硕士专业学位论文

注意力增强包表示的远程监督关系抽取方法研究

Research on Distant Supervision Relation Extraction with Attention Enhanced Bag Representation

作者：赵静菲

导师：尹传环

北京交通大学

2020 年 6 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：  　　导师签名： 

签字日期：2020 年 6 月 13 日　　签字日期：2020 年 6 月 13 日

学校代码：10004 密级：公开

北京交通大学

硕士专业学位论文

注意力增强包表示的远程监督关系抽取方法研究

Research on Distant Supervision Relation Extraction with Attention Enhanced Bag Representation

作者姓名：赵静菲 学　　号：18125297

导师姓名：尹传环 职　　称：副教授

工程硕士专业领域：计算机技术 学位级别：硕士

北京交通大学

2020 年 6 月

致谢

不知不觉，在交大的生活结束了。说漫长其实也很长，但是一转眼间，我们就面临了告别。毕业季的匆匆来临，意味着自己学生时代的终结。在画上句号的此刻，仅以此代表我深深地感恩之情！

在两年的研究生活中，非常感谢导师对我学业上的悉心指导。实验室的每周例会，与导师的每一次讨论，都使我获益匪浅，受益良多。在选题、开题和结题阶段，导师的鞭策让我始终保持前进的动力，在科研的道路上愈战愈勇，顺利通过各个阶段的考核。导师是一位身负学识，品德兼修，负责任的好老师，祝愿导师在职业生涯中收获更多的学术成果。

同时，也非常感谢实验室提供的宝贵的计算资源。如果没有实验室的环境，研究课题难以施展。感谢师兄师姐，在我遇到难题和疑惑时都及时回应，耐心解答，热心相助。感谢同届的实验室同学，因为有你们的陪伴、谈心和探讨，让实验室的生活增添了许多色彩。

感谢认识的朋友，相识相知是一种缘分。在路上的偶遇，在宿舍楼的每一次交谈，都让人感受到温暖和真诚。祝愿你们在往后的日子里前程似锦，永远保留热忱的心。

感谢交大，感谢学校里每一位默默付出的辛勤劳动者，是你们用双手创造了交大温暖的大家庭。食堂里的大叔和阿姨，宿管阿姨们，谢谢你们的微笑和热情，谢谢你们为交大做的一切。

感谢我的家人，你们在背后的支持和鼓励是我勇往直前的动力源泉。我不会辜负你们的期待，我是你们永远的骄傲。

初夏已至，心存留念。我会无比怀念校园的生活，带着知行合一的精神步入下一个人生阶段。以梦为马，不负韶华，加油吧！

摘要

关系抽取作为信息抽取技术的一项关键环节，在自动构建知识图谱、自然语言处理领域具有重要的理论意义和广阔的应用前景。为多种应用提供重要支持，主要表现在智能问答系统和智能化搜索场景中。其深厚的实用价值备受学界和业界的广泛关注，涌现出众多的理论成果和应用产品。远程监督关系抽取技术通过外部知识库作为监督源，自动标注语料数据，节省了人工标注成本，成为了关系抽取的研究热点。由于远程监督的强烈假设前提，导致回标语料中存在大量的噪声数据。因此，目前的研究重点主要集中在如何削弱噪声数据的消极影响上。本文基于 Transformer 预训练语言模型，提出了两种注意力增强包表示的远程监督关系抽取模型，以此缓解噪声的影响，具体如下：

（1）提出了基于 Transformer 预训练语言模型和实体注意力机制的关系抽取模型 DISTRE-EA。关系抽取任务意在识别两个实体所要表达的关系，所以实体表示的指导作用不可忽视。在发挥语言模型优势的基础上，用实体本身的信息指导注意力。通过计算实体表示与包内部每个句子之间的相关特性，深入挖掘二者的语义相似度，筛选预训练模型得到的句子嵌入表示信息，降低噪声句子影响的同时有助于优化包表示。相关实验数据表明：DISTRE-EA 优于主流方法，有效验证了模型的抽取效果。

（2）提出了基于 Transformer 预训练语言模型和句子级自注意力机制的关系抽取模型 DISTRE-SA。由于包内部的句子嵌入表示是通过预训练模型分别获得的，彼此之间缺乏紧密的关联性。为了打破包内部互相独立的句子表示，我们在句子级别上使用自注意力技术。自注意力不仅可以捕获输入序列的相互依赖关系，并且多头注意力能够丰富不同表示子空间的隐含信息。使用自注意力对句子嵌入表示进行变换后，结合选择性注意力进一步增强了包表示。在 NYT 数据集的结果表明：DISTRE-SA 相比原方法 DISTRE 具有较高的抽取准确率。

关键词：关系抽取；远程监督；语言模型；注意力机制；信息抽取

ABSTRACT

As a key link of information extraction technology, relation extraction has important theoretical significance and broad application prospects in the field of automatic construction of knowledge graphs and natural language processing. It provides important support for a variety of applications, mainly in the intelligent question answering system and intelligent search scenarios. Its profound practical value has aroused widespread concern in academia and industry, and numerous theoretical results and application products have emerged. The distant supervision relation extraction technology uses external knowledge base as the supervision source to automatically label corpus data, which saves the cost of manual labeling and has become a research hotspot in the field of relation extraction. Due to the strong assumptions of distant supervision, there is a large amount of noisy data in the annotated corpus. Therefore, the current research focus is mainly on how to weaken the negative effects of noise data. Based on the pre-trained Transformer language model, this paper proposes two distant supervision relation extraction models with attention enhanced bag representation to mitigate the impact of noise as follows:

(1) We proposed a relation extraction model DISTRE-EA based on pre-trained Transformer language model and entity attention mechanism. The relation extraction task is intended to identify the relation to be expressed by two entities, so the guiding role of entity representation cannot be ignored. Based on the advantages of the language model, the information of the entity itself is used to guide attention. By calculating the relevant characteristics between the entity representation and each sentence inside the bag, the semantic similarity between the two is deeply explored, and the representation information of the sentence obtained by the pre-training model is screened to reduce the impact of noise sentences and help optimize the bag representation. Relevant experimental data shows that DISTRE-EA is superior to mainstream methods, and effectively validates the extraction effect of the model.

(2) We presented a relation extraction model DISTRE-SA based on pre-trained Transformer language model and sentence-level self-attention mechanism. Since the sentence representations within the bag are obtained through pre-trained models, there is a lack of close correlation between each other. In order to break the independent representation of sentences within the bag, this article uses self-attention techniques at the sentence level. Self-attention can not only capture the interdependence of the input

sequence, but multi-head attention can also enrich the implicit information of different representation subspaces. After using self-attention to transform the sentence representation, combined with selective attention, the bag representation is further enhanced. The results in the NYT dataset show that DISTRE-SA has higher extraction accuracy than the original method DISTRE.

KEYWORDS：Relation Extraction; Distant Supervision; Language Model; Attention Mechanism; Information Extraction

目录

1 绪论

1.1 关系抽取研究背景

近些年来，知识是人工智能领域聚焦的重点，它是连接数据和人工智能的桥梁。在大数据时代的背景下，日益增长的海量数据使得互联网包含的信息愈发复杂多源，碎片化的多源异构数据需要转化成机器能够理解的知识类数据。如何从非结构化的网络文本中抽取中丰富的语义信息，辅助人类做出智能决策，这是自然语言处理（Natural Language Processing, NLP）亟需突破的困境。机器的感知能力在某些任务上已经超越了人类，例如宣布停赛的 ImageNet 项目。但另一方面，机器的认知能力还有巨大的上升空间，与人类的理解水平依旧相差甚远。要使机器真正地融入到世界当中，与我们自然地互动交流，就必须灌溉足够的知识给机器。当机器拥有大量的知识存储后，形成对各个场景的分析决策的能力。储存记忆信息的关键技术就是知识图谱（Knowledge Graph），如果大脑是人类的思维决策系统，那么知识谱图就是机器的核心。

2012 年 Google 首次提出了知识图谱的概念，其目的是改善现有搜索引擎的方式，提升用户查询和返回高质信息的体验。知识图谱支持用户按主题搜索，精准定位问题描述，将搜索结果知识系统化，以图形化方式反馈给用户。知识图谱把知识存储到计算机，构建紧密互联的网状结构并保持动态更新。作为智能化前沿发展的热门技术，知识图谱在学术界和业界掀起了一阵狂潮，它有十分广泛的应用场景，典型应用比如深度问答、智能语义搜索、社交网络等，在金融、 医疗、电商等垂直行业的实际应用技术日趋成熟。知识图谱按照领域划分，可分为通用知识图谱和特定领域的知识图谱。其中，Google 所提出的知识图谱是面向全领域的通用型知识图谱，以常识性知识为主，强调知识的广度，是一个结构化的百科知识库。特定领域的知识图谱是面向某个专业领域的，对精准率要求高，强调知识的广度。从学术的角度上看，知识图谱本质上一个结构化的语义知识库，揭示了物理世界中的实体及其相互关系[1]。它将数据的粒度从文件级别降到数据级别，聚集和融合大量的知识，从而达到为现实世界做出快速响应和高效处理的目的。大规模知识图谱构建的关键技术主要是信息抽取、知识融合、知识加工和知识更新[2]。信息抽取（Information Extraction）从非结构化文本中抽取实体、关系和属性等元素。通过使用知识融合技术，可以消除这类知识元素的指称项与事实对象的歧义关系。知识加工对消除歧义的知识元素进行质量评估和筛选等操作，将合格的部分加入到高

质量的知识库中。知识更新技术保证知识图谱的时效性，为知识图谱的构建保持迭代更新。其中，信息抽取包含三个子任务，分别是命名实体识别、关系抽取和事件抽取。

关系抽取技术（Relation Extraction）是指从结构化、半结构化或非结构化文本中抽取对应实体对的关系。实体抽取的输出通常是一个三元组（实体 1，关系，实体 2）。在句子“Obama was born in the United States.”中，识别出的关系组合是（Barack Obama, BornIn, United States）。关系抽取是信息抽取技术中极为关键的一环，具有深刻的研究意义和富含价值的应用前景。关系抽取为自动构建知识图谱提供技术支持，人工构建知识图谱耗时耗力，存在覆盖率低、数据稀疏等问题。关系抽取提取的结构化三元组表示可通过设置批量导入，直接自动构建庞大的知识图谱，例如使用率高的图数据库 Neo4j。关系抽取为信息获取提供技术知识，目前自然语言理解发展正在突破真正地语义理解的阶段，关系抽取是篇章理解的核心部分，在改善自然语言处理性能上有极大的贡献作用。

1.2 研究目的和挑战

现存的大型知识图谱 WordNet、Wikidata、DBpedia 以结构化方式存储海量数据，涵盖大量丰富的信息。由于现实世界爆炸式增长的复杂数据，研究者们努力探索快速利用高质知识数据的办法，为自动构建知识图谱提供技术支持。其中，关系抽取技术是一项快速且高效的识别方式，可以解决目标实体之间的关系分类问题，是构建知识图谱的重要步骤。关系抽取是一个经典任务，此项研究持续 20 多年，取得了一系列阶段性的成果。随着深度学习的复苏，神经网络模型在特性提取的深度和模型的准确率上已经超越了基于特征工程和核函数的传统机器学习方法。

关系抽取主要面临三大挑战：自然语言表达的多样性、关系表达的隐含性和实体关系的复杂性。

自然语言表达的多样性：由于文本表述的方式多种多样，一种关系的说法有多种不同的体现。比如，“X 出生在 Y 地”，“X 回到了他的故乡”，“X 一直住在 Y 地”等等，都表达了 Y 是 X 的“出生地”这个意思。从文本中识别的实体关系需要映射到结构化的三元组表示上，要求关系抽取具有准确性。

关系表达的隐含性：不是每个关系的表述都能在文本中找到一个与之相关联的准确信息，我们只能通过常识性的知识去理解出句子要表达的意思。如今，通过词的上下文信息可以整合出一个词的向量表示，该向量表示本身能涵盖丰富的语义信息。但是在一些专用领域上，并不能提供大量丰富的语料来对词向量进行训练。

实体关系的复杂性：现实世界的事物之间是随着时间的变化而改变的，比如某

个会议的举办地点可能每年都在变化。人物的关系更加复杂，一对配偶在某年之后离婚了，那么他们的关系就变成了前夫和前妻。因此实体关系的时效性难以保证，实体关系需要保持动态更新。

1.3 关系抽取研究现状

从 20 世纪 90 年代以来，关系抽取技术得益于系列国际权威评测和会议的推动。如在消息理解系列会议 MUC，自动内容抽取评测 ACE 和文本分析会议系列评测 TAC 上[3]，关系抽取的检测和识别任务均占据了一席之地。近年来，在国际会议 ACL、NAACL、EMNLP、AAAI 上，关系抽取任务是持续关注的对象。关系抽取的研究内容主要分为限定域关系抽取和开放域关系抽取。限定域关系抽取是指按照事先预定义关系的类别，基于人工或基于启发式的规则自动构建标注语料，主要利用有监督和弱监督方法识别定义的关系类型。有监督方法致力于如何挖掘更多的特征相关表示上。弱监督方法主要解决自动构建语料产生的噪声问题。开放域关系抽取是指在无事先定义关系类别的情况下，由系统自动识别抽取关系，主要利用无监督的方法用关系指示代词表示关系类型。

随着深度学习时代的来临，关系抽取逐渐从人工特征工程过渡到端到端的神经网络模型方法。从是否依赖监督数据的角度上看，关系抽取分为有监督、无监督、半监督和弱监督方法。有监督方法依靠大量的标注语料进行训练，标注语料的质量直接影响到模型的性能。有监督方法虽然能获得最好的抽取效果，但是人工标注代价昂贵，且无效率保证，很难适应到大规模的任务当中。无监督方法通过聚合模板的方式，把具有语义表示相同的实体词进行聚集，此过程无需人工标注的数据。因为没有预定义关系类别，所以它通过层次聚类、主题模型等技术可以发现新的关系，但是新发现的关系有很大概率不具备语义信息，缺乏规整的结构化三元组表示，导致最终的关系类别难以归一化。半监督方法通过反复迭代更新种子集的方式，在系统中逐渐扩大关系数据集。由于不断增加模板而带入噪声，容易造成“语义漂移”的现象。因此，弱监督或远程监督的方式成为关系抽取任务的主要研究方向，远程监督利用知识库回标文本获得大量的监督数据，可以自动生成训练预料。远程监督需要结构化的知识库作为种子，生成的语料不可避免地存在噪声数据，所以该任务主要研究如何降低语料的噪声问题。

Mintz[4]在 2009 年提出了远程监督（Distant Supervision）的思想，它假设在知识库的两个实体如果存在某种关系，那么所有包含这两个实体的训练句子都表达了这种关系，即所有训练句子的关系标签都是知识库的实体对所对应的关系。然而，远程监督的假设太过强烈，产生了错误标签的问题，一个句子中出现的这两个实体

并不一定表示了它们在知识库的关系，有可能这两个实体只是共享一个主题而已。比如知识库存在的三元组（Apple, /business/company/founders , Steve Jobs），训练语料有两个句子“Steve Jobs was the co-founder and CEO of the Apple and formerly Pixar.”和“Steve Jobs passed away the day before Apple unveiled iPhone 4S in late 2011.”可以明显发现，句子 1 表达了正确的关系，但是句子 2 并不表示“创始人”的意思。为了解决句子噪声的问题，2010 年 Riedel[5]把远程监督关系抽取任务视为多示例问题，多示例学习（Multi-instance Learning）把包含相同实体对的所有示例（句子）看成一个包（Bag），根据知识库定义每个包的关系，从句子级别粒度转换为包级别粒度。多示例学习有一个“At-Least-One”的前提，即包含两个实体的所有句子中，至少有一个句子可以体现该关系，即至少有一个标注正确的句子。在此后的一系列研究中，都是以多示例为基础实现关系抽取模型。2015 年 Zeng[6]创造性地提出了分段卷积神经网络（Piecewise Convolutional Neural Network, PCNN）的思想，以两个实体为界把句子分成三段，分别最大池化后再组合成句子的向量表示，该方法有效地捕获两个实体之间的结构信息，避免了对隐层节点的过度削减。2016 年 Lin[7]实现了选择性注意力机制（Selective Attention）模型，对包中的每个句子分配权重来缓和噪声的影响，自此分段卷积结合注意力机制成为了远程监督关系抽取的标杆性存在。

2019 年，Alt[8]发表在 ACL 会议上的论文提出了结合 Transformer 预训练语言模型和选择性注意力机制的远程监督关系抽取方法 DISTRE（Distantly Supervised Transformer for Relation Extraction）。谷歌团队提出的 Transformer 模型最早应用于机器翻译任务[9]中，能够有效解决 Seq2Seq 的序列问题。它抛弃了传统的卷积神经网络（Convolutional Neural Network, CNN）和循环神经网络（Recurrent Neural Network, RNN），整个网络结构完全是由自注意力机制（Self-attention）和前馈神经网络（Feed Forward Neural Network）组成[10]，自注意力机制能够实现快速并行，改进了 RNN 训练慢的特点。通过堆叠 Transformer Block 的形式使模型增加到非常深的深度，充分发掘深度神经网络的特性，提升模型的鲁棒性和泛化能力。2018 年在 11 项自然语言处理任务中取得卓越成绩的 BERT 算法[11]，改变了基于预训练语言模型生成的词向量和下游特定自然语言处理任务之间的关系，而 BERT 最关键的部分便是 Transformer 的概念。

虽然，基于注意力机制的 Transformer 模型在自然语言理解各任务（包括关系抽取）上取得了较大的成功。但是，现有的基于 Transformer 语言模型的关系抽取方法 DISTRE 没有考虑实体和句子对包表示的重要性。因此，本文基于 DISTRE 将研究不同级别的注意力机制对关系抽取的影响。

1.4 主要研究内容及创新

本文以 Alt 提出的方法 DISTRE 为基础模型，提出了实体注意力机制模型和基于句子级的自注意力机制模型，主要内容如下：

（1）研究了基于 Transformer 预训练语言模型和实体注意力机制的关系抽取模型（Distantly Supervised Transformer for Relation Extraction based on Entity Attention, DISTRE-EA）。预训练语言模型本身涵盖更丰富的语法和语义信息，Transformer 相比于其他预训练语言模型属于轻量级，适合快速微调（fine-tuning）的任务。虽然句子表示信息丰富，但是关系抽取任务不单单是句子分类任务，我们应该聚焦实体 1 和实体 2 相关的句子信息。预训练语言模型不适合改变内部结构，本文将针对任务的设计放在预训练语言模型之外，由此我们提出了实体注意力机制模型。通过 Transformer 编码得到包中的句子嵌入表示后，再计算实体和句子之间的相关性，判断哪些句子的表示包含了更多与实体相关的信息。实验表明，我们的模型在 NYT（New York Times）数据集[7]上比基础模型占据绝对地优势，AUC 值达到了 0.43。

（2）研究了基于 Transformer 预训练语言模型和句子级自注意力机制的关系抽取模型（Distantly Supervised Transformer for Relation Extraction based on Self-attention, DISTRE-SA）。目前仍然存在的问题是，包中的句子表示之间相互独立，没有语义联系。本文在句子粒度上使用自注意力机制，使得每个句子的表示都与其他句子有关。通过计算句子之间的相关性，充分利用多头注意力的优势捕捉各个表示子空间的信息，进一步增强包的嵌入表示。基于自注意力机制的句子表示能丰富地挖掘语义语法上的相关性，有效地提升模型的抽取效果。最终实验中 AUC 值达到 0.427。

1.5 论文组织结构

第 1 章 绪论。主要阐述关系抽取的研究背景、研究目的和存在的挑战。根据现有的研究状况提出远程监督关系抽取任务，在本章的最后介绍了本文的主要研究内容及创新。

第 2 章 技术方法和研究现状。探讨国内外对关系抽取任务的相关理论综述，分类介绍有监督、无监督和半监督关系抽取研究进展，重点阐述在神经网络模型技术背景下的远程监督关系抽取方法。

第 3 章 注意力增强包表示的关系抽取模型。对本文提出的两种模型 DISTRE-EA 和 DISTRE-SA 做详细介绍，深度剖析网络架构的实现细节和作用原理，讲述模型的设计思想及目标优化过程。

第 4 章 实验设计与结果评估分析。介绍远程监督关系抽取的标准数据集，给出实验优化方案和常用评估指标。针对特定的参数进行一系列的探讨分析，通过对比领域内具有代表性的模型，验证本文提出的两种方法的有效性。

第 5 章 总结与展望。总结本文的主要研究内容，并对下一步的研究方案做出了展望。

2 技术方法和研究现状

目前，基于机器学习的关系抽取方法占据了主导地位。机器学习发展已有几十年的历史，是一门多领域交叉的学科，涵盖概率论知识、统计学知识和复杂算法知识。近些年来，得益于计算机软硬件性能，深度学习技术蓬勃发展，至今已有数种深度学习框架，如卷积神经网络、循环神经网络、生成式对抗网络、图卷积网络，各式各样的神经网络模型应用于图像处理、机器视觉、多媒体计算、自然语言处理领域，并获得卓越的理论研究成果以及富有创造性的实际应用产品。深度学习作为机器学习的一个重要组成部分，以人工神经网络为基本架构，通过分层特征提取的高效算法进行自主学习。

在关系抽取任务中，传统的抽取方法需要专家设计手工特征，这种使用 NIP 工具提取特征的方式存在传播误差问题。因此，越来越多的学者将深度学习模型应用到关系抽取任务当中。基于深度学习的关系抽取模型能够自动抽取特征，相比于传统的方法而言，无需复杂的特征工程。因此，基于机器学习的关系抽取任务主要分为有监督关系抽取、无监督关系抽取、半监督关系抽取和远程监督关系抽取四大类。远程监督关系抽取是本文的重点，将详细介绍近些年的理论研究进展。

2.1 有监督关系抽取

有监督关系抽取通常把关系抽取任务视为端到端的分类任务，依靠人工标注的语料进行模型训练，对指定的关系进行匹配识别，可以抽取出更加有效准确的特征信息，所以在精准率和召回率上有很大的优势。有监督方法可通过数据特征学习数据的分布式表示，将底层特征进行组合，形成更加抽象的高级特征，接着使用有监督的分类器实现关系抽取。有监督关系抽取任务一般使用标注完全准确的 SemEval2010 Task 8 数据集[12]和美国国家标准技术研究院组织的 ACE 2005 评测数据集[13]。

神经网络模型是一个“黑箱”操作，输入低维向量之后，经过训练就能得到输出结果。端到端模型的特点是通过前向传播网络计算结果，梯度反传更新模型参数，每层神经网络横向可多个神经元共存，纵向可有多层神经网络连接。CNN 是传统神经网络的扩展，它的优势是可共享卷积核，消除参数膨胀的问题。Liu[14]第一次把 CNN 应用到关系分类任务中，借助 WordNet 知识库聚类构造同义词列表，把每个词映射到同义词列表上，使用独热编码形成词向量的表示。除此之外，还加入了实体类型、词性（POS）标注特征。整个模型架构非常简单，甚至没有使用池化层

对每个卷积核选取最优的特征表示，因此产生的噪声比较明显，降低了模型的泛化能力。由于仅仅使用一个尺寸的窗口导致提取出的特征比较单一，Nguyen[15]在卷积阶段使用多个尺度的卷积核，卷积后加入了激活函数。而 Santos[16]把方向转移到目标函数上，用排名损失函数代替交叉熵损失函数，给予正样本更高的评分，负样本更低的评分，该方式易于区分容易混淆的关系类别。

CNN 模型基于空间扩展，用于静态输出，捕获的是局部 n-gram 特征，无法处理有变化的时间序列问题。但是样本的先后顺序在 NLP、机器视觉等领域有着十分重要的影响，因此出现了循环神经网络。RNN 基于时间扩展，不仅接收当前时刻的输入，同时也接收上一个时刻的输出，具有时间记忆功能，对序列的非线性特征进行学习时具有一定的优势。Zhang[17]把双向循环神经网络用于关系分类，在不使用任何词法特征的前提下，与 Nguyen[15]使用多窗口方法相比，达到了相似的实验效果。然而循环神经网络存在长期依赖的缺陷，在对序列进行建模时，由于反向传播路径太长，会出现严重的梯度消失（Gradient Vanishing）和梯度爆炸（Gradient Explosion）现象，难以捕捉文本中的长期时间关联。因此，结合不同的长短时记忆网络（Long Short-Term Memory Networks, LSTM）可以很好地解决这个问题，LSTM 通过门机制控制特征的流通和损失，Zhou[18]使用 BiLSTM 结合注意力机制的方式，应用到关系分类任务中，在 SemEval2010 Task 8 数据集上表现良好，F1 值达到了 84%，与 Santos 排名损失函数的结果类似。Zhou 对 LSTM 的每一步输出都做了一个加权操作，这样可以缓和句子中的噪声词语的影响，增强关键词的作用。Wang[19]在 CNN 基础模型上采用多个注意力机制，效果有了巨大的提升，F1 直达 88%。文中主要使用两个注意力，一个是输入层注意力，另一个是池化层注意力。输入层注意力对两个实体分别设计了两个对角矩阵（本质上是实体的向量表示），将句子中的每个词向量分别与两个实体表示做内积计算，以权衡哪些词对实体的贡献大小。最后使用 softmax 进行归一化，得出第一层的注意力权重。第二个池化层注意力取代了卷积后的最大池化层，Wang 定义了一个随机初始化的注意力参数矩阵，把它和关系标签的向量表示、卷积后的特征表示三者进行矩阵运算，以此表征第i个元素与第j个关系的相似性。Wang 在另一方面改进了 Santos 的排名损失函数，使用关系向量表示来逼近最后的输出，以此衡量正确的标签与模型输出的相关性。在此之后，Zhu[20]对输入层注意力进行了改动，直接对句中的单词与关系标签计算相似度，以表征单词对关系的贡献作用。

有监督关系抽取方法一般默认句子中已经标注了候选实体，但是在实际任务中，应该由系统自动地发现实体。并且端到端的网络需要大量已标注的关系标签对模型进行训练，所以有监督方式对数据标注依赖性强，跨领域难，完备性差。

2.2 无监督关系抽取

无监督关系抽取技术基于分布式假设理论，是一种自顶向上的构建方法，主要应用于开放领域。分布式假设的核心思想是：如果两个词具有相似的上下文信息，就认为这两个词共享某种语义关系。那么在关系抽取任务中，若两个实体出现在相似的上下文语境中，那么这两个实体就倾向于表达某种特定的关系。无监督关系抽取分为关系实例聚类和关系类型词选择这两个过程，通过将实体对进行聚类，然后选择具有代表性的词语作为实体的关系标签。

Hasegawa[21]首先实现了无监督关系抽取模型，其关键思想是把命名实体之间的上下文词的相似性作为语义关系的特征表示，通过层次聚类的方法，将成对的命名实体进行聚集。然后设置阈值识别潜在的语义关系，选取出现频率最高的词语作为关系类别。Hasegawa 等人收集一整年的纽约时报预料作为实验语料，检测命名实体关系的同时自动为关系提供适当的标签。Bollegala[22]提出了一种使用 Web 搜索引擎的关系相似度度量，以计算实体对所隐含的语义关系相似度。通过自动提取的词法模式来表示实体对存在的各种语义关系，然后对提取的词法模式进行聚类，确保每个聚类代表某个特定的关系类别。之后，Bollegala[23]又提出了一种新颖的联合聚类算法（Co-clustering Algorithm），利用实体对空间和模板空间的对偶性，训练了一个$L1$正则化逻辑回归模型，以识别每个聚类簇代表的关系模式。针对一个模板可能对应多个语义关系的问题，Yao[24]提出了主题模型的思想，使用局部和全局特征将与模式关联的实体对划分为不同的聚类簇，然后再使用层次化聚类将聚类簇映射到语义关系上。

无监督关系抽取方法极大地减少了对标注语料的依赖性，在构建知识杂糅、多领域的大规模文本库中比其他监督方法有较高的优势。虽然无监督方法节约了人力物力资源，可移植性强，但是模型的精准率和召回率较低，抽取效果差，其挖掘的语义关系难以归一化。

2.3 半监督关系抽取

半监督关系抽取是一种自扩展技术，根据预设的关系类型以少量的人工标注的实例模式作为初始种子集，通过循环反复学习的机制来标注大量未标注语料，以此逐渐扩大关系数据集和序列模式。

Bootstrapping 算法是经典的半监督关系抽取算法，由 Brin[25]率先提出后构建了 DIPRE 系统。Bootstrapping 算法从关系种子的上下文语义信息中，识别能够表达实体关系的序列模式，利用模式相似性度量函数把置信度高的新关系模式加入

初始种子集中。该过程反复迭代，通过种子集对剩余的未标注语料进行预测分类。从万维网非结构化文本中抽取新的实例，同时学习新的抽取模板，不断更新关系模式种子集。Agichtein[26]在 Brin 的基础上实现了 Snowball 抽取系统，该系统标注了实体和属性值的实体类型，采用限定实体类型的办法实现半监督关系抽取。Snowball 系统从原始文档中抽取实例模式和提取实体关系元组，对每次迭代提取得到的模式和元祖进行评估，最后只保留高质和可靠的部分以提升整体的抽取性能，该评估过程无需人工干预。他们使用 300000 多份的报纸文档，对 Snowball 技术进行了全面的实验评估。之后，Liu[27]提出了一种可以同时抽取多种关系类型的 MultiSnowball 系统，改变了仅支持单一实体类型识别的现象。MultiSnowball 系统采用 Bootstrapping 框架，不仅可以迭代地查找新的关系元组和抽取模式，而且可以标识新的关系类型。抽取模式在所有类型中可共享，以此获得更多的关系元组。此后的系统一般沿着 Bootstrapping 的思路，加入更合理的模式描述、限制条件和评分策略。另一方面，Liu 基于先前的抽取结果构建大规模的模式库，如 NELL（Never-Ending Language Learner）系统，目前已获得 280 万个事实。

半监督关系抽取方法最大的优势是仅需少量的标注语料，大大减少了人工干预的成分。在更新模式的过程中，新模型的加入无可避免地带来噪声影响，因此如何对模型进行过滤是一个难点，否则过多的噪声将会造成“语义漂移”现象。所以基于半监督的方法虽然精准率较高，但是召回率往往偏低。

2.4 远程监督关系抽取

远程监督关系抽取的概念最早由 Mintz[4]提出，主要是利用知识库回标文本数据的思想，解决了大量无标签语料的标注难题。Mintz 假设文本中的实体对与知识库的实体对一致，那么就标注包含相同实体对的所有句子为同一种关系。

上述自动收集数据集的方式，需要把待标注的语料对齐到恰当的结构化知识库。Mintz 使用维基百科中每个词条的页面构成待标注语料，并将其对齐到 Freebase 知识库；Riedel[5]使用纽约时报的文本内容作为待标注预料，并同样将其对齐到 Freebase 知识库，由此构成了 NYT 数据集；Zeng[28] 同样使用纽约时报的文本内容作为待标注预料，但将其对齐到了 Wikidata 知识库[29]。其中，最常用于比较远程监督关系抽取模型效果的公开数据集是上述纽约时报和 Freebase 知识库组合形成的 NYT 数据集。

由于上述构建数据集时的假设太强，导致回标噪声问题，即大量语句被标注了错误的关系标签。表 2-1 展示的是关系抽取任务的分类，与其他类别相比远程监督最大的优势便是无需标注语料，但是回标语料存在噪声。为了解决噪声的影响，

Riedel[5]把远程监督关系抽取视为多示例问题。多示例学习（Multi-instance Learning）把包含相同实体对(e_1, e_2)的所有句子当成一个包（Bag），包中的每个句子称为示例，然后根据知识库定义每个包的关系。若知识库中存在关系定义，则该示例标注为正样本，否则标注为负样本。

表 2-1 关系抽取分类
Table2-1 Relation extraction classification

分类	数据集标注方法	优点	缺点
有监督	人工标注	获得最好的抽取效果	标注费时费力，难以大规模推广
无监督	无需人工标注	可以发现新关系	关系不具备语义信息，难以归一化
半监督	少量人工标注	减少人工干预，可自扩展学习	模式存在噪声，召回率偏低
远程监督	远程对齐知识库	自动生成训练语料	需要知识库作为种子，语料存在噪声

在 Riedel 之后，学者从优化特征提取和降噪两个角度不断完善远程监督关系抽取的解决方案。其中，优化特征提取存在以下两个思路：（1）特征提取器，随着深度神经网络的飞速发展，使用更有效的网络作为特征提取器；（2）实体相关信息，结合关系抽取任务特点，将实体的相关信息引入到特征提取中。降噪也存在以下两个思路：（1）抑制噪声，基于不同的注意力机制，为句子表示分配权重，抑制句子表示中噪声的影响；（2）过滤噪声，通过捕捉数据集的潜在分布，剔除或修正噪声样本。

2.4.1 优化特征提取

特征提取（Feature Extraction）是将原始数据转化为机器学习算法可以处理的数值特征的过程。在建模过程中，特征提取具有关键性的作用，为后续的标签分类奠定了基础。虽然通过深度神经网络可以自动地提取特征，但也需要针对自然语言和任务特点的结构设计，才能更好地获取符合任务需求的特征。因此，优化特征提取的方法可以分为设计更有效的特征提取器和融入实体相关信息这两类，具体如下：

（1）特征提取器

基于统计机器学习的传统方法提取的特征通常来自现有 NLP 系统的输出，容易造成现有工具的错误传播现象，其抽取的性能大部分取决于提取特征的质量。2014 年 Zeng[30]首次把端到端的神经网络模型应用到远程监督关系抽取任务中，利用卷积神经网络提取词汇特征和句子级别的特征，无需复杂的预处理就可以将所有单词作为输入。通过查找表（Lookup table）将单词标记（Token）转换为词向量（Word embedding），词向量特征连接上相对位置特征（Position feature）输入到卷积深度神经网络。最后通过 softmax 分类器，预测两个标记实体之间的关系。其中，相对位置特征是指句中的单词与两个实体的相对距离：

$$S\text{: } [People]_{e_1}\ have\ been\ moving\ back\ into\ [downtown]_{e_2}$$

在句子S中，“moving”到实体“People”和“downtown”的相对距离分别是-3 和 3。将相对距离映射到随机初始化的向量中，获得当前单词到实体的距离向量d_1和d_2，$PF = [d_1, d_2]$即为相对位置特征。

通常，深层神经网络能抽取更深的语义特征，但也会存在梯度消失和梯度爆炸的问题导致难以训练。深度残差网络（Deep Residual Network）是 2015 年提出的深度卷积网络，一经问世便在 ImageNet 中斩获图像分类、检测、定位三项任务的冠军。ResNet 在每两个卷积层增加一个捷径（shortcut），构成一个残差块，浅层网络的特征可跳跃传递至深层网络。ResNet 在众多计算机视觉任务中取得了最先进的性能，但是在 NLP 领域还未有较深的涉足。Huang[31]设计了一种带有残差学习的新型卷积神经网络。虽然 ResNet 解决了深度 CNN 难以训练的问题，但是模型的深度并不是越深越好。Huang 发现 9 层的残差网络可达到与 PCNN-ATT[7]相似的效果。

Transformer 作为一种预训练语言模型，通过预训练的方式保留了大量的文本知识，在多项 NLP 任务中体现了网络结构的优越性，是一种易用且有效的特征提取器。Alt[8]使用 Transformer 编码句子的表示，通过堆叠模块的形式加深网络的深度。在良好的特征提取前提下，可以提升关系的预测准确性。

（2）实体相关信息

对于关系抽取任务，优化特征提取的方式除了选择合适的特征提取器之外，引入实体相关信息也是有效的解决思路。在模型中强调实体的影响作用有利于识别两个实体之间的关系。

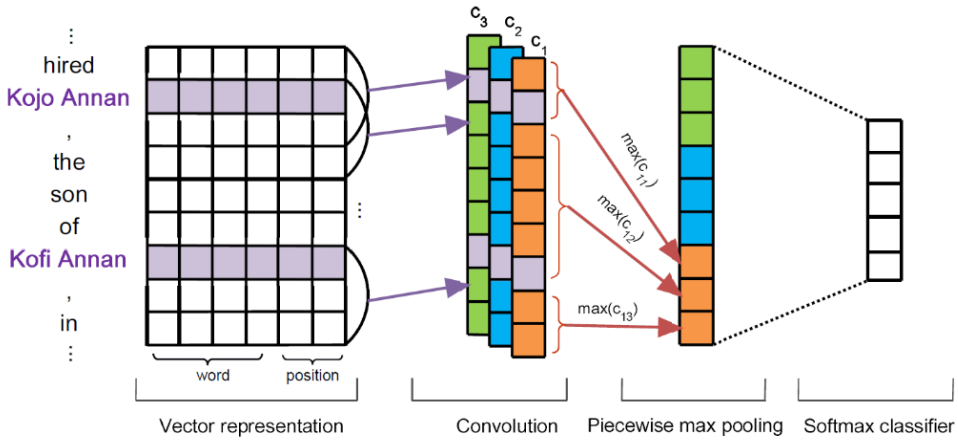


图 2-1 远程监督关系抽取 PCNN 框架[6]

Figure 2-1 The architecture of PCNNs used for distant supervised relation extraction[6]

Zeng[6]使用了考虑实体位置的卷积网络结构，模型框架见图 2-1。以两个实体的位置信息为界分成三部分：实体 1 前，实体 1 和实体 2 之间，实体 2 后。对这三部分卷积的结果取最大值：

$$p_{ij} = \max(c_{ij}), \quad 1 \le i \le n, 1 \le j \le 3 \tag{2-1}$$

即使用局部的最大池化层取代全局池化层。同时，针对远程监督的特殊性，在 PCNN 的基础上应用了多示例学习的技巧，定义一个基于包级别的损失函数，取包中所有示例的最高后验概率作为整个包的关系预测值。PCNN 模型减弱了噪声，但是只取置信度最高的示例作为包的表示，毫无疑问损失了大部分信息。

为了更好地利用先验知识，Ji[32]引入了实体描述信息，加强对实体表示的学习。如图 2-2，该模型从 Freebase 和 Wikipedia 页面中提取实体描述信息后，使用基础的卷积神经网络和池化层对其进行建模得到实体描述向量d_i。

Descriptions
[Nevada]: Nevada is a state in the Western, Mountain West, and Southwestern regions of the United States.

[Las Vegas]: officially the City of Las Vegas and often known as simply Vegas, is a city in the United States, the most populous city in the state of Nevada, the county seat of Clark County, and the city proper of the Las Vegas Valley.

图 2-2 实体 *Nevada* 和 *Las Vegas* 的描述信息[32]

Figure 2-2 The descriptions of *Nevada* and *Las Vegas*[32]

该模型有两个主要贡献，一是借鉴 TransE[33]的思想，使用两个实体向量的差值作为实体关系的表示$v_{relation} = e_1 - e_2$，与分段卷积 PCNN 得到的句子向量$b_1, \dots, b_q$进行拼接。w_i用来计算句子在包中的权重α_i：

$$w_i = W_a^{\intercal}(\tanh[b_i; v_{relation}]) + b_a \tag{2-2}$$

$$\alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^{q} \exp(w_j)} \tag{2-3}$$

二是在模型的损失函数L中加入一个先验的惩罚项λ，使用二范数来约束实体描述信息d_i和实体表示e_i之间的距离：

$$L_e = \sum_{i=1}^{|D|} \|e_i - d_i\|_2^2 \tag{2-4}$$

$$min\ L = L_A + \lambda L_e \tag{2-5}$$

上式中，$D = \{(e_i, d_i)|i = 1, \dots, |D|\}$代表（实体，描述）的组合，$L_A$是 PCNN-ATT[7]（不添加实体描述信息）得到的交叉熵损失。该模型充分利用了知识库中的监督信息，以补充任务的背景知识。背景知识不仅为预测关系提供了更多信息，而且还为注意力模块带来了更好的实体表示[34]。

Jat[35]为了强调实体的影响，在实体与单词之间进行相似度计算。为了表征词与实体的关联程度，例如“Former President Barack Obama was born in the city of Honolulu, capital of the U.S. state of Hawaii.”句子中“President”暗示实体“Barack Obama”是一个人物。该方法设计了一种基于单词与实体的注意力机制模型 EA，如公式（2-6）所示，u_{i,j,q^k}代表第j^{th}个单词与第k^{th}个实体的注意力权重：

$$u_{i,j,q^k} = [x_{i,j}, e_{q^k}^{emb}] \times A_k \times r_k \tag{2-6}$$

同时提出了基于双向门控循环单元（Bi-GRU）的注意力模型 BGWA，以 Bi-GRU 替代简单的 CNN 进行特征提取，转化了每个句子的语法语义的建模形式。把简单模型 PCNN 和以上提出的两种模型形成一个组合模型，使用加权投票方法综合三个模型的预测结果。

$$P_{i,ENSEMBLE} = \alpha * P_{i,PCNN} + \beta * P_{i,EA} + \gamma * P_{i,BGWA} \tag{2-7}$$

其中，$P_{i,model} \in R^{1*d_r}$代表第$i$个句子的所有关系得分，采用线性回归学习参数$\alpha, \beta, \gamma$。

综合以上分析，Transformer 作为特征提取器可以取得较好的分类性能，且实体的相关信息对于关系抽取任务同样重要。本文提出的基于 Transformer 和实体注意力的方法正是受到上述启发，将两者结合，发挥各自的优势。

2.4.2 降噪

远程监督关系抽取解决噪声的方法主要分为两大类，一是抑制噪声的影响，二是过滤噪声样本。抑制噪声是指在所有的训练样本中，丰富真正例（True Positive）的语义表示，降低假正例（False Positive）和假负例（False Negative）对模型性能的影响。过滤噪声是指借助方法识别有效示例，将无效示例直接摘除，构建另一份干净的训练集；或将知识库回标的关系标签在训练过程中进行更改，重新指派。

（1）抑制噪声影响

2016 年 Lin[7]在 Zeng[6]的基础上实现了选择性注意力机制模型（Selective Attention），如图 2-3，改善了 Zeng 只利用包中单个句子信息的缺陷。

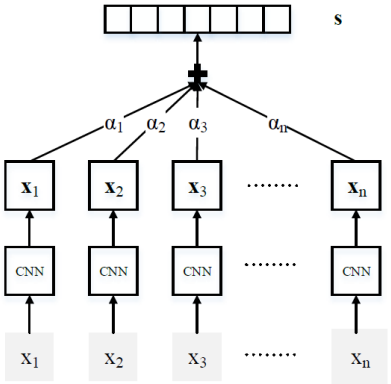


图 2-3 远程监督关系抽取 PCNN-ATT 框架[7]

Figure 2-3 The architecture of sentence-level attention-based CNN[7]

该文使用随机初始化的关系向量r指导注意力，为包中的每个示例句子x_i赋予不同的注意力权重α_i。标签正确分类的示例句子贡献较大，分配较高的权重；标签错误分类的示例贡献较小，分配较低的权重。最终句子向量使用该权重加权求和构建一个包的向量s：

$$\alpha_i = \frac{\exp(x_i A r)}{\sum_k \exp(x_k A r)} \tag{2-8}$$

$$s = \sum_i \alpha_i x_i \tag{2-9}$$

公式中，α_i代表权重大小，x_i代表卷积后的句子向量表示。在测试时，由于不知道包的真实关系，所以要遍历每个关系来计算注意力，得到每个关系的概率值后取最大值。PCNN-ATT 模型是一个经典方法，它在没有丢弃句子的情况下，有效地缓解了噪声句子的影响，实现了从使用句子级表示预测标签到使用包级别表示预测标签的转变，被之后的研究者广泛采纳。

大多数现有方法都在独立处理每个关系，而不管关系层次结构中是否存在丰富的语义相关性。2018 年，Han[36]注意到 Freebase 存储的关系是多层次结构化的，因此提出了一种新颖的从粗粒度到细粒度的分层选择性注意力（Hierarchical Selective Attention）关系抽取模型。在图 2-4 中，底层的关系是“/people/deceased person/place of death”，它的父关系是“/people/deceased person”，祖先关系是“/people”。在最顶层“/people”的关系下，还存在分支“/deceased person/place of burial”，所以底层的两个分支结构 “/place of burial”和“/place of death”可以共享父关系和祖先关系。

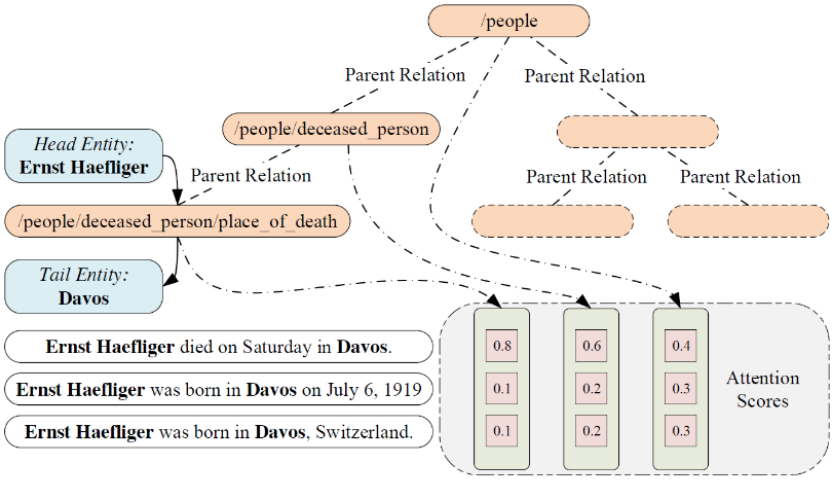


图 2-4 层次关系抽取的例子[36]

Figure 2-4 An example of hierarchical relation extraction[36]

因此，利用关系的层次结构从上到下构建类似前缀树的树状结构，顶层可捕获由几个相关子关系共享的共同特征，属于粗粒度的维度信息；最底层可捕获关系的更具体特征，属于细粒度的维度信息。在对应的三个层次关系中分别使用选择性注意力进行注意力计算，加权求和后分别得到包的嵌入表示$r_{h,t}^{i}$，这三个包的表示进行拼接形成最终的表示$r_{h,t}$：

$$r_{h,t} = [r_{h,t}^{0}; \ldots; r_{h,t}^{k-1}] \tag{2-10}$$

$r_{h,t}$被用来计算条件概率$p(r|h,t,S_{h,t})$，以预测关系类别。实验结果表明，层次关系注意力对提取长尾关系非常有效，可以提供比单层注意力更多的信息。

先前的研究主要通过设计具有包内部注意力的神经网络，专注于句子级别的降噪。Ye[37]同时考虑了包内部和包之间的注意力，以便分别处理句子级别和包级别的噪声。在处理包内注意力时，将所有的关系指导注意力计算权重，而不仅仅使用 Lin[7]中的单一目标关系：

$$e_{kj}^{i} = \{r_k s_j^{i}\}^{\intercal} \tag{2-11}$$

式中，权重e_{kj}^{i}代表在包b^{i}中第k个关系向量与第j个句子的匹配度。

如果两个包b^{i_1}和b^{i_2}都标记为关系k，则它们的表示$b_k^{i_1}$和$b_k^{i_2}$应该彼此靠近。给定一组具有相同关系标签的包群（a group of bags），为那些与其他包接近的包分配更高的权重β_{ik}：

$$g_k = \sum_{i=1}^{n} \beta_{ik} b_k^{i} \tag{2-12}$$

g_k代表包群的向量表示，受 Vaswani[10]的自注意力的启发，在衡量相似度时仅考虑包群内部的点积运算：

$$\beta_{ik} = \frac{\exp(\gamma_{ik})}{\sum_{i'}^{n} \exp(\gamma_{i'k})} \tag{2-13}$$

$$\gamma_{ik} = \sum_{i'=1,\ldots,n,i'\neq i} b_k^i b_k^{i'^{\top}} \tag{2-14}$$

该模型在 NYT 数据集上 AUC 达到了 0.422 的效果，比 Lin[7]的基础模型高出 3.4 个百分点，证明加入包间的注意力可以有效提升模型的性能。

在上述方法中，大多数的模型基于注意力机制实现抑制噪声的影响过程，包括选择性注意力、层次注意力和包间包内注意力。其中，后两种注意力的模型框架或多或少结合了选择性注意力机制。它们的主要思路都是对优质句子分配高的权重，对嘈杂句子分配低的权重。这种选择性分配权重的方法可以抑制噪声的影响，改善包的表示，为分类器提供一个优化后的输入。

（2）过滤噪声

过滤噪声是另一种降噪思路，通过更改原始数据集的方式筛选噪声样本。例如，修改误标的关系标签，重新指派置信度高的关系标签；或构建另一份干净的数据集，使用新的数据集对神经网络进行训练，提高分类的效果。

Liu[38]基于 Lin[7]的基础上提出了软标签降噪的方法。远程监督回标语料产生的标签称为硬标签（DS label），在有噪声的硬标签下训练的模型难免受影响。因此 Liu 设计了一种联合评分函数，结合了实体对表示$< h_i, t_i >$和硬标签置信度的关系评分，根据评分改变实体对的关系标签，称为软标签r_i（Soft label）。

$$r_i = argmax(o + \max(o) A \odot L_i) \tag{2-15}$$

式中，$o, A, L_i \in R^{d_r}$，L_i代表关系类别的独热编码，关系置信度向量A代表硬标签的可靠性，A中的每个元素都是 0 到 1 之间的小数。o代表实体对$< h_i, t_i >$经过 PCNN-ATT[7]得到的关系预测概率得分，$\max(o)$是一个限制硬标签效果的缩放常数。在训练过程中，软标签r_i替代了硬标签。因为软标签是动态改变的，所以在每个训练回合中，相同实体对的软标签有可能不一样。该方法修正了假负例和假正例的标签，对模型性能的提升有一定的效果。

2019 年，Jia[39]把焦点转移到注意力的正则化方法上。他通过观察 Zhou[18]的 BiLSTM-ATT 模型，发现大部分的注意力权重落在了两个实体的位置上，而忽略了其他位置的单词。为此，他提出了一种基于 BiLSTM 的注意力正则化的降噪框架。模型的一个重要工具是关系模式提取器（Relation Pattern Extraction），初始抽取器只是将原始训练集中的两个实体之间的单词作为关系模式，并将其计数。对于每种关系类型，只保留前 10%（最多 20 个）模式。对于T个单词的句子，如果一个词属于关系模式 M 或者是实体词，则将它对应的位置设为 1，由此构建一个注意力引导向量a^m：

$$b_i = \begin{cases} 1, & x_i \in \{e_1, e_2, m\} \\ 0, & else \end{cases} \tag{2-16}$$

$$a^m = \left\{ \frac{b_k}{\sum_{i=1}^{T} b_i} \right\}_{k=1}^{T} \tag{2-17}$$

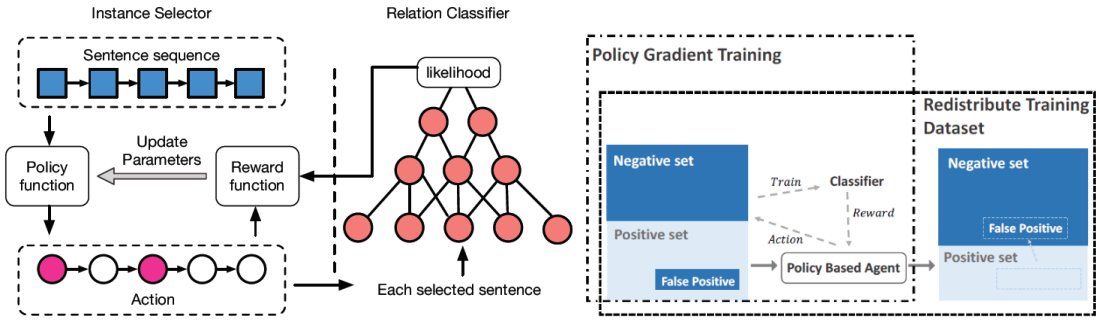
然后利用 KL 散度（Kullback-Leibler divergence）估计注意力权重a^s和注意力引导向量a^m分布的差异，希望最小化两者的差异，所以将 KL 散度定义为正则化损失$loss_a$：

$$loss_a = KL(a^m|a^s) = \sum a^m log \frac{a^m}{a^s} \tag{2-18}$$

该损失与分类器的损失相加作为模型的最终损失。该方式迫使模型关注能解释关系标签的那些模式，帮助模型准确计算示例的置信度c，筛选可信赖的示例用于下一轮的训练。

受生成式对抗网络（Generative Adversarial Networks）的启发，Qin[40]介绍了一种生成式学习框架，以学习句子级别的真正例生成器。把生成器（Generator）生成的正样本视为训练判别器（Discriminator）的负样本，直到判别器的鉴别能力下降到最大，就获得了最佳的生成器。接下来使用该生成器对训练数据集进行过滤，将假正例重新分配到负样本集中，从而为关系分类提供一份干净的数据集。

另一种构建干净数据集的方式是结合强化学习（Reinforcement Learning）。强化学习的主要思想是智能体在与环境交互的过程中，为了获得最高的奖励值，以不断试错的方式进行学习。在过滤噪声时，将过滤样本的过程建模为一种序列决策过程，利用强化学习训练过滤的策略。



a) 示例选择器和关系分类器[41] b) 深度强化学习过程[42]

a) Instance selector and relation classifier[41] b) Deep reinforcement learning process[42]

图 2-5 基于强化学习的关系抽取过程

Figure 2-5 Process of relation extraction based on reinforcement learning

Feng[41]和 Qin[42]两人分别提出了基于强化学习的远程监督关系分类方法。Feng 的模型有两个模块：示例选择器（Instance Selector）和关系分类器（Relation Classifier），见图 2-5（a）。示例选择器根据 Policy Function 对包内的句子逐个筛选，执行“选”或“不选”两种动作，筛选完成后生成一份干净的数据集。关系分类器对这份数据集进行评估，预测结果的平均似然值作为奖励（Reward）回馈给示例选

择器，并使用 Policy Gradient 更新 Policy Function 的参数。两个模块交替训练，从而共同优化示例选择器和关系分类器。

Qin 认为多示例和注意力机制并非最理想的降噪方法，被错误标记的样本仍然存在于训练集中，作为训练数据影响抽取效果。因此，Qin 提出了一个根本性的解决方案，如图 2-5（b）。通过强化学习训练一个假正例的判别器，它可以识别出数据集中的假正例，并加入到负样本集中，无需额外的监督信息。不同于之前研究中将负例直接移除的方式，Qin 把正例和负例正确分类，充分利用正负样本的信息。与 Feng 方法的区别是，前者的奖励来自于预测概率，后者的奖励采用的是本轮与前一轮的 F1 差值，该值反应了关系分类器抽取性能的改变。

本小节针对降噪的两个角度——抑制噪声和过滤噪声，分别进行了梳理和总结。受上述方法的启发，本文也试图利用自注意力，优化句子的表示，抑制包中的噪声，提出了基于 Transformer 和句子级自注意力机制的关系抽取方法。

2.5 本章小结

在本章中，主要围绕关系抽取任务分为有监督、无监督、半监督和远程监督四个方面讨论近些年来的理论研究进展和成果，对本文的重点远程监督关系抽取部分展开详细描述。远程监督关系抽取的研究方向可以大致分为优化特征提取和降噪两个方面。前者，优化特征提取包括选择合适的特征提取器和结合实体的相关信息两个思路；后者，针对远程监督的回标噪声，可分为抑制噪声和过滤噪声两种解决办法。在分析前人的基础上，本文提出了两种注意力机制模型提升关系抽取的效果。

3　注意力增强包表示的关系抽取模型

当前的远程监督关系抽取方法试图通过多示例学习和上下文信息来减轻噪声，更有效地指导关系分类。但是这些模型偏向于以高精度识别有限的一组关系，而忽略了长尾实体的关系。为了解决这个问题，在本文中我们以 Alt[8]提出的 DISTRE 作为基础模型。DISTRE 利用了 OpenAI 团队的 Transformer 预训练语言模型 GPT[43]，GPT 不仅可以捕获语义和句法特征，还涵盖海量的常识性知识，并且它的微调效率和合理的硬件要求是一大优势。通过将 GPT 扩展到包级别上，结合选择性注意力机制，较合理地处理多示例训练和远程监督数据集的关系预测。尽管 DISTRE 在 NYT 数据集上取得了 SOTA（state-of-the-art）的效果，但是我们注意到该模型忽视了两个实体与句子之间的相关性，提出了实体注意力机制模型 DISTRE-EA。另一方面，虽然通过 Transformer 编码得到的句子表示语义信息丰富，但是从包级别的概念上分析，包内部的句子本质上是相互独立的个体，因此为了建立句子之间的联系，我们在句子级别上实现了自注意力机制的关系抽取模型 DISTRE-SA。上述两种模型中，都使用了注意力机制，以此增强了包的嵌入表示，有效优化了模型提取的特征并降低了噪声带来的影响。

3.1　问题定义

远程监督假设如果两个实体对在知识库中存在某种关系，那么包含这一对实体的所有训练句子都标记为知识库中的关系。如图 3-1 所示，在知识库 Freebase 中存储的三元组中，以实体对为标识，对语料中所有包含该实体对的句子打上关系标签。

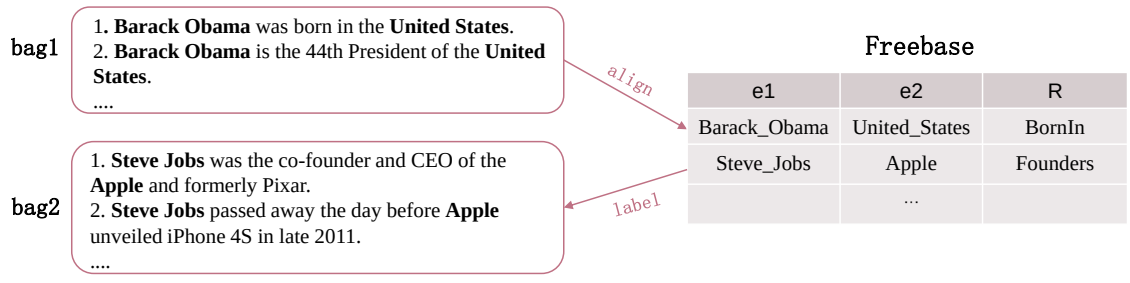


图 3-1　知识库回标语料

Figure 3-1 Labeled corpus by knowledge base

假设一个已标记的数据集$D = \{(x_i, head_i, tail_i, label_i)\}_{i=1}^{N}$，每个句子序列（输入序列）由$m$个单词组成$x_i = [x^1, \ldots, x^m]$。$head_i$和$tail_i$是输入句子$x_i$中两个实体

的位置索引，$label_i$是自动回标知识库时分配的关系标签。因为远程监督的假设太过强烈，产生了错误标签的问题，导致关系标签$label_i$是置信度较低的。为了降低噪声的影响，引入了多示例学习[5]的方法。在多示例学习中，以每一对实体对$(head, tail)$作为唯一标识，构建集合$B = \{x_1, \dots, x_n\}$。在集合B中，每个示例句子x_i同时包含相同的实体对，这个B称为一个包。包中句子的数目是不尽相同的，换句话说n是变化的。此时，我们并不关注包中每个句子表示的具体关系，只关注整个包的关系标签。因此，模型的关系预测从句子级别转到了包级别上。

预定义一个关系集合$R = \{r_1, \dots, r_{d_r}\}$，$d_r$代表关系的总数目。实体对$(e_1, e_2)$以及$e_1$和$e_2$之间的关系标签$r_{e1,e2} \in R$。包含实体对$(e_1, e_2)$的所有句子$x_1, \dots, x_n$组合成一个包$B$。远程监督关系抽取的任务是给出实体对$(e_1, e_2)$在包$B$中存在关系 r 的概率$p(r_i|B)$，其中$1 \le i \le d_r$，$r_i \in R$。

3.2 基于 Transformer 的关系抽取模型

近年来，预训练语言模型[11]在多项 NLP 任务中均取得了非常优秀的成果，刷新了多项记录。语言模型（Language Model）根据上下文信息预测目标词的表示，为一段文本生成它的概率分布。这个过程无需人工标注语料，所以语言模型可以从无限制的大规模语料中学习每个词的丰富语义表示。预训练语言模型中的预训练是指，首先该模型依据特定的任务进行训练过程，然后将获得的一套模型参数对该模型进行初始化。最后使用统一的模型，或者是当成特征直接加到一些简单模型上，以对接不同的 NLP 任务。

DISTRE（Distantly Supervised Transformer for Relation Extraction）正是利用了 OpenAI 团队的 Transformer 预训练语言模型 GPT（Generative Pre-trained Transformer）编码句子嵌入表示。并且，在对接关系抽取任务时使用选择性注意力机制，降低了噪声的影响并实现多示例学习。

3.2.1 Transformer 解码器

基于 Transformer 的关系抽取模型使用单向的解码组件，通过堆叠 TF 块（Transformer Block）的方式搭建深层的网络，并在解码组件中大量使用多头自注意力机制（Multi-headed Self-attention），引入多头自注意力的目的是为了捕获不同表示子空间上的相关信息。由于没有使用原始 Transformer 的编码层，所以模型的解码层没有对应的编码-解码注意力层。

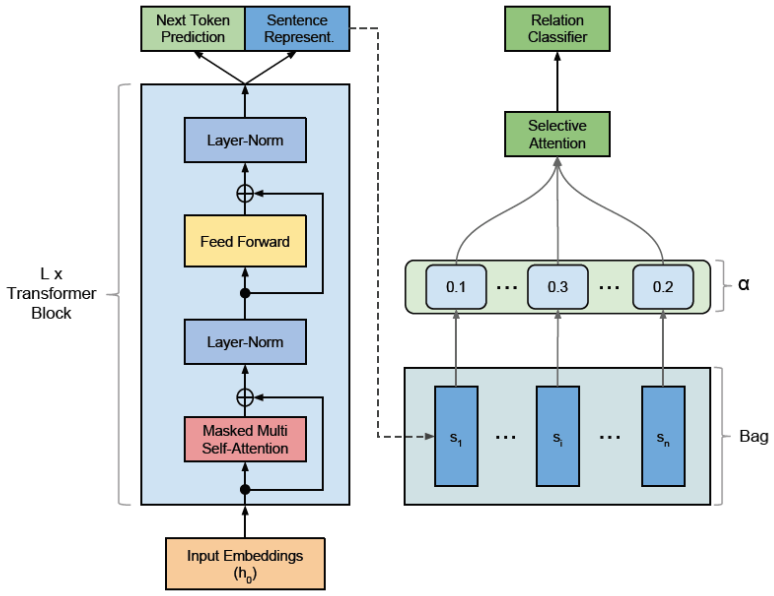


图 3-2 DISTRE 模型框架[8]

Figure 3-2 DISTRE model framework[8]

如图 3-2 左侧所示，解码组件由相同数量的解码器构成，每个解码器结构相同，但是参数不共享。在这里，一个 TF 块相当于一个解码器。每个 TF 块都接受一个向量列表并输出一个向量列表。其中，词嵌入过程只发生在最底层的 TF 块中，该层的输入向量列表由词向量和位置向量组成。而在其他 TF 块中，输入向量列表就是紧邻的前一层 TF 块的输出向量列表。

最底层的 TF 块的输入向量列表h_0计算如下：

$$h_0 = TW_e + W_p \tag{3-1}$$

其中，W_e表示词向量查找表矩阵，模型的词向量表示是随机初始化的，所以对应的W_e初始值也是随机初始化的。T矩阵由每个单词的索引对应的独热编码（One-hot encoder）组成。TW_e表示输入序列通过索引即可构成词向量列表，其行向量$e_t^p \in \mathbb{R}^d$代表输入序列的第p个位置的词向量。W_p是一个位置向量表示矩阵，其行向量$e_p \in \mathbb{R}^d$代表输入序列的第p个位置的位置向量。因为 Transformer 模型本身不捕获输入序列中单词的前后顺序关系，因此在最低层中可加入单词的位置信息。

进一步，在其他 TF 块中，输入向量列表h_l就是紧邻的前一层 TF 块的输出向量列表：

$$h_l = tf\_block(h_{l-1}) \quad \forall\, l \in [1, \mathrm{L}] \tag{3-2}$$

其中，L是 TF 块的数量。h_l也可以认为是第l层的隐状态序列，其中的向量$h_l^p \in \mathbb{R}^d$代表第p个位置的单词对应于第l层的隐状态。

具体来说，一个 TF 块由自注意力层和前馈神经网络层组成，如图 3-2 所示。在自注意力层里，每个词都要和该句子中的所有词进行缩放点积相似度计算，目的

是学习句子内部的词依赖关系，捕获句子的内部结构。我们将在 3.4 节中详细介绍自注意力机制的实现原理。然后，自注意力层的输出传递到前馈神经网络层。前馈神经网络层由多个前馈神经网络组成，每个单词对应的位置上都使用一个相同结构但互不共享参数的前馈神经网络，其形式化表示如下：

$$FFN(x) = \max(0, xW_1 + b_1)\,W_2 + b_2 \tag{3-3}$$

其中，W_1和W_2为线性变换的参数矩阵，b_1和b_2为偏置项，$\max(0,\cdot)$表示 ReLU 激活函数。为了更好地优化深层网络，在注意力层和前馈神经网络层之后都设有一个残差模块，对每层进行求和并且归一化，达到加速模型训练并提升稳定性的效果。

在最后的第 L 层，获得输入序列对应的隐状态序列h_L后，以最后一个位置m的隐状态h_L^m作为输入序列最终的嵌入表示。

3.2.2　无监督预训练语言模型表示

对于给定的一个语料$C = \{c_1, \ldots, c_n\}$，语言模型的目标是最大化似然函数，如公式（3-4）所示：

$$L_1(C) = \sum_i \log p(c_i | c_{i-1}, \ldots, c_{i-k}; \theta) \tag{3-4}$$

其中，上下文窗口大小为k，根据前k个单词预测当前词c_i出现的条件概率。此时的条件分布概率p使用上述的 Transformer 解码器定义为：

$$p(c) = softmax(h_L W_e^{\mathrm{T}}) \tag{3-5}$$

式中，h_L是最顶层的隐状态序列，W_e是词向量查找表矩阵，θ是整个语言模型的优化参数。$h_L W_e^T$表示每个隐状态h_L^p与每个词相近的程度，进而归一化后代表预测为每个词的概率。最大概率值对应的单词就是当前输入序列要预测的词，对应图 3-2 的“Next Token Prediction”。

3.2.3　基于 Transformer 的多示例学习

本小节将阐述如何利用 Transformer 解码器编码的句子嵌入表示，结合选择性注意力（Selective Attention）机制，实现多实例学习。最后，通过优化语言模型的目标与多分类任务的目标，训练基于 Transformer 的关系抽取模型。

远程监督的噪声是指包中的每个句子不一定表达出知识库中对应实体对的关系。如表 3-1 中，包的关系类别是“place of birth”，最后一列的“Yes”或“No”意味着该句子是否表达了这种关系。可以看到，包 B1 中的第二个句子不体现实体对的关系类别。

表 3-1 噪声句子举例
Table3-1 Examples of noise sentences

bag	sentence	correct
B1	1.**Barack Obama** was born in the **United States**.	Yes
	2.**Barack Obama** is the 44th President of the **United States**.	No

为了缓和回标噪声的影响，DISTRE 模型加入了选择性注意力机制[7]。如图 3-2 右侧所示，为了利用所有句子的信息，模型使用包$B = \{x_1, \dots, x_n\}$预测关系r。显然，包B的表示取决于所有的句子嵌入表示$s_1, \dots, s_n$，每个句子嵌入表示s_i包含有关实体对(e_1, e_2)是否表达关系r的信息。

在一个包中为每一个句子赋予不同的注意力权重，加权求和得到包B的嵌入表示b_s：

$$b_s = \sum_i \alpha_i s_i \tag{3-6}$$

上式中，α_i是分配给对应的句子嵌入表示s_i的权重。将一个句子中的词向量表示输入到 Transformer 预训练语言模型中，在第L层取最后一个位置的隐状态表示h_L^m作为句子的嵌入表示s_i。

注意力权重α_i有两种计算方式。一是平均注意力（Average Attention），二是选择性注意力。在平均分配权重的方法中，每个句子的嵌入表示对包的表示是贡献均等的。由于不可避免地会出现错误的标签问题，如果对每个句子一视同仁，那么错误的标记句子将在训练和测试过程中带来更加巨大的噪音。相比之下，选择性注意力机制赋予贡献较大的句子高的权重，赋予贡献较小的句子低的权重，这种不均匀的分配方式有明显的优势。选择性注意力机制可以学会识别具有最清晰表达关系的特征的句子，而不再强调那些噪声句子。所以，DISTRE 使用选择性注意力来降低噪声句子的影响，注意力权重α_i定义为：

$$\alpha_i = \frac{\exp(s_i e_r)}{\sum_{j=1}^{n} \exp(s_j e_r)} \tag{3-7}$$

其中，$e_r \in \mathbb{R}^d$是目标关系的嵌入表示。那么$s_i e_r$代表的是句子与目标关系的相关性程度。值得注意的是，在测试阶段时并不知道目标关系的标签，所以把每个句子与所有的关系类别都进行了相似度计算。

为了计算包B属于关系标签$label$的概率$p(label)$，把包的嵌入表示b_s输入到分类器中。分类器由线性层和 softmax 函数组成：

$$p(label|B, \theta) = softmax(W_r b_s + b) \tag{3-8}$$

在上式中，$b \in \mathbb{R}^{d_r}$是一个偏置项，d_r是关系的总数目。$W_r \in \mathbb{R}^{d_r \times d}$是由所有关系

的嵌入表示e_r组成的矩阵。

在微调阶段，需要最大化多分类任务的目标函数：

$$L_2(D) = \sum_{i=1}^{|B|} log\ p(label_i|B_i, \theta) \tag{3-9}$$

根据 Radford[43]所做的实验表明，在微调阶段把语言模型作为辅助目标可以提升模型的泛化能力，并且加快模型的收敛。因此，最终的目标函数由公式（3-4）和公式（3-9）组合而成：

$$L(D) = \lambda * L_1(D) + L_2(D) \tag{3-10}$$

在上式中，λ代表语言模型的权重，在实验过程中作为超参数进行调整。通过随机梯度下降算法（Stochastic Gradient Descent Algorithm）可更新模型参数。

3.3　实体注意力机制模型

在上一节中，选择性注意力机制把重点放在了如何降低噪声句子带来的影响上。虽然经过 Transformer 预训练语言模型能够得到具有丰富信息的句子嵌入表示，但是关系抽取不是单纯的句子分类任务，最终目标是正确地预测实体 1 和实体 2 的关系，因此我们提出了基于 Transformer 预训练语言模型和实体注意力机制的关系抽取模型 DISTRE-EA。本节从研究动机和网络结构两方面介绍。

3.3.1　研究动机

基于 Transformer 预训练语言模型和实体注意力机制的关系抽取模型 DISTRE-EA 的提出主要基于以下动机：

作为关系抽取任务，我们认为实体的作用不容忽视。通过实体的嵌入表示，可以从外部知识库中补充实体的描述信息[32]或者添加实体的类型信息[44]。在使用了实体的相关信息后，关系抽取模型的效果能够获得提升。因此，我们想在没有明确考虑实体信息的 DISTRE 模型中，同样引入实体的相关信息。

但基于 Transformer 预训练语言模型的方法不同于其他利用实体相关信息的方法。预训练语言模型的网络结构一般而言是固定的，不适合改变。并且，预训练语言模型的参数是经过大规模语料充分训练的。改变结构将一同损失已经训练好的参数，如此便失去了预训练语言模型提取特征的优势。

因此，在不改变 Transformer 预训练语言模型的基础上，我们通过注意力机制，将实体的相关信息引入模型。不改变 Transformer 编码句子嵌入表示的过程，而使用实体注意力机制增强包的嵌入表示，进而达到降低句子噪声和提取与任务相关

特征的目的。

3.3.2 网络结构

图 3-3 展示了基于 Transformer 预训练语言模型和实体注意力机制的关系抽取模型 DISTRE-EA 的网络结构。如图 3-3 右所示，经过 Transformer 编码得到包中每个句子的嵌入表示$s_1, \ldots, s_n$后，除了与目标关系标签r做选择性注意力计算外，我们还把实体的嵌入表示与每个句子做注意力计算。通过实体与句子的相关性，判断哪些句子的表示包含了更多与实体相关的信息。

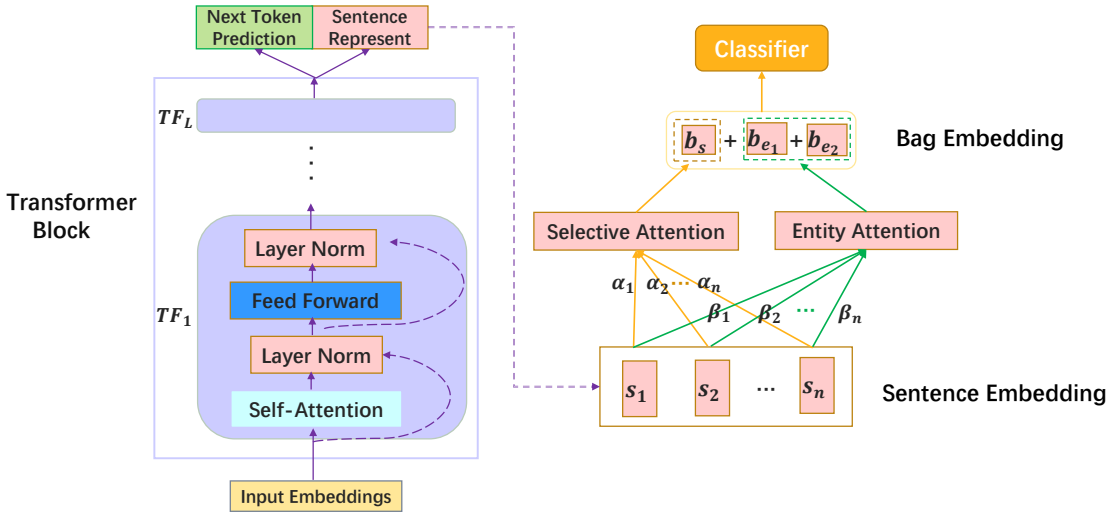


图 3-3 实体注意力机制模型 DISTRE-EA

Figure 3-3 Entity attention mechanism model DISTRE-EA

具体来说，以实体 1 为例，通过实体注意力机制得到的包的嵌入表示$b_{e_1} \in \mathbb{R}^d$如式（3-11）所示：

$$b_{e_1} = \sum_i \beta_i s_i \tag{3-11}$$

如果句子s_i中含有的内部信息与实体 1 相关度越高，那么相应的句子权重β_i的值就越高。每个句子由实体 1 指导的权重β_i定义如下：

$$\beta_i = \frac{\exp\left(f(s_i, e_1)\right)}{\sum_{j=1}^{n} \exp\left(f(s_j, e_1)\right)} \tag{3-12}$$

式中，$e_1 \in \mathbb{R}^d$是实体 1 的嵌入表示，$f(s_i, e_1)$为句子与实体 1 的相似度计算函数。包中的每个句子与实体 1 分别通过该函数计算后，归一化即可得到该句子的权重贡献大小。$f(s_i, e_1)$作为度量句子与实体相似度的函数，通常定义为如下形式：

$$f(s_i, e_1) = s_i W e_1 \tag{3-13}$$

其中，W$\in \mathbb{R}^{d \times d}$是参数矩阵。由于使用 OpenAI 团队的 Transformer 预训练语言模

型 GPT[43]，在该预训练模型中词向量的维度d设置成 768，那么 W∈ $\mathbb{R}^{768\times768}$，参数量过于庞大，不利于优化相似度计算函数。所以我们将$f(s_i, e_1)$定义成以下形式：

$$f(s_i, e_1) = s_i W_1 W_2 e_1 \tag{3-14}$$

式中，$W_1 \in \mathbb{R}^{d\times q}$，$W_2 \in \mathbb{R}^{q\times d}$。在实验过程中，最佳的 q取值为 32。与之前的 W 矩阵参数量相比，这种方式大大减少了相关度计算函数的参数量。

在上述中，我们基于实体 1 得到了加权后的包嵌入表示b_{e_1}。同理对于实体 2 也进行一样的计算，可以得到基于实体 2 的包嵌入表示b_{e_2}。另一方面，结合公式（3-6）的选择性注意力机制得到的包嵌入表示b_s，我们就能获得一个新的包嵌入表示b_{total}：

$$b_{total} = (1-2\eta) * b_s + \eta * (b_{e_1} + b_{e_2}) \tag{3-15}$$

其中，η表示基于实体注意力机制的包嵌入表示的权重。

由此，新的计算包B属于关系标签$label$的概率$p(label)$计算如下：

$$p(label|B, \theta) = softmax(W_r b_{total} + b) \tag{3-16}$$

公式（3-16）替换了上述公式的（3-8）。在对模型进行训练时，使用公式（3-10）定义相同的目标函数和优化策略。

基于实体注意力增强的包嵌入表示b_{total}，主要从实体的角度出发，一方面引入了任务导向的信息（实体相关信息），另一方面兼顾了注意力机制的降噪功能，为远程监督关系抽取任务的分类器提供更优的包嵌入表示。

3.4　句子级自注意力机制模型

在本节中，我们提出了基于 Transformer 预训练语言模型和句子级自注意力机制的关系抽取模型 DISTRE-SA，将从研究动机、网络结构、缩放点积注意力和多头注意力几个方面详细阐述其实现细节。

3.4.1　研究动机

基于 Transformer 预训练语言模型和句子级自注意力机制的关系抽取模型 DISTRE-SA 的提出主要有以下动机：

目前，包中的每个句子的嵌入表示是通过 Transformer 预训练语言模型编码获得的，因此包中句子的表示彼此相互独立，没有关联。但由于包中存在噪声句子，我们希望根据同一包中句子整体的分布，强化非噪声句子的表示，弱化噪声句子的表示，由此达到抑制噪声的效果。出于这种考虑，在句子之间使用自注意力机制，可以使得每个句子的嵌入表示都会考虑其他句子的信息，以此作为依据调整自身

的嵌入表示。所以，基于句子级的自注意力机制，能从优化句子嵌入表示的角度，增强包的嵌入表示，进而达到降低包中句子噪声和提取与任务相关特征的目的。

3.4.2　网络结构

图 3-4 展示了基于 Transformer 预训练语言模型和句子级自注意力机制的关系抽取模型 DISTRE-SA 的网络结构。其结构主要由句子编码模块、句子级别的自注意力模块和选择性注意力模块三个部分组成。

句子编码模块中，输入序列中的每个位置的单词都存在一条独特的路径流入 TF 块。在自注意力层，这些路径之间存在依赖关系。而在前馈层不存在依赖关系，因此在前馈层可以并行执行各种路径。通过 GPT 的 Transformer 预训练语言模型，我们就得到了包中每个句子的嵌入表示$s_1, \ldots, s_n$。

句子级别的自注意力模块中，自注意力机制由缩放点积注意力机制和多头注意力机制组成。其中，缩放点积注意力机制大体上分为两个步骤：（1）第一步是对于每个句子嵌入表示s_i创建三个向量，分别是查询向量（query vector）、键向量（key vector）、值向量（value vector）。并利用查询向量和键向量计算注意力的权重。（2）第二步是权重归一化后，值向量按照权重加权求和得到新的句子嵌入表示$s_1', \ldots, s_n'$。多头注意力机制是将上述过程重复多次，且为每个头保持独立的查询参数矩阵、键参数矩阵和值参数矩阵，从而产生不同的查询向量、键向量和值向量。这样做的目的是为了扩展模型专注于不同位置的能力。

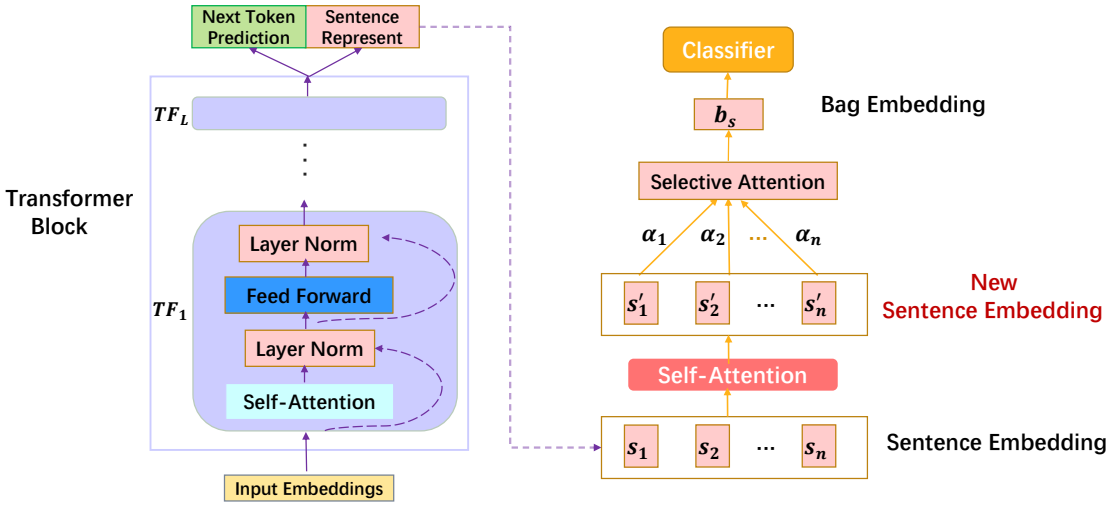


图 3-4 句子级自注意力机制模型 DISTRE-SA

Figure 3-4 Sentence-level self-attention mechanism model DISTRE-SA

选择性注意力模块中，得到新的句子嵌入表示s_i'后，通过关系标签的注意力指导，在包的内部之间计算句子嵌入表示的权重，对高质量的句子赋予较高的权重，

降低噪声句子的影响。加权求和得到包的嵌入表示b_s，传递到分类器中完成关系分类的任务。

3.4.3　缩放点积注意力

缩放点积注意力（Scaled Dot-Product Attention）机制可以描述为将查询向量和一组键值对映射到输出的过程。输出向量的表示由值向量加权求和得到，其中，分配给每个值向量的权重是通过查询向量和对应的键向量的缩放点积运算得到的。

图 3-5 展示的是缩放点积注意力的实现细节。为了计算方便，我们从矩阵的角度来进行说明。对于包中每个句子的嵌入表示$s_1, \dots, s_n$，其中$s_i \in \mathbb{R}^d$，我们将其组合成一个矩阵$S_{all} = [s_1, \dots, s_n]^\intercal$。$n$是包中句子的数目，$S_{all}$的每一行代表一个句子的嵌入表示。将$S_{all}$乘以待训练的参数矩阵$W^Q \in \mathbb{R}^{d \times d_k}$、$W^K \in \mathbb{R}^{d \times d_k}$、$W^V \in \mathbb{R}^{d \times d}$，得到查询矩阵$Q$、键矩阵$K$、值矩阵$V$：

$$Q = S_{all} W^Q \tag{3-17}$$

$$K = S_{all} W^K \tag{3-18}$$

$$V = S_{all} W^V \tag{3-19}$$

上式中，每一个句子嵌入表示s_i与三个随机初始化的参数矩阵相乘，从而创建了三个向量：查询向量$q_i \in \mathbb{R}^{d_k}$、键向量$k_i \in \mathbb{R}^{d_k}$、值向量$v_i \in \mathbb{R}^d$。

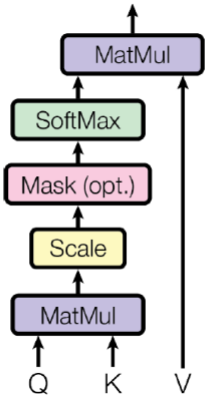


图 3-5 缩放点积注意力[10]

Figure 3-5 Scaled Dot-Product Attention[10]

进而利用上述的矩阵Q、K、V，计算包中所有句子新的嵌入表示$S'_{all} \in \mathbb{R}^{n \times d}$，如公式（3-20）所示：

$$S'_{all} = Attention(Q, K, V) = softmax(\frac{QK^\intercal}{\sqrt{d_k}})V \tag{3-20}$$

其中，$1/\sqrt{d_k}$是为了使反向传播的梯度更加稳定，也是“缩放”一词的原因。对于上述计算公式的直观理解是：$QK^\intercal$计算得到了自注意力的权重，值矩阵V中的每个

值向量v_i按照 softmax 归一化后的权重加权求和，得到新的句子嵌入表示s_i'。

显然，对于每一个句子，其自身的通过自注意力计算的权重应该在所有句子中最高，但句子的新的嵌入表示并非全部关注在自身的嵌入表示上，也会关注与之相关的其他句子。

另外，除了上述的点积注意力，还有一种常用的注意力函数，被称为加性注意力（Additive Attention）[9]，后来也常被称为一般性（general）注意力。其计算相关度函数定义为一个单层的前馈神经网络。虽然在理论上，一般性注意力和点积注意力的复杂度相似，但实际应用中，由于点积注意力可以使用高度优化后的矩阵乘法来实现，所以运算速度更快且空间利用率更高。这也是自注意力机制的一大优势。

3.4.4　多头注意力

通过增加一种多头注意力（Multi-Head Attention）的机制，进一步完善了自注意力层。它的作用是可以捕获注意力层多个表示子空间（Representation Subspaces）的信息。通过创建多组随机初始化的查询、键、值的参数矩阵，把输入向量映射到不同的表示子空间中。

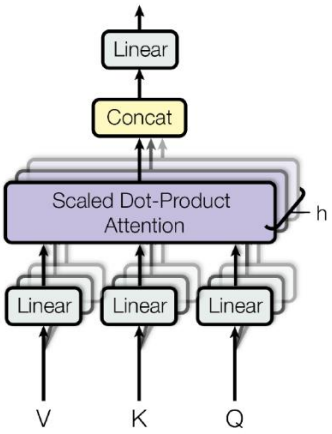


图 3-6 多头注意力由多个并行运行的注意力层组成[10]

Figure 3-6 Multi-Head Attention consists of several attention layers running in parallel[10]

多头注意力的结构如图 3-6 所示。在该模块中，为每一个头$head_i$保持独立的查询参数矩阵、键参数矩阵和值参数矩阵。对于包中句子的嵌入表示组成的矩阵S_{all}，经过公式（3-17）~公式（3-20）的计算，可以得到包中句子新的嵌入表示S'_{all}。但值得注意的是，在多头注意力机制下，这个过程需要计算h次，每次都与不同的参数矩阵集进行运算，这便是“多头”的含义：

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3-21}$$

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^O \tag{3-22}$$

式中，$i \in [1, \dots, h]$，h代表头的数目，$W^O \in \mathbb{R}^{hd \times d}$。将这$h$次多头的结果$head_i$进行

拼接，经过全连接矩阵W^O的线性变换后，我们就得到了公式（3-22）的输出结果，也就是新的句子嵌入表示$S'_{multi_all} \in \mathbb{R}^{n\times d}$。

由此，结合缩放点积注意力和多头注意力，我们就完成了自注意力机制模块。将变换得到的句子嵌入表示传递到选择性注意力中，即替换公式（3-6）中的原始句子嵌入表示s_i。完成包内部的降噪处理后，获得包嵌入表示b_s。最后将包的嵌入表示输入到分类器中，实现关系抽取任务。

3.5 本章小结

本章主要分为四部分内容。首先给出了远程监督关系抽取的问题定义。其次，从基础模型 DISTRE 出发，介绍了 Transformer 语言模型的结构，语言模型的目标函数以及如何通过结合选择性注意力机制解决多示例学习问题。接着，从研究动机出发，提出了基于 Transformer 预训练语言模型和实体注意力机制的关系抽取模型 DISTRE-EA，围绕实体注意力展开其网络结构。最后，介绍了我们提出的另一个基于 Transformer 预训练语言模型和句子级自注意力机制的关系抽取模型 DISTRE-SA，介绍了其网络结构，并重点描述网络结构中的缩放点积注意力模块和多头注意力机制模块。

4 实验设计与结果评估分析

实验环节是验证模型有效性的重要手段。通过在标准数据集上进行训练，使用常用的评估指标与主流的模型方法进行对比，从而验证我们提出的实体注意力机制模型 DISTRE-EA 和句子级自注意力机制模型 DISTRE-SA 的实验效果。在本章中，我们先介绍常用的数据集以及评估方法。然后，依据模型的特点介绍了预训练、优化方法和超参数设置的实验细节。最后，以图表的形式直观地展示了实验的结果并结合模型的特点进行了分析与总结。通过实验，验证了我们的两个模型在远程监督关系抽取任务中的有效性。

4.1 数据集简介

Riedel[5]发表的 NYT 数据集是远程监督关系抽取任务的标准数据集，通过将 Freebase 知识库与纽约时报语料的关系保持一致而生成的。其中纽约时报语料发表在 2005-2006 年之间的作为训练集，2007 年的作为测试集。我们使用由 Lin[7]等人预处理的数据集版本。

训练数据包括 522611 个句子，281270 个实体对和 18252 个相关事实。测试数据包括 172448 个句子，96678 个实体对和 1950 个相关事实。相关事实指的是真正例的数目，实体对的数目指的是包的个数。在训练阶段，两个实体和目标关系的组成的集合作为包的唯一标识；在测试阶段，两个实体的集合作为包的唯一标识。数据集的统计数目见表 4-1。数据集中包含 53 种关系，如果给定的实体对不存在任何关系，则此类关系记做 NA。

表 4-1 NYT 数据集
Table4-1 NYT10 dataset

数据集划分	句子数	实体对	相关事实	关系类别
Train	522611	281270	18252	53
Test	172448	96678	1950	53

53 种关系数目是依据 relation2id 文件定义的，该文件把关系类别转换成关系 id。每个数据样本的格式如表 4-2 所示，包含两个实体在 Freebase 中的 id 标识，实体的具体名字，关系类别和句子。

表 4-2 数据格式
Table4-2 Data format

项目	内容
Entity1_id	m.0ccvx
Entity2_id	m.05gf08
Entity1_name	queens
Entity2_name	belle_harbor
Relation	/location/location/contains
Sentence	she is a daughter... belle_harbor , queens .

4.2 评估指标

4.2.1 P-R 曲线

P-R 曲线实际上是以精准率（Precision）和召回率（Recall）这两个变量绘制出的曲线，其中召回率为横坐标，精准率为纵坐标。首先，根据是否预测为该类和是否与真实值一致两个维度，可以用一个混淆矩阵表示这四种情况，如下表 4-3 所示：

表 4-3 混淆矩阵
Table4-3 Confusion matrix

混淆矩阵		真实值 True	真实值 False
预测值	Positive	TP	FP
预测值	Negative	TN	FN

TP（True Positive）：属于该类的样例，被正确判断为该类，样例数记为 TP

TN（True Negative）：不属于该类的样例，被正确判断为其他类，样例数记为 TN

FP（False Positive）：不属于该类的样例，被错误判断为该类，样例数记为 FP

FN（False Negative）：属于该类的样例，被错误判断为其他类，样例数记为 FN

进而，可以定义关系抽取任务的精准率和召回率：

$$Precision = TP/(TP + FP) \tag{4-1}$$

$$Recall = TP/(TP + FN) \tag{4-2}$$

可以看出，精准率指的是在所有预测为正例的数据中，真正例所占的比例。召

回率是指预测为真正例的数据占所有正例数据的比例[45]。精准率和召回率是一对矛盾的度量，一般来说，精准率高时，召回率往往偏低；召回率高时，精准率往往偏低。根据分类器的预测结果的概率由高到低对测试样本进行排序，其中排在前面的是分类器认为最可能的属于预测类的样本，排在后面的是分类器认为最不可能属于预测类的样本。按此顺序逐个计算已经观测到的样本的精准率和召回率，以此绘制 P-R 曲线。

一般情况下，如果一个分类器的 P-R 曲线完全低于另一个分类器的 P-R 曲线，则可断言后者的性能优于前者。

4.2.2 AUC

AUC (Area under the Curve)，意思是曲线下的面积，介于 0 和 1 之间。在远程监督关系抽取任务中，该曲线指的是 P-R 曲线。在 P-R 曲线中，如果分类器 A 和分类器 B 的两条曲线没有绝对的高低关系，就很难区分谁的性能更优。因此，以 P-R 曲线下的面积作为衡量指标就是 AUC 指标。AUC 作为数值可以直观地评价分类器的好坏，值越大侧面反映了精准率和召回率越高，分类器的分类效果越好。

4.2.3 Precision@K

Precision@K 反映了在测试集中把预测的结果按照预测标签$\hat{y}_t$的概率从高到低排序后，前 K 个测试结果的精准率，由如下公式定义：

$$Precision@K = \frac{\sum_{t \le K} I(\hat{y}_t = y_t)}{K} \tag{4-3}$$

其中，$t \le K$表示按照预测标签$\hat{y}_t$的概率从高到低排序后的前 K 个测试结果。$I(\hat{y}_t = y_t)$是指示函数，当预测标签$\hat{y}_t$与真实标签y_t一致时，结果为 1；否则，结果为 0。本实验使用的 Precision@K 指标分别有 Precision@100，Precision@200，Precision@300，Precision@500，Precision@1000 和 Precision@2000。

4.3 实验细节

4.3.1 预训练

由于预训练对计算机的硬件资源有较高的要求，并且需要消耗大量的时间。另一方面，我们的目的是在远程监督关系抽取任务上利用预训练的参数列表，为后续

的改进奠定基础。因此我们使用 OpenAI 团队中 Radford[43]等人发表的 Transformer 语言模型。他们的模型在 BooksCorpus 语料库[46]上进行了训练，其中包含约 7000 种未出版的书籍，总共有超过 800M 不同体裁的单词。该模型设置了 12 个 TF 块和 12 个注意力头，768 维的向量列表以及 3072 维的前馈层隐状态向量。在预处理方面，使用了字节对（byte-pair）[47]对词汇表进行编码，添加了特定任务独有的 tokens，比如“start”，“end”和“delimiter”。

4.3.2 优化方法

在模型优化器上，本文采用了 Adam[48]优化算法实现梯度的反向传播。Adam 算法的对角缩放（Diagonal Rescaling）具有不变性，因此很适合求解带有大规模数据或参数的问题。同时兼顾较高的计算效率和较低的内存需求，适用于解决高噪声或稀疏梯度的任务。在本实验中，设置指数衰弱速率$\beta_1 = 0.9$，$\beta_2 = 0.999$。

为了避免由于随机初始化权重导致模型的振荡，我们选择预热学习率（warm-up）的方式。即先用较小的学习率进行训练，随着训练步长的增大，慢慢加大学习率的数值。直到学习率达到最初设置的大小后，再采用该数值训练模型。此时意味着预热学习率完成。但是在随后的训练过程中，学习率是衰减的。我们设置的学习率为 6.25e-5，衰减率为 0.2%。在预热和衰减率的控制下，可以使得模型逐渐趋于稳定，并且模型的收敛速度更快，效果更佳。

机器学习算法为了满足尽可能复杂的任务，其模型的拟合能力一般远远高于问题的复杂度，这就是过拟合（Overfitting）现象。过拟合具体表现在训练集上效果好，在测试集上效果差，模型的整体泛化能力弱。Dropout[49]是一种防止过拟合的方法，在训练阶段，每次以一定的概率随机忽略网络中的部分神经元，让部分神经元无法参与本次训练；在测试阶段，则需要保留所有的隐层节点。该方法的实际效果要比L_2正则化的方式更好，实现简单高效。在本文中，Transformer 模块中的注意力层 dropout 固定为 0.1，分类器的 dropout 在 DISTRE-EA、DISTRE-SA 两种模型下稍有变化。

4.3.3 超参设置

在实验过程中，由于预训练方法的限制，词向量和位置向量的维数$d = 768$，前馈神经网络中的状态维数是 3072。两个模型中共同的参数的取值范围是：TF 块的数量$L = \{6,8,10,12\}$，多头注意力的头数目$h = \{8,10,12,14,16\}$，语言模型权重$\lambda = \{0.4,0.5,0.6\}$，批处理的大小$batch\ size = \{8,16,24,32,36\}$，学习率

$learning\ rate = \{3.125e-5, 6.25e-5\}$，训练回合$epoch = \{3,4\}$，分类器的$dropout = \{0.2,0.25.0.3\}$。

（1）DISTRE-EA 模型。引入实体注意力机制后，模型特有参数的取值范围是：基于实体注意力机制的包嵌入表示的权重$\eta = \{0.1,0.2,0.25.0.3,0.35\}$，相似度计算矩阵$W_1$第二个维度$q = \{16,20,26,32,36\}$。模型的最优参数如表 4-4 所示。

表 4-4 DISTRE-EA 实体注意力模型参数
Table4-4 Entity attention model parameters

参数名称	参数含义	数值
L	TF 块的数量	12
h	多头注意力的头数目	12
λ	语言模型权重	0.5
η	实体嵌入表示权重	0.2
q	矩阵W_1第二个维度	32
batch size	批量数据大小	32
learning rate	学习率	6.25e-5
epoch	训练回合	3
dropout	分类器的 dropout rate	0.3

（2）DISTRE-SA 模型。在原模型的基础上增加包中句子之间的自注意力机制，模型特有参数的取值范围是：自注意机制的头数$self_head = \{4,6,8,10,12\}$，自注意力机制的$self_dropout = \{0.2,0.3,0.5,0.7,0.8\}$。模型的最优参数如表 4-5 所示。

表 4-5 DISTRE-SA 自注意力模型参数
Table4-5 Self-attention model parameters

参数名称	参数含义	数值
L	TF 块的数量	12
h	多头注意力的头数目	12
λ	语言模型权重	0.5
batch size	批量数据大小	16
learning rate	学习率	6.25e-5
epoch	训练回合	3
dropout	分类器的 dropout rate	0.2
self_head	自注意力的头数	6
self_dropout	自注意力的 dropout rate	0.8

4.4 实验结果与分析

4.4.1 对比模型

为了评估我们的模型效果，我们选取了以下几个具有代表性的方法进行对比实验：

（1）PCNN-ATT：Lin[7]发表在 2016 年 ACL 会议上的论文。该模型基于分段卷积神经网络，融合选择性注意力机制从而实现包内部的句子降噪，是一个经典的模型。

（2）RNN-adv：Wu[50]发表在 2017 年 EMNLP 会议上的论文。该模型在在词嵌入的水平上添加对抗性噪声，使用 RNN 框架对句子嵌入进行编码，后续结合选择性注意力机制提升关系抽取的准确率。

（3）PCNN-HATT：Han[36]发表在 2018 年 EMNLP 会议上的论文。该模型充分利用关系类别的层次性结构，捕获关系内部由粗粒度到细粒度的语义信息。把关系类别划分成三个层次，分别融合选择性注意力机制。

（4）PCNN-ATTRA-BAGATT：Ye[37]发表在 2019 年 NAACL 会议上的论文。该模型同时考虑了包内部和包之间的选择性注意力机制，有效地减弱句子级别和包级别的噪声。

（5）DISTRE：Alt[8]发表在 2019 年 ACL 会议上的论文。该模型基于 Transformer 预训练语言模型进行包内句子的编码表示，结合选择性注意力机制实现降噪，是本文的基础模型。

4.4.2 结果与分析

（1）与其他模型对比

图 4-1 显示了几个对比模型的实验结果。我们的两个模型 DISTRA-EA 和 DISTRE-SA 都达到了最新的 SOTA 的效果，AUC 值分别为 0.430 和 0.427。在较高召回率的情况下，两个模型的性能明显优于 PCNN-ATT、PCNN-HATT 和 PCNN-ATTRA-BAGATT；而在观测按照预测概率由高到低排序后的样例序列的前一部分时，它们的精准率略低。PCNN-ATT 模型的结果表明，它的性能仅在曲线的开端表现较好，但是精准率下降得很快，并且只能达到 0.373 的 AUC 值。类似地，PCNN-HATT 模型在最初表现更好，在召回率大约为 0.2 的位置后，精准率有所下降。包间包内注意力模型 PCNN-ATTRA-BAGATT 曲线整体相对稳定，在召回率达到 0.25 之前，它的精准率表现最优，远远高于其他的模型。这说明加入包之间的选择性注

意力机制后，可以明显利用其它包的信息降低噪声，提升模型的精准率。

我们补充了两个基于 CNN 编码的模型 CNN-ATT 和 CNN-HATT。从图中可以观察到，增加两个实体结构信息的分段卷积 PCNN 和简单的 CNN 在编码方式上，二者的抽取效果相近。与上述编码方式相比，使用 Transformer 预训练语言模型对输入序列进行编码的方式，句子的信息具备更宽泛和深度的语法语义信息，所以整体性能保持平衡和稳定。

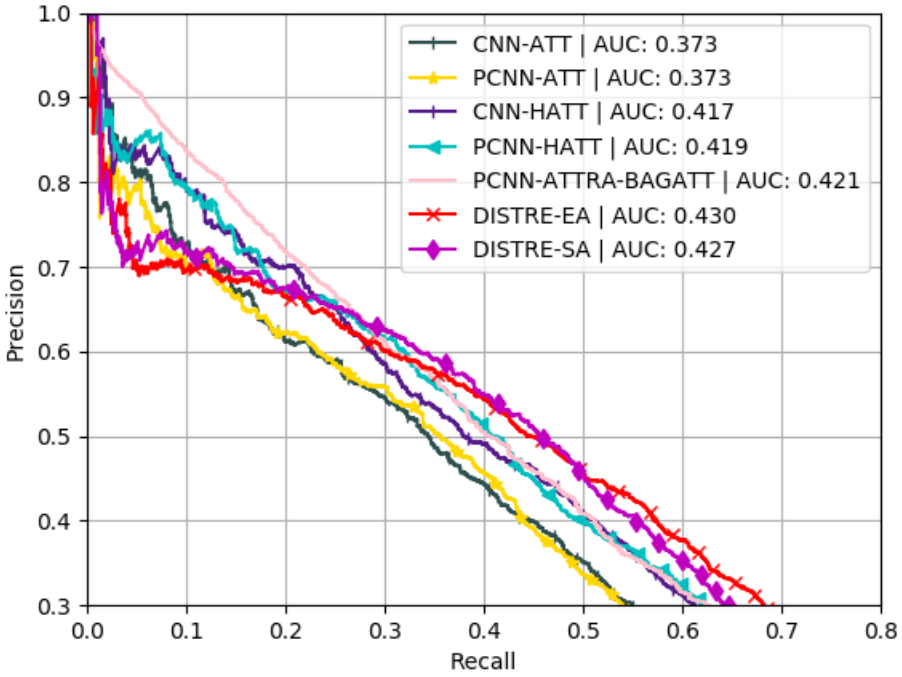


图 4-1 NYT 数据集上不同模型的结果

Figure 4-1 Results of different models on NYT dataset

表 4-6 显示了沿 P-R 曲线在不同 K 取值下的 Precision@K 指标。可以看到，前几个模型随着 K 值增大，精准率下滑非常明显，与我们的模型的差距逐渐增大。特别是在 P@1000 和 P@2000 的位置上，DISTRE-SA 都达到了 61.6%，比其他模型中最好的 PCNN-HATTRA-BAGATT 模型高出了 9.5%和 27.2%，整个模型的精准率维持较高的标准。尽管 DISTRE-EA 从 P@K 看整体稍逊色于 DISTRE-SA，但是还是优于大部分其他模型，尤其是 P@1000 和 P@2000 也都是优于其他方法的结果。除了我们的模型，其他模型有几点值得注意：PCNN-HATTRA-BAGATT 在前 300 的精确率展示了它的优势，P@100 的分数更是高达 90.3%，恰好反映了其 P-R 曲线的总体趋势，这得益于使用了包之间的注意力机制，引入了更多判断噪声句子的依据。RNN-adv 在输入序列上增加对抗性噪声，虽然能提升效果，但受限于 RNN 的特征提取能力而不能获得更好的效果。

另外，在 AUC 的整体指标衡量下，我们的两个模型高于其他模型中最好的

PCNN-HATTRA-BAGATT 模型 0.6~0.9 个百分点。DISTRE-EA 在 P@K 指标虽然不如 DISTRE-SA，但其 AUC 值是所有模型中最高的，说明其整体的效果要优于其他的模型，具有更好的稳定性。

表 4-6 对比模型的 AUC 和 P@K

Table4-6 AUC and P@K of the contrast models

System	AUC	P@100	P@200	P@300	P@500	P@1000	P@2000
CNN-ATT	.373	72.0	69.0	62.0	51.2	36.9	22.9
PCNN-ATT	.373	75.0	71.5	65.0	55.0	37.3	23.9
RNN-adv	.382	-	-	-	-	-	-
CNN-HATT	.417	83.0	72.0	69.0	57.6	40.2	25.4
PCNN-HATT	.419	82.0	73.5	68.0	59.0	41.2	25.4
PCNN-HATTRA-BAGATT	.421	**90.3**	**82.1**	**76.4**	68.4	52.1	34.4
DISTRE-EA	**.430**	76.0	70.0	69.7	67.6	**59.7**	**47.2**
DISTRE-SA	**.427**	70.0	73.5	71.0	**68.6**	**61.6**	**61.6**

（2）与基础模型对比

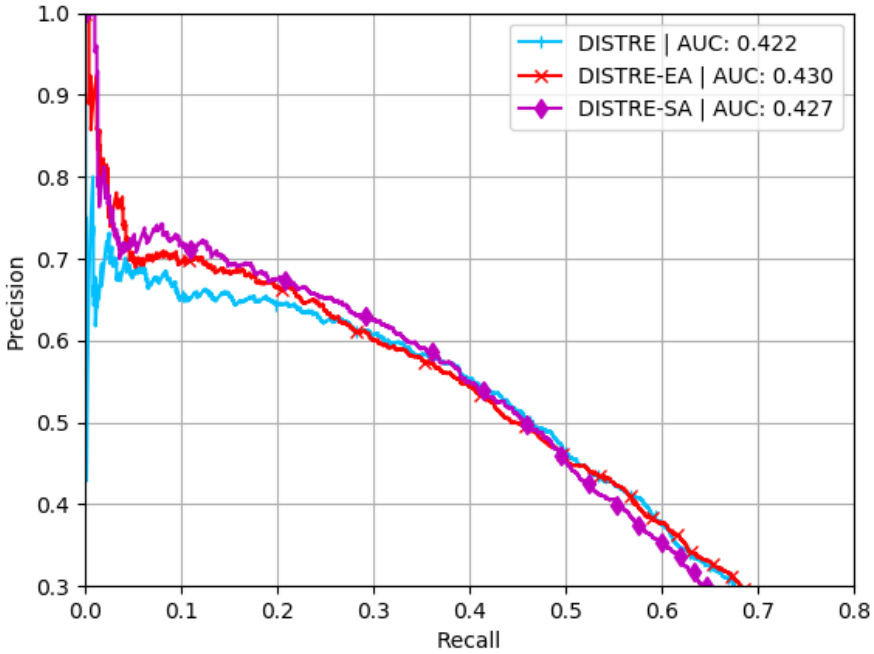


图 4-2 NYT 数据集上与基础模型的对比结果

Figure 4-2 Comparison result with the basic model on the NYT dataset

如图 4-2 所示是我们的两个模型与基础模型 DISTRE 的对比结果。DISTRE-EA 和 DISTRE-SA 模型的 P-R 曲线在大部分都要高于 DISTRE 模型的 P-R 曲线。在后

半段时，我们的模型与基础模型类似，召回率较高。在兼顾召回率的同时，可以发现对于前半段的精准率，我们的两种模型对 DISTRE 做出了明显的提升。召回率为 0.3 之前，两个模型的曲线明显高于基础模型。前半段的提升表明，对包的嵌入表示增加实体注意力机制和增加句子间的自注意力机制，可以为分类器提供更丰富的实体相关信息和更少噪声的句子嵌入表示，因此改善了基础模型的精准率。具体来说，两种模型都保留了 DISTRE 中 Transformer 预训练语言模型抽取特征的优势，DISTRE-EA 以实体注意力的形式增强了包嵌入表示中实体的相关信息，DISTRE-SA 以句子级自注意力机制，建立与其他句子之间的联系，减少了了包嵌入表示中的噪声。从实验结果看，这两种策略对关系抽取都有一定程度的提升效果。

表 4-7 中，基础模型 DISTRE 的 AUC 值为 0.422，DISTRE-EA 和 DISTRE-SA 分别提高了 8 个和 5 个千分点。在所有的 P@K 数值中，我们的模型占据了最好的结果。在 P@100 的位置上，DISTRE-EA 达到了 76%，比基础模型提高了 8 个百分点。前 500 个示例中，两个模型的结果都有很大程度地提升。尤其是 DISTRE-SA，在每个位置上都优于基础模型，并且在后续示例中仍然保持非常高的精准率。我们的两个模型中，在精准率方面，DISTRE-SA 的表现优于 DISTRE-EA，其抽取的关系更为准确；在 AUC 的评价指标上，DISTRE-EA 比 DISTRE-SA 高出 3 个千分点，其模型整体表现更稳定，能够具备高于基础模型 DISTRE 精准率的同时也具有较好的召回率。

综上，无论是实体注意力机制还是句子级别的自注意力机制，都提升了关系抽取的效果，改善了基础模型的性能。

表 4-7 对比基础模型的 AUC 和 P@K

Table4-7 Compare AUC and P@K of the basic model

System	AUC	P@100	P@200	P@300	P@500	P@1000	P@2000
DISTRE	.422	68.0	67.0	65.3	65.0	60.2	47.9
DISTRE-EA	**.430**	**76.0**	70.0	69.7	67.6	59.7	47.2
DISTRE-SA	.427	70.0	**73.5**	**71.0**	**68.6**	**61.6**	**61.6**

（3）超参数分析

在 DISTRE-EA 模型上，本文借鉴 DISTRE 基础模型的一部分超参，例如 TF 块的数量和多头注意力的头数均设为 12。由于预训练的限制，TF 块的数量 L 的上限为 12。多头注意力的头数可以在 12 附近范围内搜索，但是数值不应过大，否则造成待训练参数矩阵过大的参数量。我们针对一组预设的超参数 $L = 12, h = 12, batch\ size = 32, \eta = 0.2, \mathrm{q} = 20, \mathrm{dropout} = 0.2$。在控制其他超参数不变的情况

下，对特定的某个超参数进行调整试验，记录每次实验结果最终选取最优的数值。

图 4-3 展示四种超参数在 DISTRE-EA 模型上的实验结果。横坐标表示每个超参数的取值点，纵坐标表示对应的 AUC 值。AUC 是 P-R 曲线下的面积，具有全局代表性，所以作为衡量模型性能的评估标准。图 4-5（a）中，TF 块的数量L随着取值的增大，整体结果呈现上升的趋势，最终在上限处达到最优。图 4-5（b）中，多头注意力的头数h在小范围内波动，幅度起伏偏小，在$h = 12$左右呈对称结构。这L和h二者的最优点也恰好验证了 DISTRE 模型的取值效果。 图 4-5（c）中，不难发现随着 *batch size* 的增加，AUC 值也随之迅速提升，在达到 32 后突然下降。*batch size* 的影响起伏较大，原因是模型本身的待训练的参数数量较多，若批处理的训练数据少，模型易受数据偏差的影响，导致梯度反传的方向过于绝对，降低了鲁棒性；若训练数据过多，噪声信息混杂交错，模型无法做出针对性的修正。图 4-5（d）中，我们设置的实体嵌入表示权重η在 0.2 处表现最好。若为 0.1 的数值，则实体注意力占的比重太小，叠加在包嵌入表示上不仅没有提升作用，反而掺杂了扰乱信息。若占比大的话，模型偏向实体的表示而忽略了句子间的选择性注意力的降噪作用。总体而言，h的影响度最小，结果仅略有差异。L，$batch\ size$，η的影响幅度较大，存在 3%~5%的明显差距。

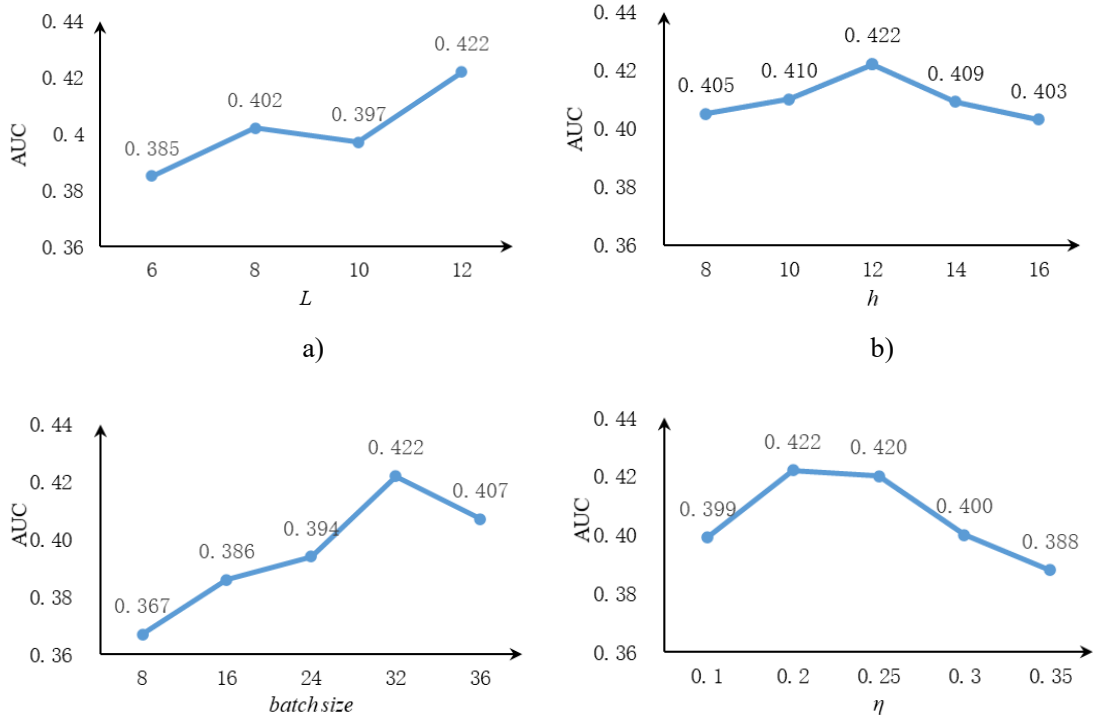


图 4-3 L、h、*batch size*、 η 四种超参数在 DISTRE-EA 模型上的影响

Figure 4-3 The influence of four hyperparameters L, h, *batch size* and η on the DISTRE-EA model

上述四种超参数可以固定其中一个，保证在该超参数设置下的模型获得最优的

效果。但是在实验过程中，我们发现注意力矩阵W_1第二个维度的q和分类器的*dropout*之间存在联动性，所以将这两者构成一个超参数组合，作为整体进行调试。

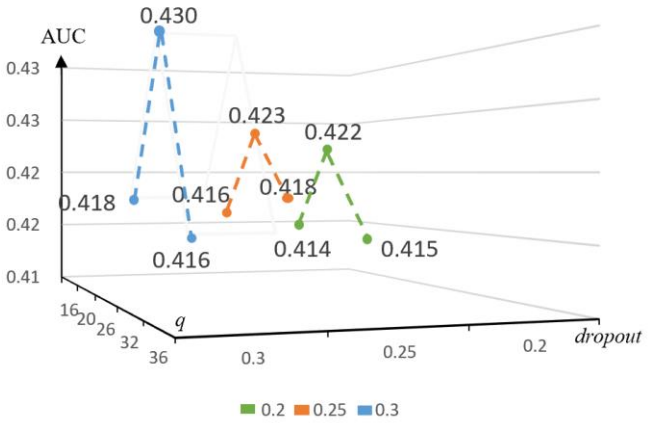


图 4-4 q、*dropout*两种超参数在 DISTRE-EA 模型上的影响

Figure 4-4 The influence of two hyperparameters q and *dropout* on the DISTRE-EA model

图 4-4 和表 4-8 是超参数q和 *dropout* 之间的组合实验结果，我们设置了九类组合，两个超参数的范围是 $dropout = \{0.2,0.25.0.3\}$，$q = \{16,20,26,32,36\}$。通过图表我们可以发现，AUC 值在二者对角线的位置上呈上升的趋势，而对角线的两侧效果略差。这体现了，q和 *dropout* 同时增加时能够提升效果，而只提升其一则效果不佳。在 dropout 为 0.2 时，三种组合的效果偏低；为 0.25 时，效果略微提升；为 0.3 时，三种组合的水平提高了一个层次。在$dropout = 0.3, q = 32$时的结果最优。由此可见，增加注意力参数矩阵W_1的第二个维度q的同时适当增大 dropout 的数值，在自注意机制复杂度高的情况下可有效防止过拟合。

表 4-8 q和 *dropout* 超参数组合的影响

Table4-8 Influence of q and *dropout* hyperparameter combination

AUC		dropout		
		0.2	0.25	0.3
q	16	0.414	-	-
	20	**0.422**	0.416	-
	26	0.415	**0.423**	0.418
	32	-	0.418	**0.430**
	36	-	-	0.416

在 DISTRE-SA 模型上，我们主要针对三类超参数进行调整，分别是批大小 $batch\ size$、句子级别自注意力的头数 *self_head* 和自注意力的 *dropout*。我们预设一组超参数 $L = 12, h = 12, batch\ size = 16, dropout = 0.2, self\_head =$

$6, self\_dropout = 0.8$。在控制变量的条件下，进行实验并记录结果。

图 4-5 是三类超参数的调试结果。横坐标表示超参数的取值点，纵坐标表示对应的 AUC 值。图 4-5（a）中，在 *batch size* 为 8 和 24 时，模型错过了局部最优解，在 loss 经过局部最小值后突然迅速增大，导致模型无法收敛。*batch size* 的最优值在 16 的位置上，与 DISTRE-EA 模型有所不同。原因在于 DISTRE-SA 是对 Transformer 编码得到的句子表示做变换，为了避免在做句子变换时造成信息损失，失去预训练模型的优势，所以同一批的训练数据不宜过多。另一方面，由于 DISTRE-SA 的超参数量庞大，模型复杂度增加，所以批处理数据也不宜过少，否则训练速度慢且难以具有较好的泛化性。图 4-5（b）中，AUC 值在达到 0.427 后逐渐下降趋于平稳，*self_head* 为 6 时获得最大的竞争力。图 4-5（c）中，句子级别自注意力的 *self_dropout* 得分呈现先下降再上升的趋势。*self_dropout* 在 0.8 处达到最优，这种较大的 dropout 取值说明在加入句子级自注意力后，由于 Transformer 语言模型中的和句子级自注意力中的多头注意力的参数过于庞大，在第二个自注意力机制上加大 dropout，可以显著提高模型的泛化能力避免过拟合。

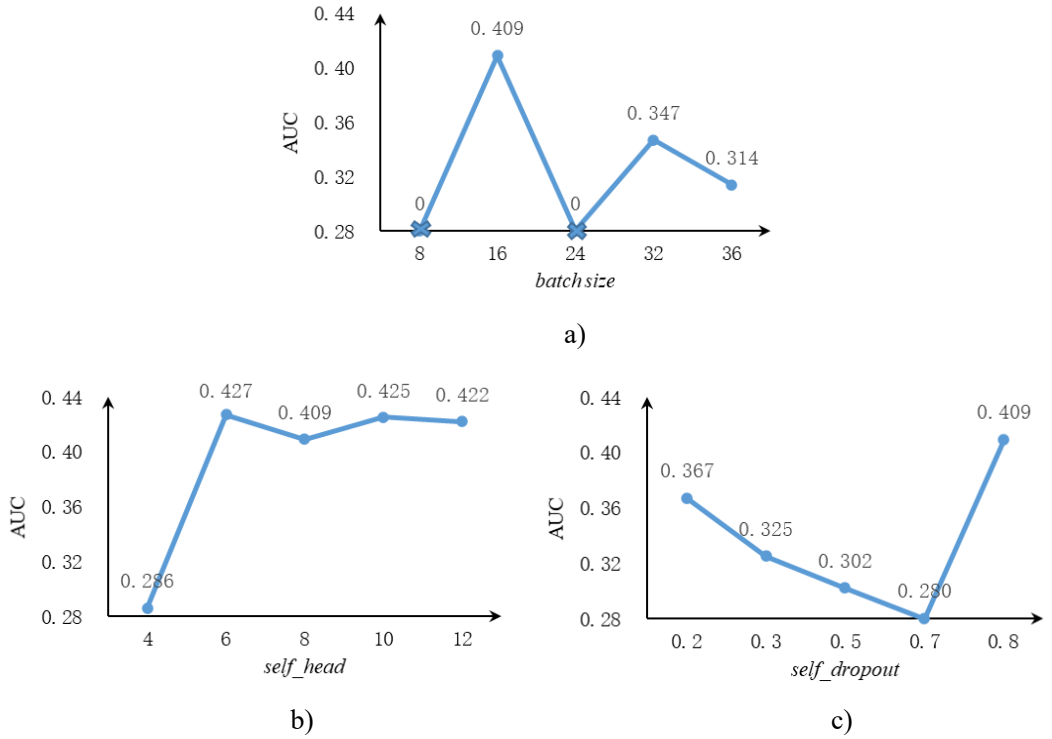


图 4-5 *batch size*、*self_head*、*self_dropout* 三种超参数在 DISTRE-SA 模型上的影响

Figure 4-5 The effect of three hyperparameters *batch size*, *self_head* and *self_dropout* on the DISTRE-SA

4.5 本章小结

在本章中，本文介绍了实验使用的数据集、评估指标、其他实验细节并对最后的实验结果进行了展示和分析。数据集的介绍涉及使用的公开数据集的名称以及样本统计信息。评估指标主要给出了关系抽取任务中常用的能反映抽取效果的三种指标及其定义。实验细节列举了预训练模型的版本、优化模型的方法，以及通过大量实验最终确定的最优超参数设置，便于其他学者在将来复现本文的工作。最后，本文介绍了用于对比的模型并着重展示和分析了实验结果。实验结果体现了我们提出的两个模型 DISTRE-EA 和 DISTRE-SA 均能有效提升远程监督关系抽取任务的效果。

5 总结与展望

关系抽取是信息抽取的关键技术，同时也是自动构建知识图谱的重要手段。目前，关系抽取技术蓬勃发展，具有良好的价值和应用前景。它的重要性和实用性受到了学术界和工业界的广泛关注。由于标注数据人工成本昂贵且数量有限，因此基于远程监督的关系抽取技术成为了研究的热点。远程监督依据知识库回标语料的方法省去了大量的前期工作，带来便捷性的同时也无可避免地产生了回标噪声问题。解决噪声的方法主要集中在两个方面，一类是抑制噪声的方法，通过融合注意力机制降低嘈杂句子对包表示的消极影响，比如选择性注意力、层次型注意力和自注意力。第二类是过滤噪声的方法，研究思路是重新构建一份干净的数据集，再结合分类器实现关系预测。本文提出的两个模型通过实体注意力机制和句子级自注意力机制强化包的嵌入表示，实现模型的降噪过程。在本章中，我们对所做的研究内容进行总结，并对关系抽取的研究方向做出展望。

5.1 研究内容总结

本文首先回顾和分类总结了关系抽取技术的研究内容，深度调研远程监督的相关技术和成果。为了解决远程监督存在的大量噪声问题，本研究提出了基于 Transformer 预训练语言模型和实体注意力机制的关系抽取模型 DISTRE-EA，基于 Transformer 预训练语言模型和句子级自注意力机制的关系抽取模型 DISTRE-SA。上述两个模型利用注意力机制从不同的角度提升了包嵌入表示的表征能力，为分类器提供了更加精准和丰富的信息。通过与多种主流方法的对比实验，验证模型的有效性。本研究的主要贡献具体如下：

（1）在 Transformer 预训练语言模型基础上实现实体注意力机制。一方面，语言模型本身涵盖海量的自然语言语法语义信息，而我们模型的设计保证了语言模型发挥其特有的优势，也可以灵活地结合其他将来出现的更好的预训练语言模型。另一方面，我们发现在以往的研究中，实体表现的部分欠缺欠佳。关系抽取的任务意在识别两个实体中表达的准确关系，因此实体本身的指导作用不容忽视。通过计算实体与包内每个句子之间的相关度，优选包内预训练模型得到的句子嵌入表示，降低噪声句子影响的同时有助于优化包的嵌入表示。实验表明，该模型的召回率和 AUC 都取得了不错的表现。与其他模型对比可以看出，我们的模型在各方面指标上有较好的提升，证明了模型的有效性。

（2）在 Transformer 预训练语言模型上实现句子级自注意力机制。自注意力机

制是 Transformer 模型的重要组成部件，它可以捕获输入序列内部之间的相互依赖信息。对于一个包而言，包内的句子嵌入表示是通过预训练语言模型得到的，彼此之间缺乏紧密的关联性。但实际上包中的句子存在噪声句子和非噪声句子的差异和关联。为了打破包内句子嵌入表示的互相独立，本文在句子级别上使用自注意力机制。利用查询向量和键向量计算注意力权重后，对句子本身的值向量加权求和，即可得到融入了其他句子信息的新的句子嵌入表示。进而，结合包内部的选择性注意力机制，可以获得更优的包嵌入表示，降低噪声句子的消极影响。实验表明，该模型也取得了显著的效果，证明了该模型的有效性。

5.2 未来工作展望

近年来围绕关系抽取技术的发展涌现出了丰富的成果，层出不穷的新技术应用其中。研究者们不断为该领域的推进和发展提供新的解决思路，展现出了巨大的活力。在大规模机器学习技术的背景下，关系抽取依然面临巨大的挑战。本研究针对远程监督关系抽取提出了两点研究内容，尽管取得不错的抽取效果，但是纵览全局，还有很多需要完善和提升的空间。结合当前的发展形势，可以开展以下几个工作：

（1）假正例和假负例。

由于自动标注语料产生的错误标记，假正例是一个固然存在的问题。除此之外，“At-Least-One”前提带来的失效性假正例也是一个难点，因为难免存在一个包中所有的示例均不能表达知识库中关系的情况。并且由于知识库本身的不完备性，导致标记数据产生假负例。假正例和假负例是远程监督的难题，也正是这两种数据的存在使得标记语料的噪声太强大，有待更多人进行深入研究。

（2）评估方法。

对于远程监督关系抽取模型的评估过程，仍存在许多不尽人意的地方。测试集同样存在错误标注的问题，自动评估方法显得不够公正和客观，无法准确评估模型的整体性能。而人工评估手段耗时费力，效率低下，代价成本偏高。如何合理地评估模型，需要进一步探索更好的评估方法和评估指标。

（3）关系抽取的局限性。

目前关系抽取这个经典任务在各个会议上仍是以句子级别的识别出现，极少将其扩展到篇章级别上。篇章级别关系抽取缺少相关的数据集，理论研究深入较少。但是在实际环境当中，篇章级别的关系抽取具有很高的使用价值。大多数的非结构化文本数据以篇章的形式出现，如果能在此基础上利用抽取技术将实体与关系组合识别出来，对下游任务更友好。

（4）面向开放领域的关系抽取。

目前绝大部分的关系抽取研究集中在限定领域中，根据预先定义好的关系类别进行关系抽取任务。这为关系抽取的发展带来了较强的局限性，往往存在不足之处，一是无法抽取目标关系类别之外的其他实体关系知识；二是如果更新语料库，现有模型的性能将会大幅度下降。虽然已有一些学者在进行开放域的关系抽取，但是受语义歧义的影响（如“苹果”指水果或公司），抽取质量不佳。

（5）具有时空特性的多元关系抽取。

现实中的关系具有时空特性和多元性。目前的研究集中在二元关系抽取上，即抽取目标为三元组表示（实体 1，关系，实体 2），在限定的框架下很难表达现实中的时空特性和多元性。 例如“奥巴马是美国总统”，奥巴马曾经是美国总统，这是时间特性；其次，美国总统在历史上有多个，体现了多元关系。具有时空特性的多元关系能建模更丰富的关系知识，是未来研究的重点方向。

参考文献

[1] 徐增林,盛泳潘,贺丽荣,王雅芳.知识图谱技术综述[J].电子科技大学学报,2016,45(04):589-606.

[2] 刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述[J].计算机研究与发展,2016,53(03):582-600.

[3] 漆桂林,高桓,吴天星.知识图谱研究进展[J].情报工程,2017,3(01):4-25.

[4] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 1003-1011.

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

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

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

[8] Alt C, Hübner M, Hennig L. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction[J]. arXiv preprint arXiv:1906.08646, 2019.

[9] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

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

[12] Hendrickx I, Kim S N, Kozareva Z, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals[C]//Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, 2010: 33-38.

[13] Doddington G R, Mitchell A, Przybocki M A, et al. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation[C]//Lrec. 2004, 2: 1.

[14] Liu C Y, Sun W B, Chao W H, et al. Convolution neural network for relation extraction[C]//International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, 2013: 231-242.

[15] Nguyen T H, Grishman R. Relation extraction: Perspective from convolutional neural networks[C]//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015: 39-48.

[16] Santos C N, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural

networks[J]. arXiv preprint arXiv:1504.06580, 2015.

[17] Zhang D, Wang D. Relation classification via recurrent neural network[J]. arXiv preprint arXiv:1508.01006, 2015.

[18] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016: 207-212.

[19] Wang L, Cao Z, De Melo G, et al. Relation classification via multi-level attention cnns[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 1298-1307.

[20] Zhu J, Qiao J, Dai X, et al. Relation classification via target-concentrated attention cnns[C]//International Conference on Neural Information Processing. Springer, Cham, 2017: 137-146.

[21] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora[C]//Proceedings of the 42nd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2004: 415.

[22] Bollegala D T, Matsuo Y, Ishizuka M. Measuring the similarity between implicit semantic relations from the web[C]//Proceedings of the 18th international conference on World wide web. 2009: 651-660.

[23] Bollegala D T, Matsuo Y, Ishizuka M. Relational duality: Unsupervised extraction of semantic relations between entities on the web[C]//Proceedings of the 19th international conference on World wide web. 2010: 151-160.

[24] Yao L, Riedel S, McCallum A. Unsupervised relation discovery with sense disambiguation[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012: 712-720.

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

[26] Agichtein E, Gravano L. Snowball: Extracting relations from large plain-text collections[C]//Proceedings of the fifth ACM conference on Digital libraries. 2000: 85-94.

[27] Liu X, Yu N. Multi-type web relation extraction based on bootstrapping[C]//2010 WASE International Conference on Information Engineering. IEEE, 2010, 2: 24-27.

[28] Zeng W, Lin Y, Liu Z, et al. Incorporating relation paths in neural relation extraction[J]. arXiv preprint arXiv:1609.07479, 2016.

[29] 白龙, 靳小龙, 席鹏弼, 等. 基于远程监督的关系抽取研究综述[J]. 中文信息学报, 2019, 33(10): 10-17.

[30] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[J]. 2014.

[31] Huang Y Y, Wang W Y. Deep residual learning for weakly-supervised relation extraction[J]. arXiv preprint arXiv:1707.08866, 2017.

[32] Ji G, Liu K, He S, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[33] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph

completion[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015.
[34] 鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述[J]. 软件学报, 2019 (2019 年 06): 1793-1818.
[35] Jat S, Khandelwal S, Talukdar P. Improving distantly supervised relation extraction using word and entity based attention[J]. arXiv preprint arXiv:1804.06987, 2018.
[36] Han X, Yu P, Liu Z, et al. Hierarchical relation extraction with coarse-to-fine grained attention[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2236-2245.
[37] Ye Z X, Ling Z H. Distant supervision relation extraction with intra-bag and inter-bag attentions[J]. arXiv preprint arXiv:1904.00143, 2019.
[38] Liu T, Wang K, Chang B, et al. A soft-label method for noise-tolerant distantly supervised relation extraction[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1790-1795.
[39] Jia W, Dai D, Xiao X, et al. ARNOR: Attention Regularization based Noise Reduction for Distant Supervision Relation Classification[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1399-1408.
[40] Qin P, Xu W, Wang W Y. Dsgan: Generative adversarial training for distant supervision relation extraction[J]. arXiv preprint arXiv:1805.09929, 2018.
[41] Feng J, Huang M, Zhao L, et al. Reinforcement learning for relation classification from noisy data[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[42] Qin P, Xu W, Wang W Y. Robust distant supervision relation extraction via deep reinforcement learning[J]. arXiv preprint arXiv:1805.09927, 2018.
[43] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf, 2018.
[44] Vashishth S, Joshi R, Prayaga S S, et al. Reside: Improving distantly-supervised neural relation extraction using side information[J]. arXiv preprint arXiv:1812.04361, 2018.
[45] 周志华. 机器学习[M]. Qing hua da xue chu ban she, 2016.
[46] Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books[C]//Proceedings of the IEEE international conference on computer vision. 2015: 19-27.
[47] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[J]. arXiv preprint arXiv:1508.07909, 2015.
[48] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
[49] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.
[50] Wu Y, Bamman D, Russell S. Adversarial training for relation extraction[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1778-1783.

作者简历及攻读硕士学位期间取得的研究成果

一、作者简历

赵静菲，女，1995 年 3 月 11 日生，广西南宁

2014.09-2018.06 北京交通大学 计算机与信息技术学院 计算机科学与技术 学士

2018.09-2020.06 北京交通大学 计算机与信息技术学院 计算机技术 硕士

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：  签字日期： 2020 年 6 月 13 日

学位论文数据集

表 1.1： 数据集页

<table>
<tr><td>关键词*</td><td>密级*</td><td>中图分类号</td><td>UDC</td><td>论文资助</td></tr>
<tr><td>关系抽取；远程监督；语言模型；注意力机制；信息抽取</td><td>公开</td><td></td><td></td><td></td></tr>
<tr><td colspan="2">学位授予单位名称*</td><td>学位授予单位代码*</td><td>学位类别*</td><td>学位级别*</td></tr>
<tr><td colspan="2">北京交通大学</td><td>10004</td><td>工程硕士专业</td><td>硕士</td></tr>
<tr><td colspan="2">论文题名*</td><td colspan="2">并列题名</td><td>论文语种*</td></tr>
<tr><td colspan="2">注意力增强包表示的远程监督关系抽取方法研究</td><td colspan="2"></td><td>中文</td></tr>
<tr><td>作者姓名*</td><td colspan="2">赵静菲</td><td>学号*</td><td>18125297</td></tr>
<tr><td colspan="2">培养单位名称*</td><td>培养单位代码*</td><td>培养单位地址</td><td>邮编</td></tr>
<tr><td colspan="2">北京交通大学</td><td>10004</td><td>北京市海淀区西直门外上园村 3 号</td><td>100044</td></tr>
<tr><td colspan="2">工程领域*</td><td>研究方向*</td><td>学制*</td><td>学位授予年*</td></tr>
<tr><td colspan="2">计算机技术</td><td>机器学习</td><td>2 年</td><td>2020 年</td></tr>
<tr><td>论文提交日期*</td><td colspan="4">2020 年 6 月</td></tr>
<tr><td>导师姓名*</td><td colspan="2">尹传环</td><td>职称*</td><td>副教授</td></tr>
<tr><td>评阅人</td><td colspan="2">答辩委员会主席*</td><td colspan="2">答辩委员会成员</td></tr>
<tr><td></td><td colspan="2">常晓林</td><td colspan="2">余贶琭、张硕</td></tr>
<tr><td colspan="5">电子版论文提交格式 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ）
推荐格式：application/msword；application/pdf</td></tr>
<tr><td colspan="2">电子版论文出版（发布）者</td><td colspan="2">电子版论文出版（发布）地</td><td>权限声明</td></tr>
<tr><td colspan="2"></td><td colspan="2"></td><td></td></tr>
<tr><td>论文总页数*</td><td colspan="4">47</td></tr>
<tr><td colspan="5">共 33 项，其中带*为必填数据，为 21 项。</td></tr>
</table>