或者反过来如何利用实体模型的损失函数来提高关系模型的性能，是联合 MRT 的一个关键问题。

本文提出一个基于神经网络的模型，可以使用联合 MRT 进行训练。该模型在实体模型中使用递归神经网络（RNN），在关系模型中使用卷积神经网络（CNN）。在 ACE05 和 NYT 数据集上，我们表明新的 RNN+CNN 结构优于以前的基于神经网络的模型。加上联合 MRT 后，本文的模型能够取得最好的性能。

综上所述，本文的主要贡献包括

- 对于联合实体关系抽取任务，提出一种基于风险最小化训练联合学习方式。
- 实现一个强大而简单的基础神经网络用来抽取实体和关系，并且可以在模型上使用所提出的 MRT 算法。
- 在两个数据集（ACE05 和 NYT）上取得最好的结果。

5.2 相关工作

给定一个句子，基于流水线方法的实体关系抽取系统首先用一个实体模型识别句子中的实体，然后基于这些实体用关系模型 [29, 107, 130] 判断实体之间的存在的关系。这种方法易于结合不同的数据源和不同的学习算法，具有很大的灵活性。但是，它们也可能遭受错误传播和数据效率低下的困扰。为了解决这个问题，最近的很多研究试图开发联合抽取算法。

实现联合学习方法的简单策略是参数共享。例如，在文献 [128] 中，实体模型是基于句子级别 RNN 序列标注模型，关系模型使用 RNN 在依存树路径上抽取特征并预测关系类型，并且将实体模型的隐藏状态作为特征（即共享参数）。本文的基础抽取模型与他们的类似，区别在于关系模型利用 CNN 作为特征抽取器。文献 [81] 使用注意力机制实现关系模型。

为了进一步探索实体解码器和关系解码器之间的交互，研究人员提出了一些联合解码算法。例如，文献 [80] 提出了一种基于 CRF 的模型，在每个位置不仅预测实体的标签，同时预测两边的单词距离，表示与该距离的单词具有某种关系，最后利用增强的转移矩阵刻画实体标签、预测距离相邻位置的转移依赖。文献 [228] 直接将关系标签编码到序列标注输出标签中. 它们都是精确的解码算法，但是在关系模型的建模上做了许多假设，例如文献 [228] 中的联合解码算法不能处理出现在多个关系中的实体。另一方面，文献 [102] 提出了一种基于集束搜索的联合解码算法。文献 [221] 研究全局归一化的联合模型。它们使用的特征是不受限制的，并且

没有过多的假设，但是解码算法是不精确的。在这里本文引入 MRT，这是一个更轻量级的联合学习方式。

风险最小化训练是一种学习框架，它可以处理具有任意评价指标的模型（即模型输出和真实答案的损失）[52, 137, 171]，已成功应用于许多 NLP 任务。文献 [60, 169] 将 MRT 应用于（神经）机器翻译；文献 [198] 提出了一种直接优化 F 分数的移进归约 CCG 解析器；文献 [7] 使用基于 MRT 的模型进行文本摘要任务。这些使用 MRT 的大多数工作都集中于单个任务，而联合实体关系抽取包括两个子任务。在联合学习场景中探究 MRT 如何应用是本章工作的研究重点。

最后，求解 MRT 的采样算法类似于强化学习（RL）中的策略梯度算法 [177]。最近的很多 NLP 工作使用了 MRT 的思想，但使用的是 RL 语言描述，并且得到了不错的结果，例如对话系统 [101] 和机器翻译 [135]。从数据中学习损失函数的想法是受逆强化学习 [1, 153] 的启发。

5.3 基于风险最小化训练方法的联合实体关系抽取系统

5.3.1 任务定义

本文定义联合实体关系抽取任务基于之前的工作 [128]。给定输入的句子 $s = w_1, \ldots, w_{|s|}$（w_i 是一个单词），目标是抽取一组实体集合 $\mathcal{E}$ 和一组关系集合 $\mathcal{R}$。一个实体 $e \in \mathcal{E}$ 是带有实体类型的单词序列（比如人名（PER）、组织机构（ORG））。$\mathcal{T}_e$ 表示可能的实体类型集合。一个关系是一个三元组 (e_1, e_2, l)，e_1 和 e_2 是两个实体，l 是两个实体之间的语义关系（比如组织从属关系（ORG-AFF））。$\mathcal{T}_r$ 表示可能的实体类型集合。

在本文的联合抽取方法（图 5.2）中，本文将实体检测视为序列标记任务（5.3.2 节），将关系检测视为分类任务（5.3.3 节）。这两个任务的模型共享参数并联合训练。与以前的联合学习算法不同 [81, 128, 221]，本文对联合抽取模型引入了最小风险训练。它优化了全局损失函数，弥合了训练和测试过程之间的差异（5.3.4 节）。

5.3.2 实体识别

为了表示句子 s 中的实体，本文采用 BILOU 标签方案为每一个单词 w_i 分配一个标签 t_i：t_i 的取值在集合 $\{(\mathtt{B}, *), (\mathtt{I}, *), (\mathtt{L}, *), \mathtt{O}, (\mathtt{U}, *)\}$ 中，其中 B，I，L 和 O 表示实体的开始、内部、结尾和外部。U 表示单个单词的实体。$* \in \mathcal{T}_e$ 表示不同的实体类型。例如，对于人名实体 (PER) “Patrick McDowell”，我们将 $(\mathtt{B}, \mathtt{PER})$ 分配给 “Patrick”，将 $(\mathtt{L}, \mathtt{PER})$ 分配给 “McDowell”。给定输入句子 s，实体模型通过从真实

标注 $\mathbf{t} = t_1, \ldots, t_{|s|}$ 学习预测标签 $\hat{\mathbf{t}} = \hat{t}_1, \ldots, \hat{t}_{|s|}$。

本文使用 biLSTM[63] + softmax 来处理序列标注任务。在句子的每个位置 i，前向 LSTM 从 s 的开始收集到当前位置 i 的信息，得到向量表示 $\overrightarrow{\mathbf{h}}_i$。类似地，后向 LSTM 从 s 的结尾收集到当前位置 i 的信息。

$$\overrightarrow{\mathbf{h}}_i = \text{LSTM}(\mathbf{x}_i, \overrightarrow{\mathbf{h}}_{i-1}; \overrightarrow{\boldsymbol{\theta}}),\ \overleftarrow{\mathbf{h}}_i = \text{LSTM}(\mathbf{x}_i, \overleftarrow{\mathbf{h}}_{i+1}; \overleftarrow{\boldsymbol{\theta}}).$$

$\mathbf{x}_i$ 是 w_i 的单词表示，包含两个部分。$\mathbf{x}_i = \mathbf{w}_i \oplus \mathbf{c}_i$（$\oplus$ 是向量拼接操作），其中 $\mathbf{w}_i$ 是词向量（来自于词向量表 $\mathbf{W}_e$）。$\mathbf{c}_i$ 是 w_i 的基于字符的表示，它是通过卷积神经网络得到，卷积神经网络的输入是单词 w_i 对应的字符序列，即 $\mathbf{c}_i = \text{CNN}(\text{char}(w_i); \boldsymbol{\theta}_c)$。

为了预测标签 $\hat{t}_i$，本文将前向和后向隐层向量拼接 $\mathbf{h}_i = \overrightarrow{\mathbf{h}}_i \oplus \overleftarrow{\mathbf{h}}_i$，并在 $\mathbf{h}_i$ 上使用 softmax 函数以得到 $\hat{t}_i$ 的后验分布。

$$P_{\text{ent}}(\hat{t}_i | s; \boldsymbol{\theta}_E) = \text{Softmax}(\mathbf{W}_E \cdot \mathbf{h}_i), \tag{5.1}$$

其中 $\boldsymbol{\theta}_E = \{\mathbf{W}_e, \boldsymbol{\theta}_c, \overrightarrow{\boldsymbol{\theta}}, \overleftarrow{\boldsymbol{\theta}}, \mathbf{W}_E\}$ 是实体模型的参数。给定输入句子 s 及真实标注序列 $\mathbf{t}$，实体识别的训练目标是最小化 $\mathcal{L}_{\text{ent}}$。[1]

$$\mathcal{L}_{\text{ent}}(\boldsymbol{\theta}_E) = -\frac{1}{|s|} \sum_{i=1}^{|s|} \log P_{\text{ent}}(\hat{t}_i = t_i | s; \boldsymbol{\theta}_E).$$

5.3.3 关系抽取

给定一个句子，从实体模型中可以得到实体的预测标签序列 $\hat{\mathbf{t}}$，然后可以得到该句子预测的实体集合 $\hat{\mathcal{E}}$。我们将 $\hat{\mathcal{E}}$ 中所有实体对视为候选关系。关系检测的目的是为每个实体对预测关系类型 $l \in \mathcal{T}_r$，[2] 并输出关系集合 $\hat{\mathcal{R}} = \{(e_1, e_2, l) | e_1, e_2 \in \hat{\mathcal{E}}, e_1 \neq e_2, l \in \mathcal{T}_r\}$。为了构建关系模型，本文抽取两种类型的特征，即关于实体 e_1, e_2 中的单词的特征和关于实体对 (e_1, e_2) 的上下文的特征。

- 为了在 e_1, e_2 中抽取单词的特征，本文使用两个卷积神经网络。以 e_1 为例，对于 e_1 中的每个单词，本文首先利用实体模型中 w_i 的 biLSTM 隐层向量。然后，本文将该隐层向量与第 i 位置的识别标签 one-hot 向量拼接起来。本文通过在序列向量 $\{\mathbf{h}_i \oplus \mathbf{v}_i | w_i \in e_1\}$ 上运行 CNN（卷积层后接最大池化层）为 e_1 构建特征向量 $\mathbf{f}_{e_1}$。用同样的方式，本文用另外一个 CNN 可以得到 $\mathbf{f}_{e_2}$。
- 对于实体对 (e_1, e_2) 的上下文特征，本文通过抽取 e_1 和 e_2 之间的单词特征

[1] 本文还尝试了 biLSTM-CRF [68] 模型作为实体模型，但是从实验结果看并没有获得性能上的提升。
[2] 我们添加 NONE 关系类型在集合 $\mathcal{T}_r$ 中。

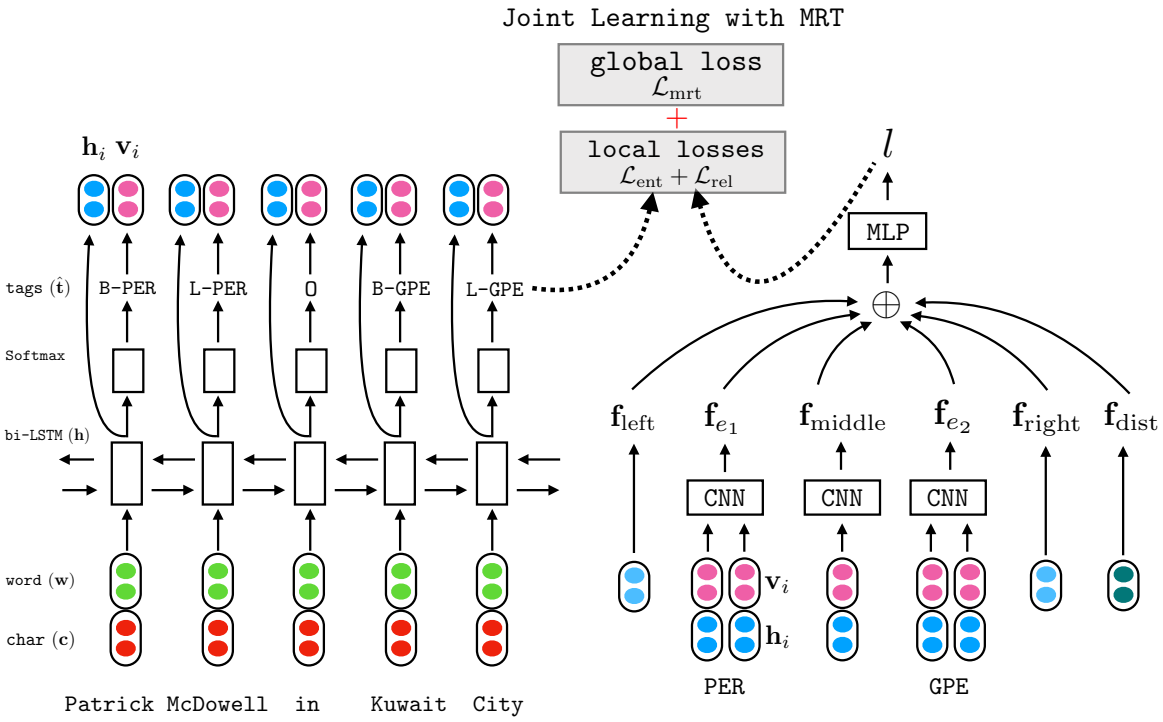

图 5.2 联合实体关系抽取的网络结构

($\mathbf{f}_{\text{middle}}$)、左边的单词特征（$\mathbf{f}_{\text{left}}$)、右边的单词特征（$\mathbf{f}_{\text{right}}$）得到三个特征向量。对于 $\mathbf{f}_{\text{middle}}$，我们在 e_1 和 e_2 之间的单词上运行 CNN。对于 $\mathbf{f}_{\text{left}}$ 和 $\mathbf{f}_{\text{right}}$，本文使用 “LSTM-Minus” 方法 [189, 221]。假设 (e_1, e_2) 的左边上下文是从位置 0 到 i，然后 $\mathbf{f}_{\text{left}} = \overrightarrow{\mathbf{h}}_i \oplus (\overleftarrow{\mathbf{h}}_0 - \overleftarrow{\mathbf{h}}_{i+1})$。相似地，如果 (e_1, e_2) 右边上下文是从位置 j to $|s|-1$，然后 $\mathbf{f}_{\text{right}} = (\overrightarrow{\mathbf{h}}_{|s|-1} - \overrightarrow{\mathbf{h}}_{j-1}) \oplus \overleftarrow{\mathbf{h}}_j$。本文使用 one-hot 特征向量 $\mathbf{f}_{\text{dist}}$ 表示句子中 e_1 和 e_2 的距离信息。

最后，$\mathbf{f}_{e_1}$，$\mathbf{f}_{e_2}$，$\mathbf{f}_{\text{middle}}$，$\mathbf{f}_{\text{left}}$，$\mathbf{f}_{\text{right}}$ 和 $\mathbf{f}_{\text{dist}}$ 被拼接为一个向量 $\mathbf{f}_{e_1,e_2}$。然后利用多层感知机得到关于关系类型标签 $\hat{l}$ 的后验分布。

$$\mathrm{P}_{\text{rel}}(\hat{l}|s, e_1, e_2; \boldsymbol{\theta}_R) = \mathrm{Softmax}(\mathbf{W}_2 \cdot \mathrm{ReLU}(\mathbf{W}_1 \cdot \mathbf{f}_{e_1,e_2})),$$

其中 $\boldsymbol{\theta}_R = \{\boldsymbol{\theta}_{e_1}, \boldsymbol{\theta}_{e_2}, \boldsymbol{\theta}_{\text{middle}}, \mathbf{W}_1, \mathbf{W}_2\}$ 是关系模型的参数（和实体模型的共享参数被忽略）。

给定输入句子 s，关系模型的训练目标是最小化

$$\mathcal{L}_{\text{rel}}(\boldsymbol{\theta}_R) = -\sum_{\substack{e_1,e_2 \in \hat{\mathcal{E}} \\ e_1 \neq e_2}} \frac{\log P_{\text{rel}}(\hat{l} = l|s, e_1, e_2; \boldsymbol{\theta}_R)}{|\hat{\mathcal{E}}|(|\hat{\mathcal{E}}| - 1)},$$

其中候选实体对 (e_1, e_2) 的真实标签 l 可以从标注数据中得到。

5.3.4 联合风险最小化训练

为了联合学习实体模型和关系模型，一个常用的策略是优化目标函数 $\widetilde{\mathcal{L}} = \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{rel}}$，这里通过共享参数实现联合训练。但是，我们认为 $\widetilde{L}$ 只是一种“局部”的损失函数，原因在于

1. $\mathcal{L}_{\text{ent}}$ 和 $\mathcal{L}_{\text{rel}}$ 都是计算局部的损失函数。例如损失函数 $\mathcal{L}_{\text{ent}}$ 是基于局部实体标签 t_i 的正确性而不是基于全局的指标（比如抽取实体的 F 分数）。

2. 实体模型和关系模型都没有显示考虑到来自另一方的信息。因为实体模型和关系模型的联系只能通过共享参数来实现，例如实体模型需要等待关系模型更新共享参数，才能产生隐式的交互，并没有考虑到关系模型的输出结果。

本文在联合模型中引入风险最小化训练框架。与优化局部损失函数相比，联合 MRT 将优化全局损失并在实体解码器和关系解码器之间提供更紧密连接。为了更好的描述算法，本文首先重整一些符号。

$\mathbf{y} \triangleq (\mathcal{E}, \mathcal{R})$ 表示真实的实体标签和关系标签。$\hat{\mathbf{y}} \triangleq (\hat{\mathcal{E}}, \hat{\mathcal{R}})$ 表示联合模型的输出。$\mathcal{Y}(s)$ 是输入句子 s 所有的可能的输出。$(\mathbf{y}, \hat{\mathbf{y}} \in \mathcal{Y}(s))$。定义联合概率分布

$$
\begin{aligned}
&P(\hat{\mathbf{y}}|s;\boldsymbol{\theta}) = P(\hat{\mathcal{E}}|s;\boldsymbol{\theta}_E)P(\hat{\mathcal{R}}|s,\hat{\mathcal{E}};\boldsymbol{\theta}_R) \\
&= \prod_i P_{\text{ent}}(\hat{t}_i|s;\boldsymbol{\theta}_E) \prod_{\substack{e_1,e_2\in\hat{\mathcal{E}} \\ e_1\neq e_2}} P_{\text{rel}}(\hat{l}|s,e_1,e_2;\boldsymbol{\theta}_R),
\end{aligned}
$$

其中 $\boldsymbol{\theta} = \boldsymbol{\theta}_E \bigcup \boldsymbol{\theta}_R$ 是联合模型的参数。

MRT 的目标是最小化以下期望损失（又名风险），

$$
\mathbf{E}_{\hat{\mathbf{y}}\sim P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}\Delta(\hat{\mathbf{y}},\mathbf{y}) = \sum_{\hat{\mathbf{y}}\in\mathcal{Y}(s)} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})\Delta(\hat{\mathbf{y}},\mathbf{y}), \tag{5.2}
$$

其中 $\Delta(\hat{\mathbf{y}}, \mathbf{y})$ 描述 $\hat{\mathbf{y}}$ 和 $\mathbf{y}$ 的差异程度。

在本文的模型中，损失函数 $\Delta(\hat{\mathbf{y}}, \mathbf{y})$ 是提高联合模型性能的关键因素。

1. 对于 $\Delta(\hat{\mathbf{y}}, \mathbf{y})$，本文考虑句子级别的实体抽取性能 F 分数和关系抽取性能 F 分数，分别表示为 $\text{F}_{\text{ent}}(\hat{\mathcal{E}}, \mathcal{E})$，$\text{F}_{\text{rel}}(\hat{\mathcal{R}}, \mathcal{R})$。为了得到相应的损失函数，我们取个负号并且归一化到 0 和 1 之间，即使用 $1 - \text{F}_{\text{ent}}(\hat{\mathcal{E}}, \mathcal{E})$ 和 $1 - \text{F}_{\text{rel}}(\hat{\mathcal{R}}, \mathcal{R})$ 作为实体损失和关系损失。利用 F 分数作为 Δ 的优点在于 F 分数刻画了输出的整体性能，并且使得训练和测试的目标一致。

2. 与在单一任务中应用 MRT 不同 [169, 198]，本文在联合抽取任务中有两个损失函数。通过在联合 MRT 优化两个任务的损失，实体模型可以根据关系模型的损失预测候选实体的合理程度，并且关系模型也可以知道实体抽取结果的可信程度。在这里，本文通过加和的方式来定义全局损失。

$$\Delta_{E+R}(\hat{\mathbf{y}}, \mathbf{y}) = 1 - \frac{1}{2}[\mathrm{F}_{\text{ent}}(\hat{\mathcal{E}}, \mathcal{E}) + \mathrm{F}_{\text{rel}}(\hat{\mathcal{R}}, \mathcal{R})].$$

为了对比 Δ_{E+R}，我们在实验中也尝试了 $\Delta(\hat{\mathbf{y}}, \mathbf{y})$ 两个替代方案，就是只考虑单个模型的损失，即 $\Delta_E(\hat{\mathbf{y}}, \mathbf{y}) = 1 - \mathrm{F}_{\text{ent}}(\hat{\mathcal{E}}, \mathcal{E})$ 和 $\Delta_R(\hat{\mathbf{y}}, \mathbf{y}) = 1 - \mathrm{F}_{\text{rel}}(\hat{\mathcal{R}}, \mathcal{R})$。

3. 除了人工制定的损失函数，本文进一步探索联合 MRT 模型是否可以自动学习损失函数。具体来说，定义 $\Gamma(\hat{\mathbf{y}})$ 是从训练集中学习到的损失函数。本文将 $\Gamma(\hat{\mathbf{y}})$，融入到 MRT 的目标函数中用来进一步增强 $\Delta(\hat{\mathbf{y}}, \mathbf{y})$，并且要求在学习过程中，正确标签 $\mathbf{y}$ 的 Γ 值（以某个间隔）小于其中的输出 $\hat{\mathbf{y}} \in Y \backslash \{\mathbf{y}\}$，

$$\begin{aligned} &\min . \sum_{\hat{\mathbf{y}} \in \mathcal{Y}(s)} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta}) \left(\Delta(\hat{\mathbf{y}}, \mathbf{y}) + \Gamma(\hat{\mathbf{y}})\right) + \xi \\ &\text{s.t.} \quad \Gamma(\mathbf{y}^*) - \Gamma(\mathbf{y}) \geq 1 - \xi,\ \xi \geq 0, \end{aligned} \tag{5.3}$$

其中 $\mathbf{y}^* = \arg\min_{\hat{y} \in Y(s)} \Gamma(\hat{\mathbf{y}})$。在这里，本文简单使用 $\Gamma(\hat{\mathbf{y}}) = 1 - P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})$ [3]，并重新规划目标函数为

$$\begin{aligned} &\sum_{\hat{\mathbf{y}} \in \mathcal{Y}(s)} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta}) \left(\Delta(\hat{\mathbf{y}}, \mathbf{y}) - P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})\right) \\ &\quad + \left[1 - P(\mathbf{y}|s;\boldsymbol{\theta}) + P(\mathbf{y}^*|s;\boldsymbol{\theta})\right]_+ \end{aligned} \tag{5.4}$$

其中 $[u]_+ = \max(u, 0)$ 是 hinge 损失函数。

优化期望损失是困难的，因为 $\mathcal{Y}(s)$ 的大小呈指数上升。实际中，我们可以通过采样一个可计算的子集 $\mathcal{Y}'(s)$ 来近似公式 5.2。具体而言，本文首先通过根据概率分布 P_{ent} 采样一个实体标签序列 $\mathbf{t}'$，从而得到一个实体集合 $\mathcal{E}'$ [4]。随后基于采样的实体可以得到所有的候选关系，我们可以为每个候选关系从关系模型概率分布 P_{rel} 中采样 l'，从而得到关系集合 $\mathcal{R}'$。算法 5.1 列出了采样过程的伪代码 [5]。在实验中，本文也尝试了算法 5.1 的变种，只采样实体模型，关系模型只使用概率最大的标签（不对关系进行采样）。

[3] 关于不同形式 $\Gamma(\hat{\mathbf{y}})$ 的研究可以作为我们未来的工作。
[4] 为了加速采样过程，本文借鉴强化学习中的 ε-greedy 算法：t'_i 以 0.9 的概率从 P_{ent} 中采样，以 0.1 的概率均匀采样。
[5] 时间复杂度是 $O(K|s|)$，与集束搜索（大小为 K）的复杂度相同 [221]。

算法 5.1 采样算法

输入: 实体模型参数 $\boldsymbol{\theta}_E$，关系模型参数 $\boldsymbol{\theta}_R$，输入句子 s，采样大小 K
输出: $\mathcal{Y}(s)$ 的子集 $\mathcal{Y}'(s)$

$\mathcal{Y}'(s) \leftarrow \{(\mathcal{E}, \mathcal{R})\}$
while $|\mathcal{Y}'(s)| \leq K$ **do**
  $i \leftarrow 1$
  while $i \leq |s|$ **do**
    以 0.9 的概率，采样 $t'_i \sim P_{\text{ent}}(\cdot|s; \boldsymbol{\theta}_E)$
    以 0.1 的概率, 均匀采样 t'_i
    $i \leftarrow i + 1$
  end while
  $\mathcal{E}' \leftarrow \mathbf{t}' = t'_1, t'_2, \cdots, t'_{|s|}$
  $\mathcal{R}' \leftarrow \emptyset$
  for $e_1, e_2 \in \hat{\mathcal{E}}, e_1 \neq e_2$ **do**
    采样 $l' \sim P_{\text{rel}}(\cdot|s, e_1, e_2; \boldsymbol{\theta}_R)$
    $\mathcal{R}' \leftarrow \mathcal{R}' \cup \{(e_1, e_2, l')\}$
  end for
  $\mathcal{Y}'(s) \leftarrow \mathcal{Y}'(s) \cup \{(\mathcal{E}', \mathcal{R}')\}$
end while

当使用采样的子集合 $\mathcal{Y}'(s)$ 时，我们重新修订了之前 MRT 的目标函数，

$$\mathcal{L}_{\text{mrt}}(\boldsymbol{\theta}) = \sum_{\hat{\mathbf{y}} \in \mathcal{Y}'(s)} Q(\hat{\mathbf{y}}|s; \boldsymbol{\theta}, \mu, \alpha) \Delta(\hat{\mathbf{y}}, \mathbf{y}), \tag{5.5}$$

其中 $Q(\hat{\mathbf{y}}|s; \boldsymbol{\theta}, \mu, \alpha)$ 是概率 $P(\hat{\mathbf{y}}|s; \boldsymbol{\theta})$ 在集合 $\mathcal{Y}'(s)$ 上的概率归一化。

$$Q(\hat{\mathbf{y}}|s; \boldsymbol{\theta}, \mu, \alpha) = \frac{1}{Z}[P(\hat{\mathcal{E}}|s, \boldsymbol{\theta}_E)^{\mu} P(\hat{\mathcal{R}}|s, \hat{\mathcal{E}}, \boldsymbol{\theta}_R)^{1-\mu}]^{\alpha}$$
$$Z = \sum_{(\mathcal{E}', \mathcal{R}') \in \mathcal{Y}'(s)} [P(\mathcal{E}'|s, \boldsymbol{\theta}_E)^{\mu} P(\mathcal{R}'|s, \mathcal{E}', \boldsymbol{\theta}_R)^{1-\mu}]^{\alpha}$$

超参数 α 控制 Q 分布的尖锐程度 [137]。μ 决定实体模型和关系模型在 Q 中的重要性。

通常使用 MRT 之前，会使用最大似然估计进行预训练，然后使用 MRT 进行微调。在这种情况下，只要联合模型存在概率分布 $P(\hat{\mathbf{y}}|s, \boldsymbol{\theta})$，就可以使用 MRT，例如全局归一化中的 P [221]。因此，我们认为 MRT 是一个灵活轻便的联合学习框架。

5.3.5 模型训练

为了训练联合抽取模型，本文首先用目标函数 $\widetilde{\mathcal{L}}$ 预训练模型（最小化局部损失函数），然而同时优化局部和全局的损失函数 $\widetilde{\mathcal{L}} + \mathcal{L}_{\text{mrt}}$。这个设置与之前的工作应用 MRT 有些不同，他们在第二阶段只优化 $\mathcal{L}_{\text{mrt}}$。我们发现在实验中加入 $\widetilde{\mathcal{L}}$ 会使得训练过程更稳定。

在预训练阶段时，本文在实体模型中使用策略采样 [12, 128]。模型使用 dropout 进行正则化并使用 Adadelta 优化器 [216]。本文使用验证集选择模型：在固定数量的训练轮数内，挑选出在验证集上具有最好关系抽取性能的模型进行测试。我们更在在意端到端关系抽取的性能，因此本文通过关系抽取结果选择模型。此外，还可以同时考虑实体识别和关系抽取的性能用来选择更好的模型，我们将在未来的工作中研究更好的模型选择算法。

5.3.6 MRT 目标函数的梯度

本文在此给出 MRT 目标的梯度的推导过程 (公式 5.5)。文献 [169, 198] 给出类似的推导。值得一提的是，MRT 损失函数的不可分解性并不能使目标不可微分。事实上，当给定采样集合 $\mathcal{Y}'(s)$，可以通过自动求导工具计算以下梯度。

$$\mathcal{L}_{\text{mrt}}(\boldsymbol{\theta}) = \sum_{\hat{\mathbf{y}}\in\mathcal{Y}'(s)} Q(\hat{\mathbf{y}}|s;\boldsymbol{\theta},\alpha)\Delta(\hat{\mathbf{y}},\mathbf{y}) = \frac{\sum_{\hat{\mathbf{y}}\in\mathcal{Y}'(s)} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})^{\alpha}\Delta(\hat{\mathbf{y}},\mathbf{y})}{\sum_{\mathbf{y}^*\in\mathcal{Y}'(s)} P(\mathbf{y}^*|s;\boldsymbol{\theta})^{\alpha}} \triangleq \frac{G(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}.$$

分子 $G(\boldsymbol{\theta})$ 的梯度是

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) &= \alpha \sum_{\hat{\mathbf{y}}\in\mathcal{Y}'(s)} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})^{\alpha-1}\nabla_{\boldsymbol{\theta}} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})\Delta(\hat{\mathbf{y}},\mathbf{y}) \\ &= \alpha \sum_{\hat{\mathbf{y}}\in\mathcal{Y}'(s)} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})^{\alpha}\frac{\nabla_{\boldsymbol{\theta}} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}\Delta(\hat{\mathbf{y}},\mathbf{y}).\end{aligned}$$

分母 $Z(\boldsymbol{\theta})$ 的梯度是

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} Z(\boldsymbol{\theta}) &= \alpha \sum_{\mathbf{y}^*\in\mathcal{Y}'(s)} P(\mathbf{y}^*|s;\boldsymbol{\theta})^{\alpha-1}\nabla_{\boldsymbol{\theta}} P(\mathbf{y}^*|s;\boldsymbol{\theta}) \\ &= \alpha \sum_{\mathbf{y}^*\in\mathcal{Y}'(s)} P(\mathbf{y}^*|s;\boldsymbol{\theta})^{\alpha}\frac{\nabla_{\boldsymbol{\theta}} P(\mathbf{y}^*|s;\boldsymbol{\theta})}{P(\mathbf{y}^*|s;\boldsymbol{\theta})}.\end{aligned}$$

因此, 有

$$\begin{aligned}\frac{G'}{Z} &= \alpha \sum_{\hat{\mathbf{y}}\in\mathcal{Y}'(s)} \frac{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})^{\alpha}}{\sum_{\mathbf{y}^*\in\mathcal{Y}'(s)} P(\mathbf{y}^*|s;\boldsymbol{\theta})^{\alpha}}\frac{\nabla_{\boldsymbol{\theta}} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}\Delta(\hat{\mathbf{y}},\mathbf{y}) \\ \frac{Z'}{Z} &= \alpha \sum_{\hat{\mathbf{y}}\in\mathcal{Y}'(s)} \frac{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})^{\alpha}}{\sum_{\mathbf{y}^*\in\mathcal{Y}'(s)} P(\mathbf{y}^*|s;\boldsymbol{\theta})^{\alpha}}\frac{\nabla_{\boldsymbol{\theta}} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}.\end{aligned}$$

根据标准的结果, 有

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{mrt}}(s;\boldsymbol{\theta},\alpha) = \frac{G'\cdot Z - G\cdot Z'}{Z\cdot Z} = \frac{G'}{Z} - \frac{G}{Z}\cdot\frac{Z'}{Z} = \frac{G'}{Z} - \mathcal{L}_{\text{mrt}}\cdot\frac{Z'}{Z}$$

表 5.1 模型超参设置

Modules	Setting
Embeddings	· $\dim(\mathbf{w}_i) = 100$, initialized with Glove vectors [142] · $\dim(\mathbf{c}_i) = \dim(\text{char embedding}) = 50$ the window sizes of CNN ($\boldsymbol{\theta}_c$) are 2 and 3
NN	· $\dim(\mathbf{h}_i) = 128$ · $\dim(\mathbf{f}_{e_1}) = \dim(\mathbf{f}_{e_2}) = \dim(\mathbf{f}_{\text{middle}}) = 50$ · the window sizes of CNNs ($\boldsymbol{\theta}_{e_1}, \boldsymbol{\theta}_{e_2}$) are 2 and 3 · training epochs: 1000 (ACE05), 50 (NYT) · Adadelta: gradient clipping with max norm 1 · batch size: 100 · dropout rate: 0.5
MRT	· without Γ: $\mu = 1.0, \alpha = 0.0001, K = 3$ · with Γ: $\mu = 1.0, \alpha = 1, K = 2$ · training epochs: 25 (ACE05), 10 (NYT)

$$= \alpha \mathbf{E}_{\hat{\mathbf{y}} \sim Q(\hat{\mathbf{y}}|s;\boldsymbol{\theta},\alpha)} \left[\frac{\nabla_{\boldsymbol{\theta}} P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})}{P(\hat{\mathbf{y}}|s;\boldsymbol{\theta})} \left[\Delta(\hat{\mathbf{y}}, \mathbf{y}) - \mathcal{L}_{\text{mrt}}(s;\boldsymbol{\theta},\alpha) \right] \right].$$

5.4 实验

5.4.1 设置

本文在 NYT 数据集和 ACE05 数据集上评估本文提出的框架。有关 NYT 和 ACE05 数据集的详细情况请参考 2.3.3 节。本文将主要讨论 ACE05 数据集的结果。

表 5.1 列出了默认模型的超参数设置。联合 MRT 中引入了超参数 μ, α, K, 本文使用验证集挑选了这些超参数。除此之外，我们没有广泛调整其它的超参数。例如, 本文在 ACE05 和 NYT 中使用相同的设置，而不是在每个数据集上调整超参数。

本文评价模型结果用精确率（P）、召回率（R）和 F 分数。具体而言，如果输出实体 e 的类型和头部区域是正确的，则输出实体 e 是正确的，如果输出关系 r 的 e_1, e_2, l 是正确的, 则输出关系 r 是正确的（即精确匹配）。

5.4.2 在 ACE05 数据集上的结果

我们首先将提出的模型与已有的实体关系抽取系统进行比较（表 5.2）。总体来看，本文的神经网络模型（NN）具有很强的竞争力。在使用 MRT 后，和现在最好的模型相比，实体识别和关系抽取的性能均获得了不小的提升。值得注意的是，本文的模型不会使用其他语言学资源，如词性标注和依存树。我们试图在按照文献 [221] 中添加句法特征，但没有观察到性能的提高。此外，我们还有以下两个结

表 5.2 在 ACE05 上的测试集结果

Model	Entity			Relation		
	P	R	F	P	R	F
L&J [2014]	**85.2**	76.9	80.8	**65.4**	39.8	49.5
M&B [2016]	82.9	**83.9**	83.4	57.2	54.0	55.6
Zhang [2017]	-	-	83.5	-	-	57.5
K&C [2017]	84.0	81.3	82.6	55.5	51.8	53.6
NN	84.0	82.9	83.4	59.5	**56.3**	57.8
MRT	83.9	83.2	**83.6**	64.9	55.1	**59.6**

论。

- 在仅依赖共享参数的系统中（文献 [81, 128] 和 NN），NN 模型取得了最好的结果。一个可能的原因是在先前的联合学习模型中没有充分探索 “RNN + CNN” 络结构。更重要的是，它表明如何构建强大的子模型和利用共享参数仍然是该任务的关键问题。我们在表 5.3 中列出了每种关系类型的性能。
- 与采用全局归一化方法的联合解码系统 [221] 相比，MRT 主要改进了关系抽取结果。我们认为改进可能来自 MRT 直接优化的是句子级别损失：两个系统都考虑了解码器之间的相互作用，并且两个目标都是通过采样来近似的，但是 MRT 优化了 F 得分，而文献 [221] 优化标签的准确率。对于文献 [102] 中的联合解码系统，虽然它的效果低于最近基于神经网络的模型，但是在未来的工作中，将传统的特征工作替换为神经网络自动抽取特征，然后再探究和 MRT 的差异，这也是值得期待的工作。

接下来，本文评估具有不同损失函数和采样方法的联合 MRT。如 5.3.4 节所述，$\Delta(\hat{\mathbf{y}}, \mathbf{y})$ 有三个方案 (Δ_{E+R}, Δ_E, Δ_R) 和自动学习的损失函数 Γ。表 5.4的前五行显示了它们在测试数据上的性能. 我们对结果有三个结论。

1. Δ_R, Δ_{E+R} 的 F 分数高于 Δ_E 和 NN. 因此，在 $\Delta(\hat{\mathbf{y}}, \mathbf{y})$ 中考虑关系损失有助于关系抽取。我们认为提升的主要原因是：模型优化关系损失可能会突出关系中出现的实体，这为关系抽取模型提供了更好的候选关系集合。
2. Δ_E 具有最佳的实体识别性能，这意味着实体损失可能对实体模型有益。在考虑关系损失 Δ_{E+R} 之后，实体性能略有下降。一个可能的原因是我们的模型选择策略仅关注关系部分，因此可能不会选择具有更好实体性能的模型。因为关系的性能通常在 57% 左右，且波动比较大，而实体的性能通常在 83% 左右，且比较稳定。
3. 自动学习的损失函数 Γ 有助于提高模型的性能，但仅使用 Γ 不如人工制定的 Δ 函数（根据评估指标量身定制）有效。这可以说明通过结合先验知识和来

表 5.3 每个关系类型下的模型结果

Relation Type	Model	P	R	F
ART (146)	M&B [2016]	36.3	55.2	43.8
	K&C [2017]	43.1	**61.1**	**50.5**
	NN	51.6	44.5	47.8
	MRT	**59.2**	41.8	49.0
PART-WHOLE (175)	M&B [2016]	56.0	**53.8**	54.8
	K&C [2017]	52.0	**53.8**	52.8
	NN	57.2	49.7	53.2
	MRT	**59.9**	52.0	**55.7**
PER-SOC (73)	M&B [2016]	67.1	67.1	67.1
	K&C [2017]	65.7	64.8	65.2
	NN	76.5	**71.2**	**73.8**
	MRT	**77.3**	69.9	73.4
PHYS (278)	M&B [2016]	48.9	**51.3**	**50.0**
	K&C [2017]	38.8	42.6	40.6
	NN	45.8	48.9	47.3
	MRT	**50.0**	42.8	46.1
GEN-AFF (99)	M&B [2016]	41.4	**64.0**	**50.2**
	K&C [2017]	48.4	51.6	50.0
	NN	56.1	37.4	44.9
	MRT	**60.9**	39.4	47.9
ORG-AFF (354)	M&B [2016]	69.2	70.4	69.7
	K&C [2017]	70.6	70.0	70.3
	NN	72.1	**72.3**	72.2
	MRT	**78.0**	70.1	**73.8**

表 5.4 不同损失函数和不同采样方法下 MRT 的结果

Settings		F of Entity	F of Relation
Default sampling	Δ_E	**83.8** +0.4	57.9 +0.1
	Δ_R	83.5 +0.1	58.9 +1.1
	Δ_{E+R}	83.6 +0.2	59.0 +1.2
	Γ	83.6 +0.2	58.3 +0.5
	$\Gamma + \Delta_{E+R}$	83.6 +0.2	**59.6** +1.8
Only sampling entity	Δ_E	**83.7** +0.3	57.4 -0.4
	Δ_R	83.5 +0.1	59.1 +1.3
	Δ_{E+R}	83.6 +0.2	57.9 +0.1
	Γ	83.6 +0.2	58.7 +0.9
	$\Gamma + \Delta_{E+R}$	83.3 -0.1	**59.2** +1.4

表 5.5 Δ_E 和 Δ_R 的 MRT 结果

Settings	F of Entity	F of Relation	α	μ
Δ_{E+R}	83.6 +0.2	59.0 +1.2	1e−4	1.0
Δ_E	83.6 +0.2	58.6 +0.8	1e−5	1.0
Δ_R	83.4 +0.0	58.8 +1.0	1e−5	0.5

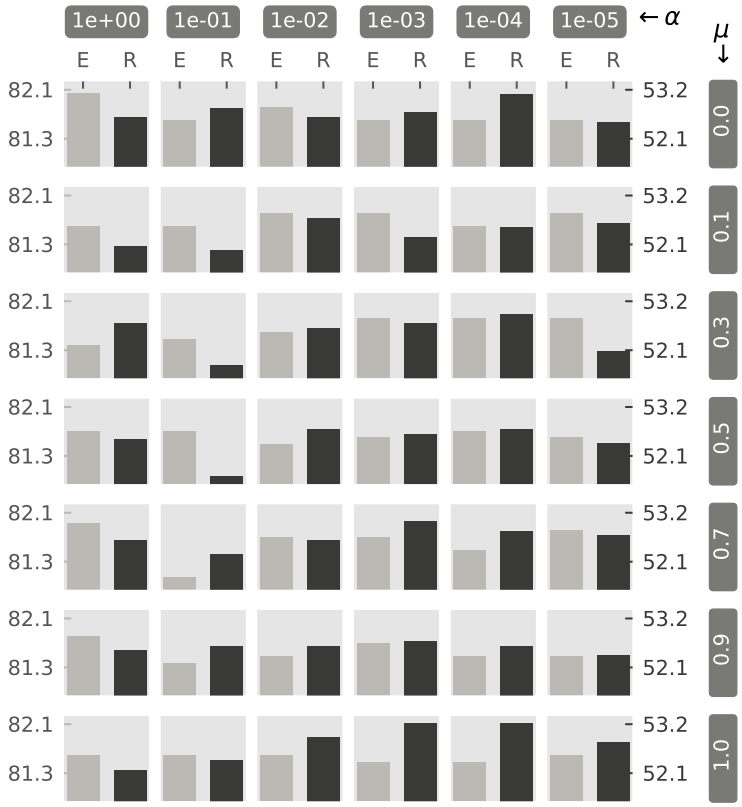

图 5.3 在验证集上随着不同的 Q 分布 MRT 的结果

自数据集的信息，取得了最好的实验结果。

关于采样方法，本文测试算法 5.1 的一个变体，它对实体进行采样，但不对关系进行采样（表 5.4 的最后五行）。与默认采样算法相比，它具有相似的实体识别性能，但其在关系抽取上的性能是有差异的。具体而言，在 $\Delta(\hat{\mathbf{y}}, \mathbf{y})$ 中添加实体损失，即 Δ_E, Δ_{E+R})，会对关系结果产生负面影响。这可能表明，相对于关系损失，实体损失的影响更大，以至于联合模型侧重于实体模型。另一方面，自动学习的损失函数 Γ 的性能对采样方法不太敏感。我们还没有清楚地理解采样算法和损失函数之间的关系。

第三，本文在图 5.3 和图 5.4 中的验证集使用 Δ_{E+R} 呈现 MRT 超参数的影响（其他设置具有类似的结果）。我们发现，对于这里的超参数，实体模型和关系模型很难相互一致：实现高关系性能的参数通常得到较低的实体性能，反之亦然。因此，如果我们仅通过查看关系抽取结果来执行模型选择，则联合模型可能牺牲实体识别性能。对于 α 和 μ （图 5.3），我们观察到在 ACE05 数据集上，模型更倾向于使用边界处的小 α （意味着更尖锐的 Q 分布），和 μ （决定 Q 接近实体模型或关系模型）。关于样本大小 K（图 5.4），在 2 到 6 范围内，随着 K 的增大，我们没有观察到抽取的性能收敛。由于随着我们增加样本量，计算成本迅速增加（在们

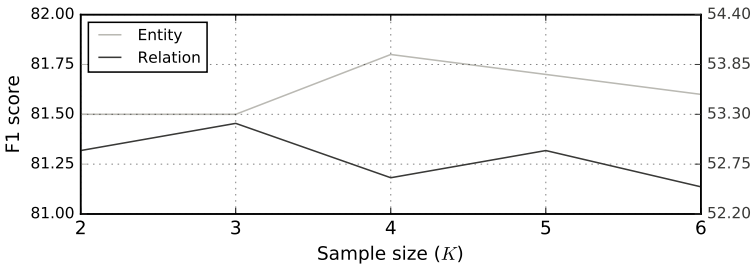

图 5.4 验证集上随着采样大小 KMRT 的结果

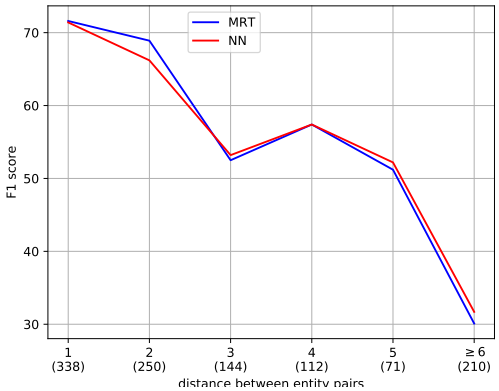

图 5.5 随着实体对距离不同 ACE05 数据集的结果

的实现中 $K = 5$ 大概比 $K = 2$ 慢 2 倍），所以我们使用较小的 K。

第四，图 5.5列出了 NN 和 MRT 相对于候选实体对之间的不同距离的性能。我们发现在距离等于 2 时 MRT 的关系抽取性能优于 NN。这意味着尽管句子级别 F 在损失函数中提供全局信息，但它仍然不足以捕获长距离依赖。因此，集成 MRT 和联合解码算法可能是一个很有前景的方向。

最后，本文列出 MRT 模型和 NN 模型的输出结果。本文定义符号“[entity span] $_{\texttt{ENT-TYPE[:REL-TYPE\_REL-ID]}}$”表示实体（“entity span”）的类型是 ENT-TYPE。如果实体之间构成了某种关系 REL-TYPE，那么两个实体的 REL-ID 是相同的。对于样例 S1，NN 识别了实体“[kennedy airport] $_{\texttt{FAC}}$”，但是 MRT 没有正确识别到“kennedy”。对于样例 S2，MRT 识别了实体“[aol] $_{\texttt{ORG}}$”，但是 NN 没有识别出。这两个例子说明了尽管这两个模型在实体识别中都出现了错误，但是 MRT 更容易识别出存在关系的实体。对于样例 S3，MRT 识别了实体“[our] $_{\texttt{ORG}}$”和实体“[founder] $_{\texttt{PER}}$”之间的关系 ORG-AFF，但是 NN 没有识别出这个关系，尽管两个实体已经识别正确。对于样例 S4，MRT 没有抽取出 PHYS 关系，然而 NN 正确找出了此关系。总体来说，MRT 和 NN 各自都有擅长的处理的情况，对于长距离的关系都还有待加强。

S1	first an update on a long running air safety investigation a year and a half near an [airline]$_{\text{VEH}}$ crashed near [new york] $_{\text{GPE,PART-WHOLE-1}}$'s [kennedy airport]$_{\text{FAC: PART-WHOLE-1}}$ there is controversy whether the disaster could have been averted .
NN	first an update on a long running air safety investigation a year and a half near an [airline]$_{\text{VEH}}$ crashed near [new york] $_{\text{GPE,PART-WHOLE-1}}$'s [kennedy airport]$_{\text{FAC: PART-WHOLE-1}}$ there is controversy whether the disaster could have been averted .
MRT	first an update on a long running air safety investigation a year and a half near an [airline]$_{\text{VEH}}$ crashed near [new york] $_{\text{GPE,PART-WHOLE-1}}$'s kennedy [airport]$_{\text{FAC: PART-WHOLE-1}}$ there is controversy whether the disaster could have been averted .

S2	the question , [i]$_{\text{PER}}$ 'm an [aol] $_{\text{ORG:ORG-AFF-1}}$ [shareholder] $_{\text{PER: ORG-AFF-1}}$ sitting at [home] $_{\text{FAC}}$, hearing this news , done this set off a few alarms ?
NN	the question , [i]$_{\text{PER}}$ 'm an aol [shareholder] $_{\text{PER:PHYS-1}}$ sitting at [home] $_{\text{FAC: PHYS-1}}$, hearing this news , done this set off a few alarms ?
NN	the question , [i]$_{\text{PER}}$ 'm an [aol] $_{\text{ORG:ORG-AFF-1}}$ [shareholder] $_{\text{PER:ORG-AFF-1,PHYS-1}}$ sitting at [home] $_{\text{FAC: PHYS-1}}$, hearing this news , done this set off a few alarms ?

S3	[our] $_{\text{ORG: ORG-AFF-1}}$ [founder] $_{\text{PER: ORG-AFF-1}}$ here at [cnn] $_{\text{ORG}}$, [ted turner]$_{\text{PER}}$, has sold more than half 0 [his]$_{\text{PER: ORG-AFF-2}}$ stake in [aol time warner] $_{\text{ORG: ORG-AFF-2}}$.
NN	[our] $_{\text{ORG}}$ [founder] $_{\text{PER}}$ here at [cnn] $_{\text{ORG}}$, [ted turner]$_{\text{PER}}$, has sold more than half 0 [his]$_{\text{PER: ORG-AFF-2}}$ stake in [aol] $_{\text{ORG: ORG-AFF-2}}$ time [warner] $_{\text{PER}}$.
MRT	[our] $_{\text{ORG: ORG-AFF-1}}$ [founder] $_{\text{PER: ORG-AFF-1}}$ here at [cnn] $_{\text{ORG}}$, [ted turner]$_{\text{PER}}$, has sold more than half 0 [his]$_{\text{PER: ORG-AFF-2}}$ stake in [aol] $_{\text{ORG: ORG-AFF-2}}$ time [warner] $_{\text{PER}}$.

S4	[john scottsdale] $_{\text{PER: PHYS-1}}$ is on the front lines in [iraq]$_{\text{GPE: PHYS-1}}$.
NN	[john scottsdale] $_{\text{PER: PHYS-1}}$ is on the front lines in [iraq]$_{\text{GPE: PHYS-1}}$.
MRT	[john scottsdale] $_{\text{PER}}$ is on the front lines in [iraq]$_{\text{GPE}}$.

表 5.6 在 NYT 数据集上的结果

Model	Relation		
	P	R	F
zheng [228]	61.5	41.4	49.5
NN	61.8	**43.3**	50.9
MRT	**67.4**	42.0	**51.7**
ren [155]	42.3	**51.1**	46.3
NN （exact match）	59.4	41.7	49.0
MRT （exact match）	**65.2**	40.6	**50.0**

5.4.3 在 NYT 数据集上的结果

本文简要列出中 NYT 数据集的结果（表 3.1）。基准方法是文献 [155]，这是基于于实体和关系的联合表示建模，以及文献 [228]，这是利用增强序列标注标签集进行联合解码。NN 和 MRT 均优于基准系统结果。特别是与文献 [228] 中的联合标签方案相比，MRT 对关系抽取模型没有任何限制，可以更有效地探索大型 NYT 训练集。同时，由于训练集是自动生成的，因此在 MRT 中观察到的全局损失是存在噪音的。像关于赌博机结构化预测工作 [88, 135] 的结果表明，当联合学习的监督是部分和带有噪音时，MRT 也是一个合理的选择。

5.5 总结

本文提出了一个简单有效的神经网络模型用来解决联合实体关系抽取任务。该网络主要利用 RNN + CNN 的架构，RNN 为实体识别自动抽取特征，CNN 为关系抽取自动抽取特征，并且 RNN 的隐层输出作为 CNN 的输入，整个网络可以联合训练。其次，本文在神经网络模型上引入了风险最小化训练方法 (MRT)，可以直接优化全局的损失函数。MRT 可以增强子模型之间的联系。在两个数据集的广泛实验证明了联合 MRT 的有效性。

第六章　基于图卷积网络的联合实体关系抽取

和前一章相同，本章从“联合模型”的角度提出一种新的针对联合实体关系抽取任务的方法。与前一章的区别在于，本文设计了一种新的联合抽取任务。具体来说，首先是识别出实体的边界，然后进行实体类型和关系类型的推理。为了解决联合类型推理，本文定义了实体-关系二分图，并且在此图上使用一个图卷积网络来学习丰富的特征，从而进行类型推理。通过引入一个二元关系分类器，本文能够一种更有效和可解释的方式来利用此二分图的结构化信息。在 ACE05 数据集上的实验结果表明，在联合抽取任务上，本文的模型在实体识别的性能取得最好，在关系抽取上的性能和当前最好的方法相当。

6.1　引言

从纯文本中抽取实体和关系是自然语言处理中的一项重要且具有挑战性的任务。给定一个句子，该任务旨在检测具有特定类型（实体）的文本片段以及那些文本片段之间的语义关系（关系）。例如在图 6.1 中，“Toefting”是一个人名实体（PER），“teammates”是一个人名实体（PER），并且这个两个实体具有 PER-SOC 关系。

为了解决实体关系抽取的任务，各种方法被提出来。大致可以分为两类：基于流水线的模型和联合模型。在本章工作中，本文仍然关注于联合实体关系抽取模型。之前联合抽取任务通常分解为实体识别和关系抽取两个子任务，其中实体识别通常可以用一个序列标注框架解决，然后被识别的实体用作关系模型的输入。这样的联合抽取方式用一个序列标注框架同时处理实体边界检测和实体类型推理，以至于在实体类型推理中没有直接考虑到关系类型的信息，同样在关系类型推理中没有直接考虑到所有实体的信息。为了能够进行联合类型推理，本文将实体边界检测从实体识别任务中分离出来，然后同时对实体关系类型进行推理。与之前联合抽取方法相比，本文重新定义的子任务更容易进行联合实体关系类型推理。

通过检查 ACE05 上现有模型 [175] 的性能，我们发现对于许多实体它们的边界被正确识别，但它们的实体类型是错误的。抽取类型实体的 F 分数约为 83%，而抽取无类型实体的 F 分数约为 90%。如果我们有一个更好的类型推理模型，可能会获得更好的联合抽取性能。同时，我们观察到对实体和关系类型的联合推断可

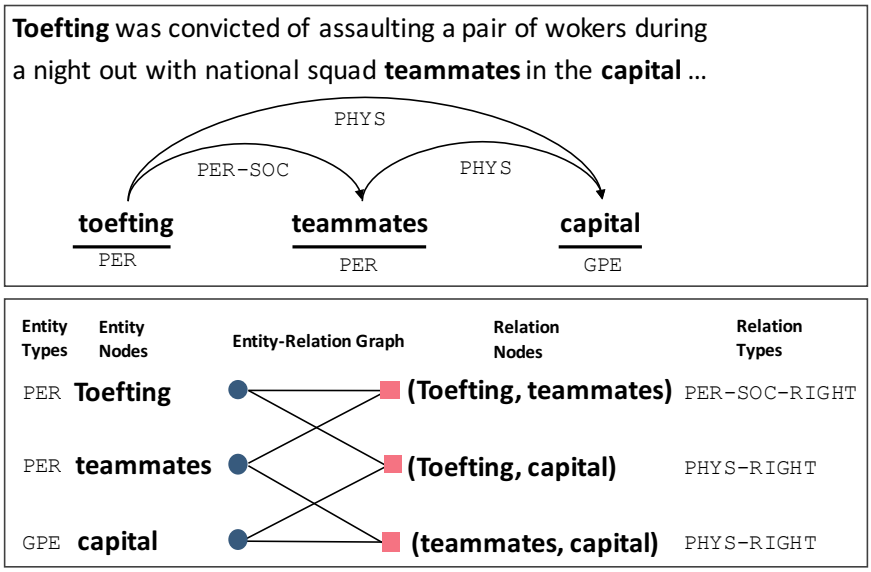

图 6.1 ACE05 标注样例

能比独立预测它们更好。例如在图 6.1中，PER-SOC 关系表明“Toefting”的类型可能是 PER，反之亦然。此外，实体 PER（“Toefting”）和关系 PER-SOC 也可能会从同一句子中其它关系获益，比如 PHY 关系。

在本章工作中，本文将联合实体关系抽取定义为两个子任务：实体边界检测和实体关系类型推断。对于实体边界检测，本文将其视为序列标注问题。对于联合类型推断，本文提出了一种基于图卷积网络（GCN）的新颖简洁的联合模型 [85]。具体而言，给定句子中所有检测到的只有边界的实体，本文定义实体-关系二分图。对于每个无类型实体，我们将其视为一个实体节点；对于每个实体-实体对，我们将其视为一个关系节点。关系节点和对应的实体节点之间连边（图 6.1的最后部分）。通过图卷积运算，二分图中的每个节点都可以收集来自邻居节点的信息，从而每个节点得到的表示都融合了图的结构化信息。这有助于我们得到更好的实体节点和关系节点的表示。例如在图 6.1中，为了预测 PER（“Toefting”），本文的联合模型可以汇集 PER-SOC，PHYS，PER（“teammates”）和 GPE（captital）的信息。

为了进一步利用二分图的结构，我们对于不同类型的边分配不同的权重。本文引入了二元关系分类任务，该任务用于确定两个实体之间是否存在关系。与以前的基于 GCN 的模型 [166, 226] 不同，在本文定义的二分图中，邻接矩阵的值是根据二元关系分类的输出而确定，这使得本文使用的邻接矩阵更具解释性。综上所述，本章工作的主要贡献是

- 本文基于图卷积网络（GCN）提出了一种新颖简洁的联合模型来处理的联合类型推理问题。
- 本文引入二元关系分类任务，以更有效和可解释的方式探索实体关系二分图

的结构。

- ACE05 数据集上的实验结果表明：本文提出的联合模型在实体识别的性能取得最好，在关系抽取上的性能和当前最好的方法相当。

6.2 相关工作

图神经网络（GNNs）具有强大的图表达能力，最近受到越来越多的关注 [10, 24, 230]。图形卷积网络（GCN）是 GNN 的典型变体之一 [22, 44, 85]。它已成功应用于许多 NLP 任务，如文本分类 [210]，语义角色标记 [120]，关系抽取 [226]，机器翻译 [9] 和知识库填充 [166]。我们注意到 GCN 的大多数先前应用程序专注于单个任务，而联合实体关系抽取包含多个子任务。在联合学习场景中探究 GCN 是这项工作的主要目的。一项密切相关的工作是 [37]，其任务是给定实体的关系抽取。本文的工作可被视为其工作的端到端扩展。

6.3 基于图卷积网络的联合实体关系抽取系统

6.3.1 图卷积网络

这里我们简要描述图卷积网络（GCN）。给定一个包含 n 个节点的图，GCN 的目标是学习图上的结构相关的节点表示，网络的输入为：

- 一个 $n \times d$ 输出节点向量 $\mathbf{H}$，其中 n 是节点个数，d 是节点向量的维度。
- 一个 $n \times n$ 图结构的矩阵表示，比如邻接矩阵 $\mathbf{A}$ [1]。

在 L 层的 GCN 中，每层都可以写成非线性函数

$$\mathbf{H}^{(l+1)} = \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}) \tag{6.1}$$

$\mathbf{H}^{(0)} = \mathbf{H}$，其中 $\hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ 是一个归一化的对称的邻接矩阵。$\mathbf{W}^{(l)}$ 是第 l 层的参数矩阵。$\mathbf{D}$ 是对角节点度数矩阵，其中 $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$。$\sigma$ 是非线性激活函数，比如 ReLU 激活函数。最后，我们可以获得节点级别的输出矩阵 $\mathbf{Z} = \mathbf{H}^{(L)}$，该矩阵的维度是 $n \times d$。

[1] 为了融入本身节点信息，我们为每个节点添加一个自循环，其中对于每个节点 i，$A_{ii} = 1.0$。

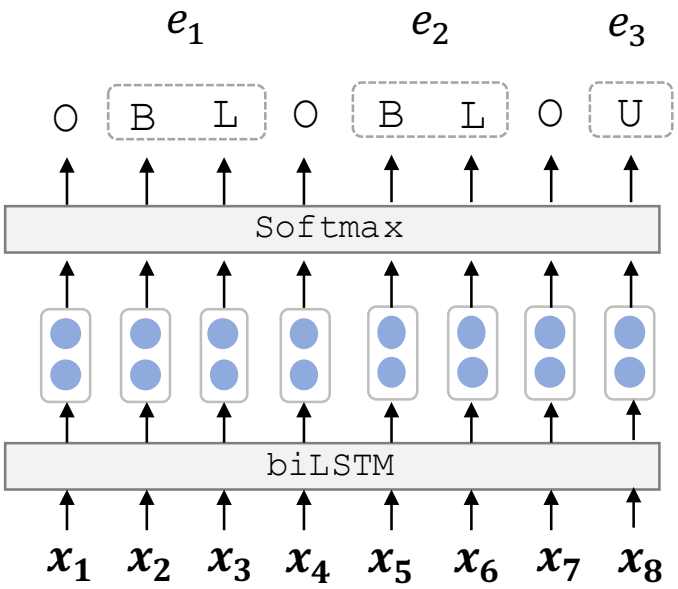

图 6.2 实体边界检测模型

6.3.2 任务定义

本文定义联合实体关系抽取任务。给定一个句子 $s = w_1, \ldots w_{|s|}$（w_i 是一个单词），该任务旨在抽取一组具有特定类型的实体边界集合 $\mathcal{E}$ 和一组关系集合 $\mathcal{R}$。一个实体边界 $e \in \mathcal{E}$ 是一个具有实体类型 y 的词序列（比如人名（PER），组织机构（ORG））。一个关系 r 是一个五元组 (e_1, y_1, e_2, y_2, l)，其中 e_1 和 e_2 是具有实体类型 y_1, y_2 的两个实体边界，l 是这两个实体之间的语义关系。（比如组织从属关系（ORG-AFF））。本文用 $\mathcal{T}_e$，$\mathcal{T}_r$ 分别表示可能的实体类型集合和关系类型集合。

在本章工作中，本文将联合实体关系抽取任务分解为两部分，即实体边界检测和实体关系类型推断。我们首先将实体边界检测视为序列标注任务（6.3.3 节），然后构造实体-关系二分图（6.3.4 节）对实体节点和关系节点执行联合类型推断（6.3.5 节）。所有子模型共享参数并共同训练。与现有的联合学习算法 [81, 128, 175, 221] 不同，本文提出了一种基于 GCN 简洁的联合模型，可以对实体和关系进行联合类型推理。它可以同时考虑同一个句子中多个实体类型和关系类型之间的交互。

6.3.3 实体边界检测

为了抽取句子中的实体边界，本文采用 BILOU 标签方案，B，I，L 和 O 分别表示实体的开始，中间，最后和外面，U 表示单个词的实体。例如，对于人名实体（PER）“Patrick McDowell”，我们将 B 分配给 “Patrick”，将 L 分配给 “McDowell”。

给定一个输入句子 s，本文使用参数为 $\boldsymbol{\theta}$ 的 biLSTM 模型 [63] 来捕获来自 s 的

前向和后向信息。

$$\mathbf{h}_i = \text{biLSTM}(\mathbf{x}_i; \boldsymbol{\theta}), \tag{6.2}$$

其中 $\mathbf{h}_i$ 是 biLSTM 的隐层状态向量（每个位置的前向 LSTM 和后向 LSTM 两个隐层状态向量拼接而成），$\mathbf{x}_i$ 是单词 w_i 的表示，是由预训练的词向量和基于字符级别的 CNN 输出向量拼接而成。然后，本文使用 softmax 函数预测 w_i 对应的标签 $\hat{t}_i$，

$$P(\hat{t}_i|s) = \text{Softmax}(\mathbf{W}_{\text{span}}\mathbf{h}_i),$$

其中 $\mathbf{W}_{\text{span}}$ 是模型参数。给定输入句子 s 和真实标签序列 $\mathbf{t} = t_1, \ldots, t_{|s|}$，实体边界模型的目标函数是最小化以下损失函数 [2]

$$\mathcal{L}_{\text{span}} = -\frac{1}{|s|}\sum_{i=1}^{|s|} \log P(\hat{t}_i = t_i|s). \tag{6.3}$$

6.3.4 实体-关系二分图

给定一个句子，从实体边界模型中可以得到无类型实体的预测标签序列 $\hat{\mathbf{t}}$，然后可以得到该句子预测的无类型实体集合 $\hat{\mathcal{E}}$。我们考虑在 $\hat{\mathcal{E}}$ 所有的实体对作为候选关系 [3]。然后本文构建了一个异构的无向二分图 $\mathcal{G}^s$，它包含两种类型的节点：

- 实体节点：对于每个无类型实体，我们将其视为一个实体节点。所以在图 $\mathcal{G}^s$ 中实体节点的数量是 $|\hat{\mathcal{E}}|$；
- 关系节点：对于每个实体-实体对，我们将其视为一个关系节点。所以在图 $\mathcal{G}^s$ 中关系节点的数量是 $\frac{|\hat{\mathcal{E}}|(|\hat{\mathcal{E}}|-1)}{2}$。

利用图 $\mathcal{G}^s$ 可以显式建模多个实体类型和关系类型的交互。对于图中所有的节点，我们有一个初始输入节点向量矩阵 $\mathbf{H}$。给定任意的一个关系 r_{12} 及对应的两个实体 e_1, e_2，本文使用 $\mathbf{H}_{r_{12}}$ 表示的关系节点向量，使用 $\mathbf{H}_{e_1}$，$\mathbf{H}_{e_2}$ 表示对应的实体节点向量。

接下来，我们在实体节点和关系节点之间连边。对于图中的边，我们将每个关系节点连接到其对应的两个实体节点，而不是直接连接任何实体（关系）节点。即本文关注二分图。原因有两个：

1. 对于识别某一个实体的类型，句子中可能存在很多的实体，并不是所有实体

[2] 本文还尝试了 biLSTM-CRF [68] 模型作为实体模型，但是从实验结果看并没有获得性能上的提升。

[3] 对于每个候选关系，第一个实体总是在第二个实体的左边。为了考虑到方向信息，本文使用总共 $2\mathcal{T}_r + 1$ 个关系类型。额外的类型是 `None`，表示两个实体对之间没有关系。

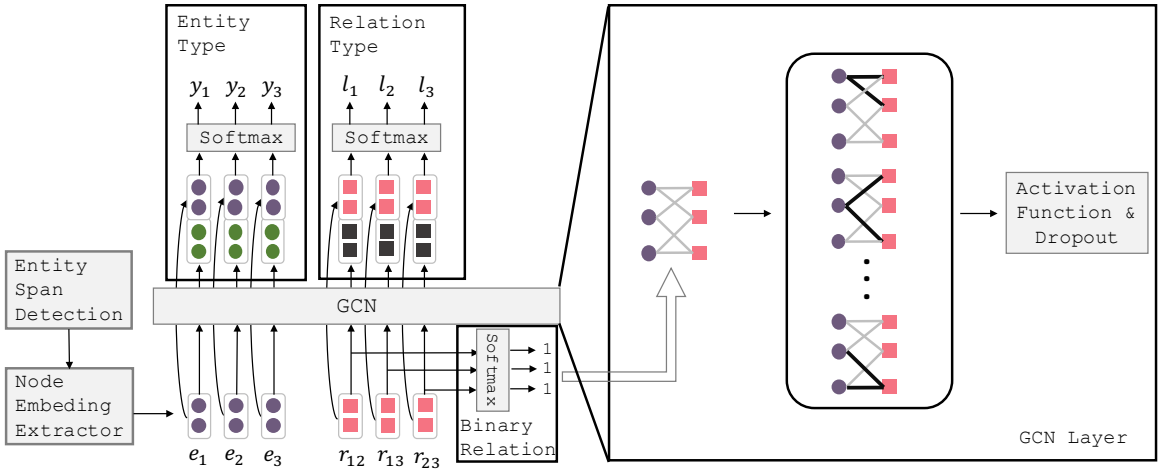

图 6.3 基于 GCN 的联合实体关系抽取网络结构

的信息对判断该实体的类型都有帮助，我们认为存在关系的实体之间的相互帮助会更大一些。即关系节点是实体节点之间的桥梁，反之亦然。

2. GCN 不适用于完全连接的图，因为在完全图中，GCN 的操作会变为很简单的计算，等价于只有一个卷积核的卷积神经网络。

这意味着对于实体节点 e，观察其他实体的唯一方法是通过其参与的关系节点。给定关系节点 r_{12} 及其两个实体节点 e_1, e_2，我们添加两条边：一个是 e_1 和 r_{12} 之间的边，另一个是 e_2 和 r_{12} 之间的边。这样方式的二分图本文称之为**静态图**。

为了进一步利用图的结构（某种先验知识）而不是使用静态图，本文还研究了 **动态图**，因为它可以用来修剪冗余边。如果两个实体存在关系，我们可以在关系节点和两个实体节点之间添加两条边，相反，如果两个实体没有关系，我们分别保留两个实体节点和关系节点，并不添加任何一条边。为此，本文引入了**二元关系分类**任务。该任务旨在预测实体对之间是否存在某种关系（忽略特定的关系类型）。本文构建了一个二元关系模型，基于关系节点向量，预测标签集合为 $\{0,1\}$，0 表示没有关系，1 表示存在关系。给定句子 s 中的关系节点 r_{ij}，为了获得二元关系标签 $\hat{b}$ 的后验，本文对关系向量 $\mathbf{H}_{r_{ij}}$ 使用 softmax 函数。

$$P(\hat{b}|r_{ij}, s) = \text{Softmax}(\mathbf{W}_{\text{bin}}\mathbf{H}_{r_{ij}}),$$

其中 $\mathbf{W}_{\text{bin}}$ 是模型参数。二分关系分类的训练目标是最小化

$$\mathcal{L}_{\text{bin}} = -\sum_{r_{ij}} \frac{\log P(\hat{b} = b|r_{ij}, s)}{\#\text{ candidate relations } r_{ij}}, \tag{6.4}$$

其中正确标签 b 可以从原始关系标注中转换得到。邻接矩阵 $\mathbf{A}$ 中的值根据以下规

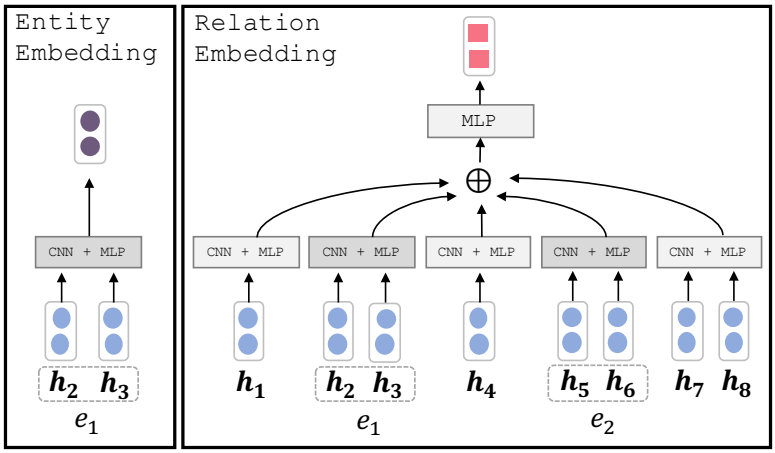

图 6.4 计算实体节点向量和关系节点向量

则确定：

- 如果 $P(\hat{b}=1|r_{ij},s)>0.5$，我们把实体节点 e_i 和关系节点 r_{ij} 之间边的权重置为 1.0，同样，也把实体节点 e_j 和关系节点 r_{ij} 之间边的权重置为 1.0；
- **A** 中对角线元素的值都置为 1.0；
- 其它的都置为 0.0。

以上邻接矩阵 **A** 中的取值只有 0 和 1 两种情况，本文称之为**硬**值 **A**。此外，本文还在实验中尝试**软**值的 **A**，即实体节点 e_i 和关系节点 r_{ij} 之间边的权重设置为概率 $P(\hat{b}=1|r_{ij},s)$。同样，实体节点 e_j 和关系节点 r_{ij} 之间边的权重也设置为概率 $P(\hat{b}=1|r_{ij},s)$，对角线元素设置为 1.0。

接下来，本文将介绍如何计算图中两种类型的节点向量：实体节点向量和关系节点向量。

实体节点向量

给定实体 $e\in\hat{\mathcal{E}}$，对于每个单词 $w_i\in e$，本文利用 biLSTM 隐层向量 $\mathbf{h}_i$ 作为该单词的表示。然后，本文在向量序列 $\{\mathbf{h}_i|w_i\in e\}$ 上使用 CNN（卷积层后接最大池化层）抽取 d 维的实体节点向量 $\mathbf{H}_e$（**H** 是6.3.1 节提到过的节点向量矩阵），如图 6.4 左侧部分。

关系节点向量

给定候选关系 r_{12}，本文抽取两种类型的特征，即关于实体 e_1,e_2 本身的特征以及关于实体对 e_1,e_2 的上下文特征。

- 对于实体 e_1 和 e_2 的本身的特征，本文使用实体节点向量 $\mathbf{H}_{e_1}$ 和 $\mathbf{H}_{e_2}$。
- 对于实体对的上下文特征，本文通过抽取 e_1 和 e_2 之间的单词特征、左边的单词特征（$\mathbf{f}_{\text{left}}$）、右边的单词特征（$\mathbf{f}_{\text{right}}$）得到三个特征向量。与实体节点向量类似，本文使用另外一个 CNN 作为这三个特征向量的抽取器。

最后，将五个特征向量拼接成单个向量。为了获得 d 维关系节点向量，本文在该单个向量上使用 MLP，如图 6.4的右侧部分所示。

6.3.5 联合类型推理

在构建实体-关系二分图之后，本文将此图输入到多层 GCN 中，以获得节点级别输出 $\mathbf{Z}$。矩阵 $\mathbf{Z}$ 中的每一行都是实体或关系节点表示，这些表示融合了图中的其他节点的信息。然后，每个节点的最终表示 $\mathbf{F}$ 由输入节点向量 $\mathbf{H}$ 和节点级别输出 $\mathbf{Z}$（$\mathbf{H}, \mathbf{Z}$ 和 $\mathbf{F}$ 是矩阵）拼接而成。

给定实体节点 e_i 和关系节点 r_{ij}，为了预测实体节点类型和关系节点类型，本文用两个 softmax 函数得到两种节点类型的概率分布

$$P(\hat{y}|e_i, s) = \text{Softmax}(\mathbf{W}_{\text{ent}}\mathbf{F}_{e_i}),$$

$$P(\hat{l}|r_{ij}, s) = \text{Softmax}(\mathbf{W}_{\text{rel}}\mathbf{F}_{r_{ij}}),$$

其中 $\mathbf{W}_{\text{ent}}$，$\mathbf{W}_{\text{rel}}$ 是模型参数。联合类型推理任务的训练目标是最小化

$$\mathcal{L}_{\text{ent}} = -\frac{1}{|\hat{\mathcal{E}}|}\sum_{e_i \in \hat{\mathcal{E}}} \log P(\hat{y} = y|e_i, s), \tag{6.5}$$

$$\mathcal{L}_{\text{rel}} = -\sum_{r_{ij}} \frac{\log P(\hat{l} = l|r_{ij}, s)}{\#\text{ candidate relations } r_{ij}}, \tag{6.6}$$

其中正确标注 y, l 可以从原始标注中得到。整个联合实体关系抽取系统的计算过程如图 6.3 所示。

6.3.6 训练

为了训练联合模型，本文优化了目标函数 $\mathcal{L} = \mathcal{L}_{\text{span}} + \mathcal{L}_{\text{bin}} + \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{rel}}$。本文在实体模型中使用策略采样 [12, 128]。模型使用 dropout 进行正则化并使用 Adadelta 优化器 [216]。本文使用验证集选择模型：在固定数量的训练轮数内，挑选出在验证

表 6.1 ACE05 测试集结果.

Model	Entity			Relation		
	P	R	F	P	R	F
L&J [2014]	85.2	76.9	80.8	65.4	39.8	49.5
Zhang [2017]	-	-	83.5	-	-	57.5
Sun [2018]	83.9	83.2	83.6	64.9	**55.1**	**59.6**
M&B [2016]	82.9	**83.9**	83.4	57.2	54.0	55.6
K&C [2017]	84.0	81.3	82.6	55.5	51.8	53.6
NN	85.7	82.1	83.9	65.6	50.7	57.2
GCN	**86.1**	82.4	**84.2**	**68.1**	52.3	59.1

集上具有最好关系抽取性能的模型进行测试。本文使用 100 维的 glove 向量 [142] 作为词向量初始化。隐层单元个数和节点向量维度都设置为 128。对于网络中的所有 CNN，卷积核大小为 2 和 3，输出通道数为 25。

6.4 实验

6.4.1 设置

本文在 ACE05 数据集上评估本文提出的框架。有关 ACE05 数据集的详细情况请参考 2.3.3 节。

本文评价模型结果用精确率（P），召回率（R）和 F 分数。具体地，如果输出实体的类型 y 和头部区域 e 是正确的，则输出实体 (e, y) 是正确的，如果输出关系 r 的 (e_1, y_1, e_2, y_2, l) 是正确的，则输出关系 r 是正确的（精确匹配）。

在本章工作中，默认设置“GCN”是具有动态硬邻接矩阵的基于 GCN 的单层联合模型，它在 ACE05 数据集上取得了最好的关系抽取性能。

6.4.2 在 ACE05 上端到端结果

首先，本文将提出的模型与已有的实体关系抽取系统进行比较（表 6.1）。总的来说，与现有的联合模型相比，本文提出的“GCN”取得了 84.2% 的最好实体识别性能。对于关系抽取性能，“GCN”优于其它所有联合模型，除了文献 [175] 的联合模型。与本文的基础神经网络“NN”相比，“GCN”在实体识别和关系抽取方面的性能均有很大的提高。这些证明了“GCN”可以从句子中捕获多个实体类型和关系类型的信息。

与采用风险最小化训练方法 [175] 相比，“GCN”具有更好的实体识别性能和相当的关系抽取性能。与现有的联合解码系统不同，本文没有使用复杂的联合解

表 6.2 不同设置下 ACE05 数据集的结果

Model	Entity			Relation			Entity Span			Binary Relation		
	P	R	F	P	R	F	P	R	F	P	R	F
Sun （NN）[2018]	84.0	**82.9**	83.4	59.5	**56.3**	57.8	-	-	-	-	-	-
NN	85.7	82.1	83.9	65.6	50.7	57.2	**91.2**	89.6	90.4	-	-	-
GCN （static）	85.0	82.6	83.8	66.6	51.3	57.8	90.8	**90.2**	**90.5**	-	-	-
GCN （dynamic + soft）	85.3	82.3	83.8	67.3	51.6	58.5	90.8	**90.2**	**90.5**	77.3	**56.4**	65.2
GCN （dynamic + hard）	**86.1**	82.4	**84.2**	**68.1**	52.3	**59.1**	**91.2**	89.5	90.4	**78.2**	56.3	**65.4**

码算法，如集束搜索 [102]、全局归一化 [221] 和最小风险训练 [175]。本文的模型仅依赖于共享参数，类似与文献 [81, 128] 值得注意的是，与所有其他方法相比，本文的 “GCN” 的精确率很高。我们认为可能的原因是 GCN 具有强大的特征抽取能力，从而得到丰富的实体节点和关系节点的特征表示。

接下来，本文在不同的设置下评估我们的模型。如 6.3.4 节所述，本文提出三种类型的图：“GCN （static）”、“GCN （dynamic + hard）” 和 “GCN （dynamic + soft）”。表 6.2 最后三行列出了对应的性能。关于表 6.2，我们得出以下几个结论。

- 与没有风险最小化训练的联合神经网络模型 “Sun（NN）” 模型相比 [175]，本文的 “NN” 在实体识别上的性能要高 0.5%。一个可能的原因是实体类型模型和关系类型模型共享更多参数（实体 CNN + MLP 参数），而 “Sun（NN）” 仅共享 biLSTM 隐层状态。但是，本文的 “NN” 在关系上的表现低 0.6%。一个可能的原因是我们没有使用输出实体类型的特征进行关系类型分类。

- 在引入图卷积网络之后，三个基于 GCN 的模型实体识别和关系抽取的性能都得到了提高。具体而言，“GCN（static）” 在关系抽取任务的性能略有改善。“GCN（dynamic + soft）” 在关系抽取上得到了 0.7% 的提高，在实体识别上的性能几乎没有变化。“GCN（dynamic + hard）提高了实体识别性能（0.4%），并在关系抽取性能方面实现了大幅提升（1.9%）。在关系抽取结果方面，它与当前最好的联合模型 [175] 相当。这些观察结果表明，提出的联合模型对于实体和关系的联合类型推断是有效的，并且也显示了所提出动态图的合理性。

- 对于本文提出的四个模型（表 6.2最后 4 行），实体边界检测和二元关系分类的性能都比较接近。一个可能的原因是这些都是更粗粒度的任务，所有模型都可以轻松抽取到有效的特征。值得注意的是，二元关系分类的表现还不是很好，本文的动态图的构建依赖于二元关系检测任务的输出。如果二元关系检测任务的性能可以继续提升，我们相信会进一步改进联合模型的性能，我们把它作为未来的工作。

第三，本文列出 GCN 层数的影响（表 6.3）。本文以 “GCN（dynamic + hard）” 为例。通常，这四个任务的性能对 GCN 层数量不敏感。特别是实体边界，实体和关

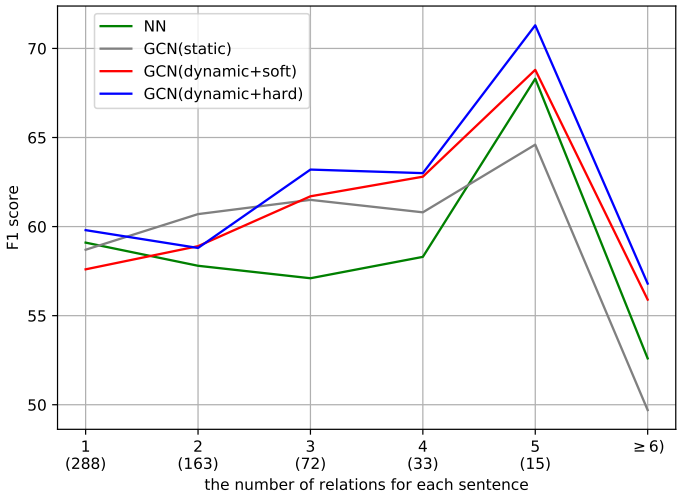

图 6.5 不同关系数量的句子对应的 F 分数

系的表现在 1.0% 内波动，二元关系在 1.4% 波动。有趣的是，我们发现一层 GCN 模型尽管在其他三个任务的表现并不是最好的，但它取得了最好的关系抽取性能。一个可能的原因是所有模型彼此密切相关，但是它们在这种联合环境中如何相互影响仍然是一个开放的问题。

表 6.3 在 ACE05 验证集上不同 GCN 层数的结果

	1-layer	2-layer	3-layer
F of Entity Span	90.4	90.5	**90.7**
F of Binary Relation	61.5	**62.9**	62.8
F of Entity	81.6	82.1	**82.2**
F of Relation	**53.8**	53.5	53.6

第四，本文检查关于不同关系数量句子的性能表现（图 6.5）。总体来说，当关系数量大于 2 时，基于 GCN 的模型几乎优于 NN。它证明了基于 GCN 的模型更适合处理具有多个关系的句子。在现实应用中这样的多关系数据非常普遍，我们认为本文提出的模型在复杂的多关系数据集上表现会更好。

最后，本文在一些具体的例子中将 “NN” 模型与 “GCN” 模型进行比较，如表 6.4所示。♡ 是正确标注结果，♣，♠ 分别是 “NN”，“GCN” 的模式输出结果。对于样例 S1，“NN” 正确抽取了 “[legislature]$^{\text{ORG}}$” 和 “[chairman]$^{\text{PER}}$” 之间的 ORG-AFF 关系，但是错误判断了 “[legislature]$^{\text{ORG}}$” 和 “[north korea]$^{\text{GPE}}$” 之间的 GEN-AFF 关系。对于样例 S2，“NN” 没有检测出 PART-WHOLE 关系，然而 “GCN” 正确找到了。这两个样例说明了 “GCN” 更适合处理多个关系的句子。对于样例 S3，“GCN” 识别了 “[units]$^{\text{PER}}$” 和 “[captial]$^{\text{GPE}}$” 之间的 PHYS 关系，但是 “NN” 没有找到及时实体已经被正确识别。然而，两个模型都没有识别 “[units]$^{\text{PER}}$” 和 “[weapons]$^{\text{WEA}}$” 之间的 ART 关系。我们认为使用更强大的图神经网络可能会在这种情况下有所帮助。

表 6.4 模型输出结果样例

S1	the [british]$^{\text{GPE:}\heartsuit\clubsuit\spadesuit}_{\text{GEN-AFF-2:}\heartsuit\clubsuit\spadesuit}$ [arm]$^{\text{ORG:}\heartsuit\clubsuit\spadesuit}_{\text{PART-WHOLE-1:}\heartsuit\spadesuit\text{|GEN-AFF-1:}\heartsuit\clubsuit\spadesuit}$ of french distributors [pathe]$^{\text{ORG:}\heartsuit\clubsuit\spadesuit}_{\text{PART-WHOLE-2:}\heartsuit\spadesuit}$ to show four releases .
S2	…[chairman]$^{\text{PER:}\heartsuit\clubsuit\spadesuit}_{\text{ORG-AFF-1:}\heartsuit\clubsuit\spadesuit}$ of [north korea]$^{\text{GPE:}\heartsuit\clubsuit\spadesuit}_{\text{PART-WHOLE-2:}\heartsuit\spadesuit\text{|GEN-AFF-2:}\clubsuit}$'s [legislature]$^{\text{ORG:}\heartsuit\clubsuit\spadesuit}_{\text{PART-WHOLE-1:}\heartsuit\spadesuit\text{|ORG-AFF-2:}\heartsuit\clubsuit\spadesuit\text{|GEN-AFF-1:}\clubsuit}$, the supreme people 's assembly .
S3	a red line may have been drawn around the [capital]$^{\text{GPE:}\heartsuit\clubsuit\spadesuit}_{\text{PHYS-2:}\heartsuit\spadesuit}$ with [republican gurad]$^{\text{ORG:}\heartsuit\clubsuit\spadesuit}_{\text{ORG-AFF-2:}\heartsuit\clubsuit\spadesuit}$ [units]$^{\text{PER:}\heartsuit\clubsuit\spadesuit}_{\text{PHYS-1:}\heartsuit\spadesuit\text{|ORG-AFF-1:}\heartsuit\clubsuit\spadesuit\text{|ART-1:}\heartsuit}$ ordered to use chemical [weapons]$^{\text{WEA:}\heartsuit\clubsuit\spadesuit}_{\text{ART-2:}\heartsuit}$ once u.s. and allied troops cross it .

表 6.5 在 ACE05 数据集上给定正确实体后关系抽取的结果

Model	**Relation**		
	P	R	F
M&B [2016]	**70.1**	61.2	65.3
C&M [2019]	69.7	59.5	64.2
NN	68.5	62.8	65.5
GCN （static）	69.1	63.8	66.4
GCN （dynamic + soft）	68.7	63.4	65.9
GCN （dynamic + hard）	68.7	**65.4**	**67.0**

6.4.3 在 ACE05 上给定正确实体的结果

本文在表 6.5 列出了给定实体后在 ACE05 数据集上关系抽取的结果。为了和已有的模型进行比较 [37, 128]，本文使用相同的数据划分方式。在这个实验中，本文没有细微调整 GCN 的超参数，而是使用和端到端模型相同的超参数。用于比较的基准系统是文献 [37, 128]。

总的来说，与基于依存树的目前最好的模型相比，本文的“NN”取得相当的关系抽取性能 [128]。它表明基于 CNN 的神经网络能够提取更强大的特征来帮助关系分类任务。添加 GCN 后，基于 GCN 的模型更进一步提升了关系抽取的性能。这表明本文所提出的模型在没有借助任何外部句法工具的情况下可以取得比较大的提升，同时证明了 GCN 模型在关系抽取中的有效性 [4]。

6.5 总结

本文将实体关系抽取任务分为两个子任务：实体边界检测和实体关系类型推断。同时，本文提出了一种基于 GCN 的新颖简洁的联合模型用于实体关系类型推断任务。与现有的联合方法相比，它提供了一种在句子中明确捕获多个实体类型和关系类型上的交互的新方法。为了使用 GCN，本文提出实体-关系二分图，并设

[4]为简单起见，本文没有在模型中明确使用实体类型特征。我们相信在使用这些特征时性能会有进一步的提升。

计了二元关系分类任务用于修建图中冗余的边。在 ACE05 数据集上的实验表明了该方法在端到端的实体关系抽取任务和给定正确实体的关系抽取任务均可获得不错的性能。

第七章　结语与展望

实体关系抽取任务是自然语言处理领域一大重要任务，一直是工业界和学术界的研究热点，为各种下游任务以及直接应用提供了支持。此外，实体关系抽取为自然语言理解提供了基础，是计算机理解自然文本的主要方式。因此，实体关系抽取的研究有着非常重要的理论价值。

当今实体关系抽取主要面临着两个方面的挑战：一是数据，二是联合模型。在数据方面，由于人工标注数据的费时费力，想要得到大规模的标注数据通常不太现实。基于此种考虑，远程监督的方法被提出来。刚开始的远程监督是利用知识库和大规模自由文本对齐，从而自动生成训练数据。然而由于是启发式的对齐，使得训练数据中存在大量的噪音样本。针对这些问题，本文主要提出两种方法来应对：

1. 针对远程监督中的噪音问题，本文研究了利用少量人工标注的高质量异构数据集缓解此问题，提出了一个基于多任务学习的融合框架，可以实现异构数据集之间的知识迁移，从而提高远程监督学习的性能。

2. 由于某些领域不存在知识库，比如在线评论。本文提出利用语言学规则进行远程监督，同时利用神经网络分类器在这些自动得到的数据集上进行训练，最终利用新得到的分类器可以找出规则覆盖不到的关系。

实体关系抽取包含两个子任务，简单的基于流水线的方法会存在错误传播等明显的缺点。为了缓解这个问题，联合模型的方式得到近些年来的关注。联合模型的难点是如何加强两个子模型之间的交互。针对这个问题，本文也提出两个解决方案来应对：

1. 本文提出基于风险最小化训练方法的联合模型，一方面可以优化全局的损失函数，另一方面同时加强了实体模型和关系模型之间的交互，提供了一种新的方式进行联合学习。

2. 本文构建了一个新颖的实体-关系二分图，并使用图神经网络对实体和关系同时建模，可以同时捕获实体类型和关系类型的信息，直接在特征层面加强了实体模型和关系模型之间的交互。

本文提出的所有方法，均设计了实验证明其有效性，同时对结果也做了详细的分析。

本文的工作仍有很多可以改进的地方，在此基础上，列出以下未来可能研究的方向。

1. 预训练：最近一段时间预训练模型（比如 Elmo，BERT，GPT，XLNet 等）在各种自然语言处理任务中取得了非常惊人的进步。主要思想是在大规模无监督文本中训练语言模型等任务，从而得到一个预训练好的句子编码器，针对不同的任务，只需要接上对应的解码器，然后微调整个网络。在实体关系抽取任务中，除了句子级别的编码器之外，关系抽取的编码器也尤为重要，而在这点之前提到的预训练模型很难使用。针对此点，可以考虑设计其他的预训练任务，从而使得关系抽取的编码器更加强大。

2. 负样本：在实体模型抽取到实体之后，任意两个实体对形成候选关系，但是候选关系大部分都是负样本，可能只有少量的正样本。通常来说，负样本的数量是正样本的 10 倍以上。这在训练过程中存在正负样本严重不平衡的问题。如何选择合适的负样本对最后训练得到的模型尤为重要。针对此点，可以考虑设计一个模型用来过滤负样本。

3. 子模型训练：通常在联合模型的训练中，只使用一个优化器。然而实体任务和关系任务本质上是两类任务。对于共享的参数和每个任务特有的参数，可能有着不同的特性。所以，针对这种联合模型，可以考虑设计一种训练方法将这些因素考虑进来。

参考文献

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

[2] Steven Abney. Understanding the yarowsky algorithm. *Computational Linguistics*, 30(3):365–395, 2004.

[3] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM, 2000.

[4] Chinatsu Aone, Lauren Halverson, Tom Hampton, and Mila Ramos-Santacruz. Sra: Description of the ie2 system used for muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*, 1998.

[5] Douglas E Appelt, Jerry R Hobbs, John Bear, David Israel, Megumi Kameyama, David Martin, Karen Myers, and Mabry Tyson. Sri international fastus system: Muc-6 test results and analysis. In *Proceedings of the 6th conference on Message understanding*, pages 237–248. Association for Computational Linguistics, 1995.

[6] Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. *arXiv preprint arXiv:1802.09913*, 2018.

[7] Shiqi Shen Ayana, Zhiyuan Liu, and Maosong Sun. Neural headline generation with minimum risk training. *arXiv preprint arXiv:1604.01904*, 2016.

[8] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *Ijcai*, volume 7, pages 2670–2676, 2007.

[9] Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Simaan. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967, 2017.

[10] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[11] Oliver Bender, Franz Josef Och, and Hermann Ney. Maximum entropy models for named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 148–151. Association for

Computational Linguistics, 2003.

[12] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.

[13] Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, 2016.

[14] Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high-performance learning name-finder. *arXiv preprint cmp-lg/9803003*, 1998.

[15] Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. An algorithm that learns what's in a name. *Machine learning*, 34(1-3):211–231, 1999.

[16] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[17] William J Black, Fabio Rinaldi, and David Mowatt. Facile: Description of the ne system used for muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*, 1998.

[18] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. Citeseer, 1998.

[19] Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. Nyu: Description of the mene named entity system as used in muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*, 1998.

[20] Sergey Brin. Extracting patterns and relations from the world wide web. In *International Workshop on The World Wide Web and Databases*, pages 172–183. Springer, 1998.

[21] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. Association for Computational Linguistics, 2010.

[22] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[23] Razvan C Bunescu and Raymond J Mooney. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 724–731. Association for Computational Linguistics, 2005.

[24] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637, 2018.

[25] Rui Cai, Xiaodong Zhang, and Houfeng Wang. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 756–765, 2016.

[26] David Campos, Sérgio Matos, and José Luís Oliveira. Biomedical named entity recognition: a survey of machine-learning tools. *Theory and Applications for Advanced Text Mining*, pages 175–195, 2012.

[27] R Caruana. Multitask learning: A knowledge-based source of inductive bias. machine learning. 1997.

[28] Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. An investigation of recurrent neural architectures for drug name recognition. *arXiv preprint arXiv:1609.07585*, 2016.

[29] Yee Seng Chan and Dan Roth. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 551–560. Association for Computational Linguistics, 2011.

[30] Hongshen Chen, Yue Zhang, and Qun Liu. Neural network for heterogeneous annotations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 731–741, 2016.

[31] Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. Unsupervised feature selection for relation extraction. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*, 2005.

[32] Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-criteria learning for chinese word segmentation. *arXiv preprint arXiv:1704.07556*, 2017.

[33] Hai Leong Chieu and Hwee Tou Ng. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

[34] Laura Chiticariu, Yunyao Li, and Frederick R Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 827–832, 2013.

[35] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

[36] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[37] Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. A walk-based model on entity graphs for relation extraction. *arXiv preprint arXiv:1902.07023*, 2019.

[38] Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[39] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[40] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

[41] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, page 423. Association for Computational Linguistics, 2004.

[42] James Curran and Stephen Clark. Language independent ner using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 164–167, 2003.

[43] Oier Lopez De Lacalle and Mirella Lapata. Unsupervised relation extraction with general domain knowledge. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 415–425, 2013.

[44] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016.

[45] Lingjia Deng and Janyce Wiebe. Mpqa 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1323–1328, 2015.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[47] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3): 361–365, 1996.

[48] Salah El Hihi and Yoshua Bengio. Hierarchical recurrent neural networks for long-term dependencies. In *Advances in neural information processing systems*, pages 493–499, 1996.

[49] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[50] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165 (1):91–134, 2005.

[51] Murthy Ganapathibhotla and Bing Liu. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 241–248. Association for Computational Linguistics, 2008.

[52] Kevin Gimpel and Noah A Smith. Softmax-margin crfs: Training log-linear models with cost functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 733–736. Association for Computational Linguistics, 2010.

[53] Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 13–18. Association for Computational Linguistics, 2007.

[54] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[55] Matthew R Gormley, Mo Yu, and Mark Dredze. Improved relation extraction with feature-rich compositional embedding models. *arXiv preprint arXiv:1505.02419*, 2015.

[56] Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 427–434. Association for Computational Linguistics, 2005.

[57] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

[58] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel:a large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *EMNLP*, 2018.

[59] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, page 415. Association for Computational Linguistics, 2004.

[60] Xiaodong He and Li Deng. Maximum expected bleu training of phrase and lexicon translation models. In *Proceedings of the 50th Annual Meeting of the Association*

for Computational Linguistics: Long Papers-Volume 1, pages 292–301. Association for Computational Linguistics, 2012.

[61] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[62] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics, 2009.

[63] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[64] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics, 2011.

[65] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics, 2011.

[66] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[67] Xuanjing Huang et al. Attention-based convolutional neural network for semantic relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2526–2536, 2016.

[68] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[69] Kevin Humphreys, Robert Gaizauskas, Saliha Azzam, Charles Huyck, Brian Mitchell, Hamish Cunningham, and Yorick Wilks. University of sheffield: Description of the lasie-ii system as used for muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*, 1998.

[70] Hideki Isozaki and Hideto Kazawa. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Compu-*

tational linguistics-Volume 1, pages 1–7. Association for Computational Linguistics, 2002.

[71] Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep K Ravikumar. A dirty model for multi-task learning. In *Advances in neural information processing systems*, pages 964–972, 2010.

[72] Soufian Jebbara and Philipp Cimiano. Aspect-based relational sentiment analysis using a stacked neural network architecture. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pages 1123–1131. IOS Press, 2016.

[73] Zongcheng Ji, Aixin Sun, Gao Cong, and Jialong Han. Joint recognition and linking of fine-grained locations from tweets. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1271–1281. International World Wide Web Conferences Steering Committee, 2016.

[74] Jing Jiang and ChengXiang Zhai. A systematic exploration of the feature space for relation extraction. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 113–120, 2007.

[75] Wenbin Jiang, Liang Huang, and Qun Liu. Automatic adaptation of annotation standards: Chinese word segmentation and pos tagging: a case study. In *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: Volume 1-Volume 1*, pages 522–530. Association for Computational Linguistics, 2009.

[76] Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang. Relation extraction with multi-instance multi-label convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1471–1480, 2016.

[77] Richard Johansson and Alessandro Moschitti. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509, 2013.

[78] Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 22. Association for Computational Linguistics, 2004.

[79] Jagat Narain Kapur. *Maximum-entropy models in science and engineering*. John Wiley & Sons, 1989.

[80] Arzoo Katiyar and Claire Cardie. Investigating lstms for joint extraction of opinion entities and relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, 2016.

[81] Arzoo Katiyar and Claire Cardie. Going out on a limb: Joint extraction of entity mentions and relations without dependency trees. In *Proceedings of the 55th An-*

nual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 917–928, 2017.

[82] Kazuya Kawakami. Supervised sequence labelling with recurrent neural networks. *Ph. D. thesis*, 2008.

[83] Jun’ichi Kazama and Kentaro Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 698–707, 2007.

[84] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[85] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[86] Roman Klinger and Philipp Cimiano. The usage review corpus for fine-grained, multi-lingual opinion analysis. In *Proceedings of the Language Resources and Evaluation Conference*, 2014.

[87] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1065–1074, 2007.

[88] Julia Kreutzer, Artem Sokolov, and Stefan Riezler. Bandit structured prediction for neural sequence-to-sequence learning. *arXiv preprint arXiv:1704.06497*, 2017.

[89] Vijay Krishnan and Christopher D Manning. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1121–1128. Association for Computational Linguistics, 2006.

[90] GR Krupka and KH IsoQuest. Description of the nerowl extractor system as used for muc-7. In *Proceedings of the 7th Message Understanding Conference, Virginia*, pages 21–28, 2005.

[91] Shantanu Kumar. A survey of deep learning methods for relation extraction. *arXiv preprint arXiv:1705.03645*, 2017.

[92] Onur Kuru, Ozan Arkan Can, and Deniz Yuret. Charner: Character-level named entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 911–921, 2016.

[93] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[94] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recogni-

tion. *arXiv preprint arXiv:1603.01360*, 2016.

[95] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10): 1995, 1995.

[96] Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. Mit at semeval-2017 task 10: Relation extraction with convolutional neural networks. *arXiv preprint arXiv:1704.01523*, 2017.

[97] Joohong Lee, Sangwoo Seo, and Yong Suk Choi. Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing. *Symmetry*, 11(6):785, 2019.

[98] Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. Structure-aware review mining and summarization. In *Proceedings of the 23rd international conference on computational linguistics*, pages 653–661. Association for Computational Linguistics, 2010.

[99] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *arXiv preprint arXiv:1812.09449*, 2018.

[100] Jiwei Li, Minh-Thang Luong, Dan Jurafsky, and Eudard Hovy. When are tree structures necessary for deep learning of representations? *arXiv preprint arXiv:1503.00185*, 2015.

[101] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[102] Qi Li and Heng Ji. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–412, 2014.

[103] Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. Component-enhanced chinese character embeddings. *arXiv preprint arXiv:1508.06669*, 2015.

[104] Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham. Svm based learning system for information extraction. In *International Workshop on Deterministic and Statistical Methods in Machine Learning*, pages 319–339. Springer, 2004.

[105] Wenhui Liao and Sriharsha Veeramachaneni. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 58–65. Association for Computational Linguistics, 2009.

[106] Nut Limsopatham and Nigel Henry Collier. Bidirectional lstm for named entity recognition in twitter messages. 2016.

[107] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In *Proceedings of the*

54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2124–2133, 2016.

[108] ChunYang Liu, WenBo Sun, WenHan Chao, and Wanxiang Che. Convolution neural network for relation extraction. In *International Conference on Advanced Data Mining and Applications*, pages 231–242. Springer, 2013.

[109] Kang Liu, Liheng Xu, and Jun Zhao. Extracting opinion targets and opinion words from online reviews with graph co-ranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 314–324, 2014.

[110] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*, 2017.

[111] Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries. *Information*, 6(4):848–865, 2015.

[112] Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia. Neural relation extraction via inner-sentence noise reduction and transfer learning. *arXiv preprint arXiv:1808.06738*, 2018.

[113] Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1790–1795, 2017.

[114] Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 359–367. Association for Computational Linguistics, 2011.

[115] Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. A dependency-based neural network for relation classification. *arXiv preprint arXiv:1507.04646*, 2015.

[116] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444, 2002.

[117] Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM, 2011.

[118] Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. *arXiv preprint arXiv:1705.03995*, 2017.

[119] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

[120] Diego Marcheggiani and Ivan Titov. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, 2017.

[121] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2015.

[122] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics, 2003.

[123] Paul McNamee and James Mayfield. Entity extraction without language-specific resources. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–4. Association for Computational Linguistics, 2002.

[124] Andrei Mikheev. A knowledge-free method for capitalized word disambiguation. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 159–166. Association for Computational Linguistics, 1999.

[125] Andrei Mikheev, Marc Moens, and Claire Grover. Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics, 1999.

[126] Tomáš Mikolov. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*, 80, 2012.

[127] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *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: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[128] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*, 2016.

[129] Makoto Miwa and Yutaka Sasaki. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869, 2014.

[130] Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. A rich feature vector for protein-protein interaction extraction from multiple corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing:*

Volume 1-Volume 1, pages 121–130. Association for Computational Linguistics, 2009.

[131] Raymond J Mooney and Razvan C Bunescu. Subsequence kernels for relation extraction. In *Advances in neural information processing systems*, pages 171–178, 2006.

[132] Arjun Mukherjee and Bing Liu. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1*, pages 339–348. Association for Computational Linguistics, 2012.

[133] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[134] David Nadeau, Peter D Turney, and Stan Matwin. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Conference of the Canadian society for computational studies of intelligence*, pages 266–277. Springer, 2006.

[135] Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. Reinforcement learning for bandit neural machine translation with simulated human feedback. *arXiv preprint arXiv:1707.07402*, 2017.

[136] Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, 2015.

[137] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics, 2003.

[138] Sachin Pawar, Girish K Palshikar, and Pushpak Bhattacharyya. Relation extraction: A survey. *arXiv preprint arXiv:1712.05191*, 2017.

[139] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[140] Hao Peng, Sam Thomson, and Noah A Smith. Deep multitask learning for semantic dependency parsing. *arXiv preprint arXiv:1704.06855*, 2017.

[141] Nanyun Peng and Mark Dredze. Multi-task multi-domain representation learning for sequence tagging. *arXiv preprint arXiv:1608.02689*, 2016.

[142] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[143] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[144] Thai-Hoang Pham and Phuong Le-Hong. End-to-end recurrent neural network models for vietnamese named entity recognition: Word-level vs. character-level. In *International Conference of the Pacific Association for Computational Linguistics*, pages 219–232. Springer, 2017.

[145] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, 2015.

[146] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30, 2016.

[147] Pengda Qin, Weiran Xu, and William Yang Wang. Dsgan: generative adversarial training for distant supervision relation extraction. *arXiv preprint arXiv:1805.09929*, 2018.

[148] Pengda Qin, Weiran Xu, and William Yang Wang. Robust distant supervision relation extraction via deep reinforcement learning. *arXiv preprint arXiv:1805.09927*, 2018.

[149] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1): 9–27, 2011.

[150] Xipeng Qiu, Jiayi Zhao, and Xuanjing Huang. Joint chinese word segmentation and pos tagging on heterogeneous annotated corpora with multiple task learning. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 658–668, 2013.

[151] Changqin Quan, Meng Wang, and Fuji Ren. An unsupervised text mining method for relation extraction from biomedical literature. *PloS one*, 9(7):e102039, 2014.

[152] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[153] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736. ACM, 2006.

[154] Yael Ravin and Nina Wacholder. *Extracting names from natural-language text*. Citeseer, 1997.

[155] Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. Cotype: Joint extraction of typed entities and relations

with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1015–1024. International World Wide Web Conferences Steering Committee, 2017.

[156] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

[157] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112, 2003.

[158] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[159] Tim Rocktäschel, Michael Weidlich, and Ulf Leser. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640, 2012.

[160] Tim Rocktäschel, Torsten Huber, Michael Weidlich, and Ulf Leser. Wbi-ner: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 356–363, 2013.

[161] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

[162] Cicero Nogueira dos Santos and Victor Guimaraes. Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*, 2015.

[163] Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580*, 2015.

[164] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[165] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 107–110, 2004.

[166] Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. End-to-end structure-aware convolutional networks for knowledge base completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3060–3067, 2019.

[167] Yan Shao, Christian Hardmeier, and Joakim Nivre. Multilingual named entity recognition using hybrid neural networks. In *The Sixth Swedish Language Tech-*

nology Conference (SLTC), 2016.

[168] Rahul Sharnagat. Named entity recognition: A literature survey. *Center For Indian Language Technology*, 2014.

[169] Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Minimum risk training for neural machine translation. *arXiv preprint arXiv:1512.02433*, 2015.

[170] Yusuke Shinyama and Satoshi Sekine. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 304–311. Association for Computational Linguistics, 2006.

[171] David A Smith and Jason Eisner. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 787–794. Association for Computational Linguistics, 2006.

[172] Rion Snow, Daniel Jurafsky, and Andrew Y Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in neural information processing systems*, pages 1297–1304, 2005.

[173] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211. Association for Computational Linguistics, 2012.

[174] Anders Søgaard and Yoav Goldberg. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, 2016.

[175] Changzhi Sun, Yuanbin Wu, Man Lan, Shiliang Sun, Wenting Wang, Kuang-Chih Lee, and Kewen Wu. Extracting entities and relations with joint minimum risk training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2265, 2018.

[176] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 455–465. Association for Computational Linguistics, 2012.

[177] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 2. MIT press Cambridge, 1998.

[178] György Szarvas, Richárd Farkas, and András Kocsor. A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. In

International Conference on Discovery Science, pages 267–278. Springer, 2006.

[179] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

[180] Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In *proceedings of ACL-08: HLT*, pages 308–316, 2008.

[181] Antonio Toral and Rafael Munoz. A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, 2006.

[182] Serhat Uzunbayir. *A Comparison Between Relational Database Models and NoSQL Trends On Big Data Design Challenges Using A Social Shopping Application*. PhD thesis, 06 2015.

[183] Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. Reside: Improving distantly-supervised neural relation extraction using side information. *arXiv preprint arXiv:1812.04361*, 2018.

[184] Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. Combining recurrent and convolutional neural networks for relation classification. *arXiv preprint arXiv:1605.07333*, 2016.

[185] Guanying Wang, Wen Zhang, Ruoxu Wang, Yalin Zhou, Xi Chen, Wei Zhang, Hai Zhu, and Huajun Chen. Label-free distant supervision for relation extraction via knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2246–2255, 2018.

[186] Linlin Wang, Kang Liu, Zhu Cao, Jun Zhao, and Gerard De Melo. Sentiment-aspect extraction based on restricted boltzmann machines. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 616–625, 2015.

[187] Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, 2016.

[188] Shaolei Wang, Yue Zhang, Wanxiang Che, and Ting Liu. Joint extraction of entities and relations based on a novel graph scheme. In *IJCAI*, pages 4461–4467, 2018.

[189] Wenhui Wang and Baobao Chang. Graph-based dependency parsing with bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2306–2315, 2016.

[190] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polar-

ity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

[191] Yi Wu, David Bamman, and Stuart Russell. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783, 2017.

[192] Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 3-volume 3*, pages 1533–1541. Association for Computational Linguistics, 2009.

[193] Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. Structural opinion mining for graph-based sentiment representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1332–1341. Association for Computational Linguistics, 2011.

[194] Minguang Xiao and Cong Liu. Semantic relation classification via hierarchical recurrent neural network with attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1254–1263, 2016.

[195] Kai Xu, Zhanfan Zhou, Tianyong Hao, and Wenyin Liu. A bidirectional lstm and conditional random fields approach to medical named entity recognition. In *International Conference on Advanced Intelligent Systems and Informatics*, pages 355–365. Springer, 2017.

[196] Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv preprint arXiv:1506.07650*, 2015.

[197] Liheng Xu, Kang Liu, Siwei Lai, Yubo Chen, and Jun Zhao. Mining opinion words and opinion targets in a two-stage framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1764–1773, 2013.

[198] Wenduan Xu, Michael Auli, and Stephen Clark. Expected f-measure training for shift-reduce parsing with recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–220, 2016.

[199] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. Classifying relations via long short term memory networks along shortest dependency paths. In *proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1785–1794, 2015.

[200] Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. Improved relation classification by deep recurrent neural networks with data aug-

mentation. *arXiv preprint arXiv:1601.03651*, 2016.

[201] Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, 2018.

[202] Vikas Yadav, Rebecca Sharp, and Steven Bethard. Deep affix features improve neural named entity recognizers. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 167–172, 2018.

[203] Yadollah Yaghoobzadeh, Heike Adel, and Hinrich Schütze. Noise mitigation for neural entity typing and relation extraction. *arXiv preprint arXiv:1612.07495*, 2016.

[204] Xu Yan, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. Classifying relations via long short term memory networks along shortest dependency path. *arXiv preprint arXiv:1508.03720*, 2015.

[205] Yulan Yan, Naoaki Okazaki, Yutaka Matsuo, Zhenglu Yang, and Mitsuru Ishizuka. Unsupervised relation extraction by mining wikipedia texts using information from the web. In *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: Volume 2-Volume 2*, pages 1021–1029. Association for Computational Linguistics, 2009.

[206] Bishan Yang and Claire Cardie. Extracting opinion expressions with semi-markov conditional random fields. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1335–1345. Association for Computational linguistics, 2012.

[207] Bishan Yang and Claire Cardie. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, 2013.

[208] Bishan Yang and Claire Cardie. Joint modeling of opinion expression extraction and attribute classification. *Transactions of the Association for Computational Linguistics*, 2:505–516, 2014.

[209] Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. Multi-task cross-lingual sequence tagging from scratch. *arXiv preprint arXiv:1603.06270*, 2016.

[210] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. *arXiv preprint arXiv:1809.05679*, 2018.

[211] Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. Structured relation discovery using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1456–1466. Association for Computational Linguistics, 2011.

[212] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu,

Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of ACL 2019*, 2019.

[213] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

[214] Jianxing Yu, Zheng-Jun Zha, Meng Wang, Kai Wang, and Tat-Seng Chua. Domain-assisted product aspect hierarchy generation: towards hierarchical organization of unstructured consumer reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 140–150. Association for Computational Linguistics, 2011.

[215] Mo Yu, Matthew Gormley, and Mark Dredze. Factor-based compositional embedding models. In *NIPS Workshop on Learning Semantics*, pages 95–101, 2014.

[216] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[217] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106, 2003.

[218] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. Relation classification via convolutional deep neural network. 2014.

[219] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, 2015.

[220] Dongxu Zhang and Dong Wang. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*, 2015.

[221] Meishan Zhang, Yue Zhang, and Guohong Fu. End-to-end neural relation extraction with global optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1730–1740, 2017.

[222] Ningyu Zhang, Shumin Deng, Zhanlin Sun, Xi Chen, Wei Zhang, and Huajun Chen. Attention-based capsule networks with dynamic routing for relation extraction. *arXiv preprint arXiv:1812.11321*, 2018.

[223] Shaodian Zhang and Noémie Elhadad. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–1098, 2013.

[224] Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 73–78, 2015.

[225] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Man-

ning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, 2017.

[226] Yuhao Zhang, Peng Qi, and Christopher D Manning. Graph convolution over pruned dependency trees improves relation extraction. *arXiv preprint arXiv:1809.10185*, 2018.

[227] Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 56–65. Association for Computational Linguistics, 2010.

[228] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. Joint extraction of entities and relations based on a novel tagging scheme. *arXiv preprint arXiv:1706.05075*, 2017.

[229] GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480. Association for Computational Linguistics, 2002.

[230] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.

[231] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, 2016.

[232] Jianhan Zhu, Victoria Uren, and Enrico Motta. Espotter: Adaptive named entity recognition for web browsing. In *Biennial Conference on Professional Knowledge Management/Wissensmanagement*, pages 518–529. Springer, 2005.

[233] Yuying Zhu, Guoxin Wang, and Börje F Karlsson. Can-ner: Convolutional attention network forchinese named entity recognition. *arXiv preprint arXiv:1904.02141*, 2019.

致谢

岁月如梭，韶光易逝。眨眼间，时光像一匹飞驰的骏马，从我们的身边飞逝而去。五年多的博士生涯即将划上完满的句号。我们的一生中面临着很多的抉择，往往有些决定会改变整个人生轨迹。回首过往，很庆幸没有为自己的选择而后悔。科研的路上充满着各种各样的可能性，有过惊奇，有过迷茫，有过充实，经历过荆棘，也收获过成果。不管是成功或是失败，是欢笑或是泪水，都是人生道路上宝贵的风景，深深烙印在记忆中。感激昨天，珍惜今天，憧憬明天。

衷心感谢我的导师吴苑斌老师！在学术研究上，循循善诱，从实验研究到撰写论文，每一步的悉心指导，无不显示出吴老师严谨的学术态度和渊博的学识。吴老师不仅是传授知识的良师，更是人生道路上的明师，教会了我许多对事物的看法和人生道理，这将使我终生受益。感谢吴老师出现在我的博士生涯和我的人生中，在未来的工作中，必将谨遵教诲，负重前行。感谢我的导师孙仕亮教授，孙老师丰富的科研经验，乐观向上的科研精神和专业的学术技巧深深地影响着我。感谢曾经帮助过我的所有老师。

感谢实验室之前已经毕业的学长学姐。感谢实验室已经毕业的同学，李赋博、何荣炜等。感谢 AntNLP 已经毕业的师弟李晨瑞，以及在读的各位师弟师妹们：纪焘、杜雨沛、刘宇芳、韦阳、黄子寅、郑淇等。感谢隔壁实验室所有学弟学妹：田俊峰、王飞翔、姜梦晓等。一个人可以走的很快，很多人才可以走的很远，很高兴能和你们一起奋斗前进，这将是一段非常难忘的经历。感谢一路上所有帮助过我的朋友。

感谢我的家人，感谢父亲和母亲，感谢这么多年的支持，我一定努力成为你们的骄傲。希望在以后的日子里，家人可以永远健康幸福。

最后，感谢自己。我就是我，是颜色不一样的烟火。

在读期间发表的学术论文情况

[1] **Changzhi Sun** , Yeyun Gong, Yuanbin Wu, Ming Gong, Daxing Jiang, Man Lan, Shiliang Sun, and Nan Duan. Joint Type Inference on Entities and Relations via Graph Convolutional Networks. ACL 2019. (**CCF A 会议**)

[2] **Changzhi Sun** and Yuanbin Wu. Distantly Supervised Entity Relation Extraction with Adapted Manual Annotations. AAAI 2019. (**CCF A 会议**)

[3] **Changzhi Sun** , Yuanbin Wu, Man Lan, Shiliang Sun, Wenting Wang, Kuang-Chih Lee, and Kewen Wu. Extracting Entities and Relations with Joint Minimum Risk Training. EMNLP 2018. (**CCF B 会议**)

[4] **Changzhi Sun**, Yuanbin Wu, Man Lan, Shiliang Sun, and Qi Zhang. Large-scale Opinion Relation Extraction with Distantly Supervised Neural Network. EACL 2017.