远程监督关系抽取综述

杨穗珠 刘艳霞 张凯文 洪吟 黄翰

(华南理工大学 软件工程, 广州 中国 510641)

摘 要 远程监督可以为关系抽取任务自动构建数据集，缓解了人工构建数据集的压力和成本，为自动关系抽取的实现奠定基础，然而使用远程监督方法构建的数据集存在错误标注以及长尾问题，严重影响关系抽取性能。目前，远程监督关系抽取任务的主要研究方向为关系模型的降噪手段以及对长尾关系的处理方法。近年来，随着深度学习技术的发展，这两个领域的研究工作也迎来了新一轮的机遇与挑战。本文对近几年远程监督关系抽取的研究进展进行综述，针对远程监督关系抽取任务定义常用工作流，将已有的研究成果进行分类和梳理，分析比较主要方法，整理其中的关键问题，介绍已有的解决方案和相关数据集，总结远程监督关系抽取任务所用评测指标与评估方式，展望未来研究趋势。

关键词 关系抽取；信息抽取；远程监督；降噪；长尾现象；错误标注

Survey on Distantly-Supervised Relation Extraction

YANG Sui-Zhu LIU Yan-Xia ZHANG Kai-Wen HONG Yin HUANG Han

(School of Software Engineering, South China University of Technology, Guangzhou 510641, China)

Abstract Relation extraction is a fundamental task in natural language processing and one of the essential parts of information extraction, whose dataset requires high cost due to manual labelling. Fortunately, distant supervision was proposed to alleviate the pressure and cost of manually annotated corpus, which can automatically build datasets for relation extraction task. Owing to its value in automatic relation extraction, it has been widely concerned by academia and business in recent years. However, the datasets constructed by distant supervision are not exactly equivalent to those generated manually. On the contrary, they suffer from the problem of wrong labelling and long tail distribution, resulting in their low quality, and thus hindering the improvement of relation extraction based on these datasets. Therefore, in order to reduce the impact, most of the existing work about distantly-supervised relation extraction (DSRE) focused on how to deal with the noise generated by wrong labelling problem and the long tail distribution. In recent years, deep learning technologies have developed rapidly such as deep neural network, attention mechanism, deep reinforcement learning and so on. Compared with traditional machine learning methods, e.g. feature-based methods, the application of deep learning methods has obvious advantages in relation extraction, as well as DSRE task. That is why DSRE is faced with a new round of opportunities and challenges. What's more, as researches continue, a common workflow of this task was generated step by step. This paper summarizes the existing work in the field of DSRE, and pays more attention to the methods based on deep learning. This paper starts with an introduction of distant supervision as well as its vanilla assumption, analyzes the major shortcoming and reviews the methods based on traditional machine learning such as topic models and pattern correlation and so on. Then this paper introduces the general

本课题得到国家自然科学基金 (No. 61876208)、广东省重点研发项目 (No. 2018B010109003)、广州市科技计划(No.201802010007, No.201804010276)资助.杨穗珠，女，1995年生，硕士研究生，主要研究领域为知识图谱、远程监督. Email: 414608192@qq.com . 刘艳霞(**通信作者**)，女，1979年生，博士，副教授，主要研究领域为知识图谱、神经网络.E-mail: cslyx@scut.edu.cn. 张凯文，男，1997年生，硕士研究生，主要研究领域为知识图谱. E-mail: nczkevin97@gmail.com . 洪吟，男，1995年生，硕士研究生，主要研究领域为联合实体关系抽取. E-mail: michael-hong@foxmai.com . 黄翰，男，1980年生，博士，教授，计算机学会(CCF)高级会员（13542S）,主要研究领域为智能算法、演化计算.E-mail: hhan@scut.edu.cn .

workflow with four modules, including sample collection, external information, encoder and classifier. According to their target problem, the existing work is divided into two categories, the noise reduction methods of DSRE and the solutions of the long tail distribution. For each category, in the light of different modules of the common workflow, the existing work is summarized from four aspects, namely sample noise reduction, external information fusion, encoder optimization and classifier optimization. Meanwhile, this paper analyzes different improvement methods of the same module, and compares their weakness and strength. It should be noted that these four aspects are not mutually exclusive, meaning that there can be two or more modules improved in one method at the same time. What's more, we introduce the datasets in common use for this task in detail, as well as their related corpus and knowledge graphs. Moreover, this paper introduces the metrics and evaluation methods used in the DSRE evaluation. Last but not least, this paper ends up with forecasting the future development trend. In order to bring this task into a new frontier, we hope that DSRE can be integrated with some popular and reasonable technologies such as joint extraction, few-shot learning, hybrid supervision and so on.

Key words relation extraction; information extraction; distant supervision; noise reduction; long tail; wrong labelling

1 引言

关系抽取（Relation Extraction，RE）的目的是对句子中实体与实体之间的关系进行识别，抽取句子中的三元组信息，即（实体 1，实体 2，关系）三元组，得到的三元组信息可以提供给知识图谱的构建、问答、机器阅读等下游自然语言处理（Natural Language Processing，NLP）任务，是 NLP 领域的一个基础任务。传统的监督学习的关系抽取方法虽然准确率较高，模型结果更为可靠，但需要人工标注数据集，构造这样的数据集需耗费大量的人力、金钱和时间。近年来，为了实现自动化关系抽取，学者们提出了远程监督（Distant Supervision）的方法。Craven 等人[1]在 1999 年尝试从现有的信息源（知识图谱、数据库或者简单的表格信息等）中提取三元组信息，再对语料集进行标注，生成提供关系抽取的“弱标记”的训练数据。在 2007 年，Wu 等人[2]也认为监督学习的方法需要过多的人力干预，提出从维基百科页面的信息框中抽取出结构化信息的方法，这个方法使得从已有非结构化数据中获取结构化信息成为可能。在 2009 年，Mintz 等人[3]总结前人的工作，参考 Wu 等人[2]的方法，提出了使用远程监督的方法进行关系抽取任务的数据集构造，这个方法将 Freebase 作为辅助的结构化信息，对维基百科的文章进行标注，生成关系抽取的数据集，大大缓解了人工构造关系抽取训练集的压力。

远程监督结合了半监督学习和无监督学习的优点，利用已有的结构化数据来对数据进行自动化标注生成训练数据，这个思想不仅应用到关系抽取任务中，还用到了 NLP 的多个领域，Go 等人[4]将其用于情感分析任务，Plank 等人[5]将其应用于词性标记任务中，而 Qin 等人[6]将其应用于对话系统的样例生成中，Lee 等人[7]将其应用到命名实体识别任务中，远程监督方法减轻了人工构造数据集的压力，降低了学术研究的成本。然而 Mintz 等人[3]提出的远程监督方法并不是完美的，在关系抽取任务上由于提出时伴随了强约束性的假设，生成的数据集存在严重的错误标注以及数据长尾问题。

近年来，研究学者针对错误标注及数据长尾问题，在远程监督噪声数据过滤以及解决长尾关系等领域都提出了切实可行的方案。深度学习技术的应用使得远程监督关系抽取任务的性能获得了突破性进展，新的研究成果和思路不断出现，已有的综述未能全面而深入地进行总结。Roth 等人[8]针对远程监督中的降噪方法进行综述，将已有的降噪方法分为 At-least-one 假设、主题模型以及模式相关性三类，并对每个类的方法都进行了详细的说明，但是对目前常用的神经网络模型没有进行深入的描述和分析。Smirnova 等人[9]针对远程监督在关系抽取的应用进行了综述，把已有的研究成果分为降噪方法、基于嵌入的方法以及利用辅助信息的方法，其中降噪方法参考了 Roth 等人[8]对降噪方法的分类，他们详述了在各个分类下，远程监督在关系抽取任务的应用情况，然而其并没有对基于深度学习方法的远程监督关系抽取方法进行详细的说明，尚未覆盖目前最新的研究进展。

本文则以基于深度学习方法的远程监督关系抽取任务为重点研究对象，针对目前常用的提升远

程监督关系抽取任务性能的解决方案进行分类说明和总结，并详尽梳理其中的关键性问题和解决方法，整理评价指标，廓清本领域的发展情况与趋势，展望发展方向，为未来的研究工作奠定基础。

1.1 方法描述

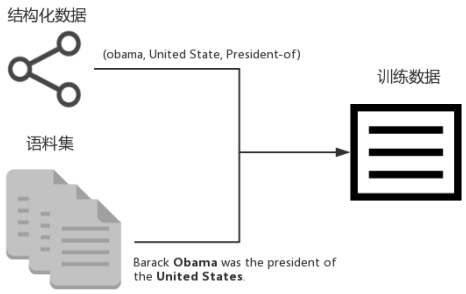

图 1 远程监督生成训练数据过程

远程监督的主要目的是减少构造数据集的成本，因此使用已有的知识图谱对语料集进行自动标注，最终生成训练数据，其产生训练数据的工作流如图 1 所示。已有的知识图谱（Freebase、DBPedia 等）或其他的结构化文本中提供表达关系 R 的实体对，如表 1 中的关系 President-of 以及实体对（Obama, United State），而语料集提供包含该实体对的句子，如表 1 中的 S1、S2。研究人员可以通过实体对齐等自然语言处理方法融合这两部分信息，从而得到带标注信息的数据，最后将这些数据作为关系抽取任务的训练样本，进行关系抽取。本文将通过这个过程完成的关系抽取任务称为远程监督关系抽取任务（Distant Supervised Relation Extraction, DSRE）。

表 1 远程监督训练样本示例

President_of(Obama, United States)	
S1	Barack Obama was the president of the United States.
S2	Obama lives in the United States with his wife.

1.2 本文框架

本文主要针对 DSRE 任务中的降噪方法以及长尾问题的解决方法进行阐述，第 2 节概述 DSRE 方法的发展以及相关深度学习技术，并梳理 DSRE 任务的关键问题；第 3 节分类介绍 DSRE 任务中的降噪方法，将模型根据优化的模块不同进行分类描述。第 4 节介绍近年来处理 DSRE 中数据的长尾问题的方法。第 5 节介绍 DSRE 任务中常用的数据集，以及常用的评测指标与评估方法。第 6 节展望 DSRE 任务的未来研究趋势。第 7 节进行总结。

2 远程监督关系抽取

2.1 基本模型

Mintz 等人[3]提出远程监督这个概念的同时，也为远程监督方法设定了基本假设，即如果两个实体 h, t 在已知的知识图谱中存在关系 R，则所有包含 (h,t) 的句子 s 都将表达关系 R，最终基于生成的训练集完成关系抽取任务，我们将这个原始工作模型称为 **Vanilla 模型**，相应的假设称为 **Vanilla 假设**。

Vanilla 模型生成训练数据的过程如图 2 所示。首先对语料集进行预处理，预处理步骤包括命名实体识别（Named Entity Recognition，NER）、词性（part-of-speech，POS）标注、依存分析等处理。这些是自然语言处理领域比较常用的分析方法，可以依赖外部工具得到，其中比较常用的工具有斯坦福大学的 CoreNLP 工具[1]，该工具可以对句子进行词性标注，以及词法特征、句法特征的挖掘等。完成语料集的预处理之后，需要对 NER 步骤中获得的实体进行实体匹配（Entity Matching），即将语料集中的实体对应到知识图谱的实体中，以便于后续关系标签的标注。这一阶段需要已有的知识图谱提供实体信息，一般可以通过对应的 API 获得实体信息，或者下载知识图谱资源文件并在上面进行查询。如图 2 所示，实体匹配将实体 Martha Washington 对应到知识图谱中编号为"Q191789"的实体，而 George Washington 则对应到编号"Q23"的实体。当句子中的两个实体都在知识图谱中找到对应的实体编号以后，则开始对句子进行特征提取。Mintz 等人[3]提取了文本特征中的词法特征（Lexical Feature）、句法特征（Syntactic feature）、实体类型标签等特征，其中词法特征包括：

1) 两个实体之间的词汇序列；
2) 每一个词例（token）的词性（POS）标签；
3) 实体的开始位置索引；

[1] https://stanfordnlp.github.io/CoreNLP/

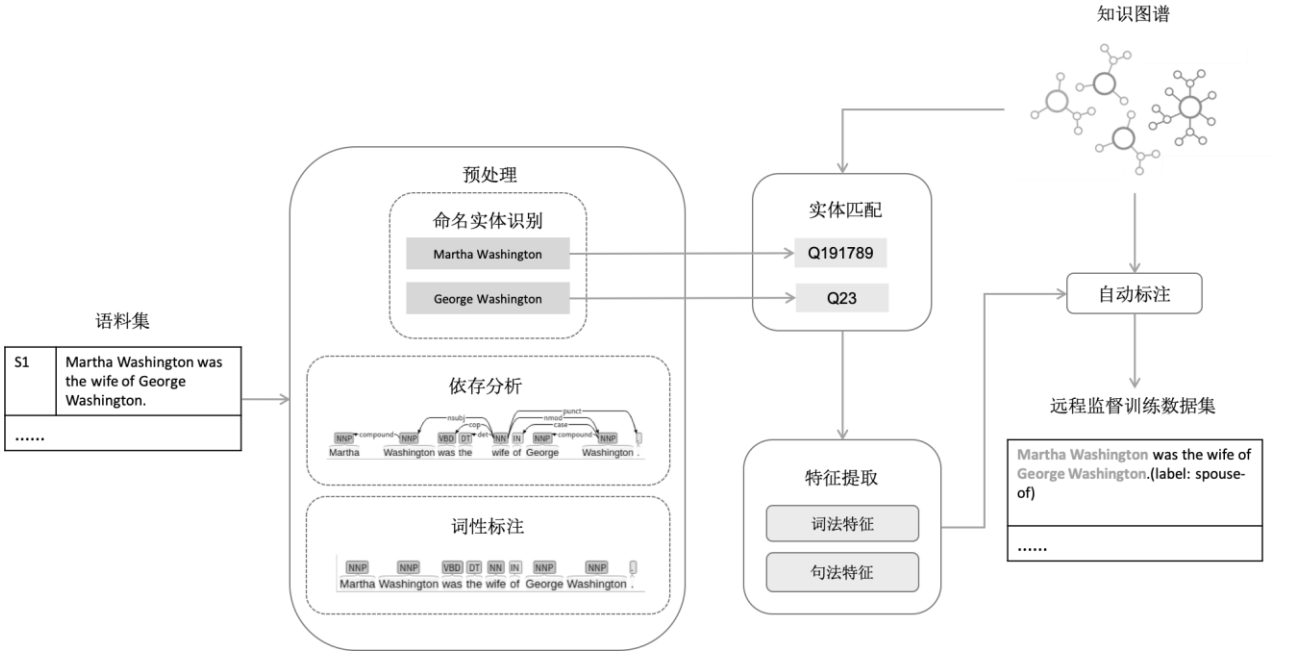

图 2 Vanilla 模型生成训练数据的过程

4) 第一个实体左边的 k 个词以及他们的 POS 标签；

5) 第二个实体右边的 k 个词以及他们的 POS 标签。

而句法特征则包括两个实体之间的句法依存树，以及与其中一个实体相连，但又不在依存路径内的一个节点。Vanilla 模型最后将这些特征进行融合，将融合后的特征作为后续关系抽取任务的输入进行训练。

生成训练数据的最后一个阶段就是对句子进行自动标记（Labelling），即从知识图谱中获取相应实体对的关系。获取的方式同样可以通过 API 或者下载的资源文件进行查询，最终得到由远程监督生成的关系抽取训练数据集。在这个数据集上，Vanilla 模型使用了一个逻辑回归多分类器进行关系抽取。

远程监督方法的应用，使关系抽取任务摆脱了对人工标注数据集的依赖，基于已经存在的结构化数据进行数据的自动标注，大大降低了构建数据集的成本，然而这个构造出来的数据集并不是完美的，依然存在着很多可以改善的地方，这些问题都会对后续关系抽取任务的性能造成影响，具体会在下一节进行讨论。

2.2 关键问题

2.2.1 错误标注问题

在 Vanilla 模型的假设中，所有包含两个实体(h,t)的句子都标为知识图谱中的关系R，这样的确能对部分句子进行正确标注，这种被正确标注的句子被称为有效示例，但在实际情况中，一对实体出现在同一个句子中，不一定表达了某一种关系，而只是与同一个主题相关，因此，这一类句子被称为假正例（False Positive Instance，FPI）。在这种情况下就会存在错误标注（Wrong Labeling, WL）问题。

不仅如此，Vanilla 假设中默认同一个实体对只存在一种关系，即不会同时存在两个三元组(r_1,e_1,e_2)和(r_2,e_1,e_2)同时有效的情况,但在实际情况中同一个实体对在不同句子中可能拥有不同的关系，例如$(Obama, United\ State)$这个实体对，既可能是 president_of 的关系（表 1 中的 S1），也有可能是 live_in 的关系（表 1 中的 S2），直接将全部包含该实体对的句子都标注为其中一种关系，则关系分类器在进行 live_in 关系的参数学习时，也会学习了 president_of 关系的示例信息，从而影响关系分类器的性能。这种情况也归为 WL 问题。

如果错误标注的样本数量占比不大，模型可以在适当的噪声中获得更高的鲁棒性，但从远程监督获得的数据集中的 WL 问题却不容忽视，主要体现在假正例的数量占比较大，导致噪声数据对关系抽取的性能造成了负面的影响。Dumitrache 等人[10]

以由远程监督得到的 TAC-KBP 2013 英文填槽（Slot Filling）评测任务数据集[⌣]为例，抽取验证集中的 16 个关系中的所有句子让 15 个人进行标注，统计每个关系包含的示例中的假正例占比，统计结果如图 3 所示，可见在 16 个关系中，有 10 个关系的假正例的占比大于 50%，尤其是 origin 和 place_of_death，这两个关系的假正例占比大于 0.9。用这些具有高比重噪声的数据进行模型训练时，模型拟合的更有可能是噪音数据，得到的关系抽取模型并不能保障基于真实数据的关系抽取性能。

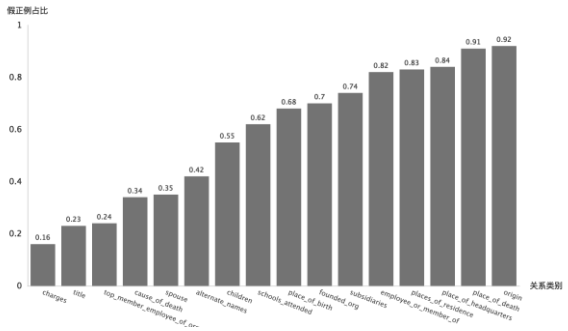

图 3 16 个关系中假正例占比

严重的 WL 问题意味着数据集的不可靠，模型可能从错误的样本中学习了过多的错误特征，从而降低模型的性能。因此将通过 Vanilla 假设获取到的数据直接用于关系抽取任务时，得到的模型的正确性是值得商榷的。

2.2.2 数据长尾现象

目前的远程监督主要是使用开放域的知识图谱以及语料集，例如常用的维基页面、纽约时报等开放域语料集以及 Freebase、WikiData 等开源知识图谱。这样的数据收集渠道比较多，成本也较低，而另外一些垂直领域知识图谱或者语料集的使用权往往掌握在企业手中，获取成本比较高。这样的数据来源对数据的质量有一定的限制，例如其中通用关系所占样本数比非通用关系所拥有的句子数多得多，类似于*place_of_birth*、*nationality*等关系的句子数量会更多，相比之下，往往那些有一定专业领域性的示例数量占比很低，例如*component_of*、*owner_of_shopping_center*等，造成远程监督生成的训练样本分布极度不均匀的现象。

以 Riedel2010 数据集[11]为例，这个数据集包含了 53 种关系（包括 NA 关系），对其中除了 NA 关系外拥有示例数量最多的前 20 个关系进行示例数量上的可视化，情况如图 4 所示。由图 4 可以看出，地点与地点之间的*contains*关系拥有的示例数比其他关系拥有的示例数多了几倍，是第二的*nationality*关系的七倍。在这 20 种关系中存在着严重的长尾（Long Tail，LT）现象：其中只有 10 种关系拥有的句子数超过了 1500 条，剩余关系拥有的示例数量都少于 1500 条，而没有在图中表示出来的关系示例数量甚至少于 206 条。我们将这种拥有示例数量较少的关系称为长尾关系。

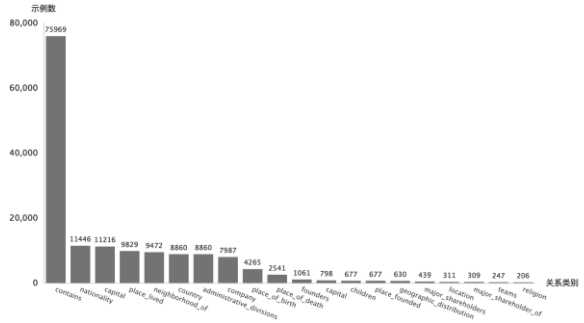

图 4 Riedel2010 数据集部分数据分布情况

长尾关系拥有的示例数量过少，不利于生成有效的关系提取器。同时由于在远程监督数据集中还存在 WL 问题（如上一小节所述），这就造成随着拥有的示例数量变少，长尾关系中包含的示例中错误标注的比例会更高，甚至会存在拥有的唯一一个示例都是错误标记的极端情况，那模型就更容易受到错误信息的影响，从而影响关系提取器在相应长尾关系的抽取性能，也降低了 DSRE 任务的整体性能。

2.3 研究进展

2.3.1 基于特征的模型

针对以上问题，研究学者为提高关系抽取性能，对模型设计及数据集构造等过程进行了改进和优化，提出了相应的解决方案。

首先为了降低远程监督过程生成的噪声数据的影响，有不少研究学者以 Vanilla 假设为切入点进行假设层面的改进，例如 At-Least-One 假设[11]以及多示例学习[12]的使用。同时针对真实语料集中存在的重叠关系的情况，学者们提出了多标签多示例学习（Multiple Instance Multiple Learning，MIML），

⌣ http://surdeanu.info/kbp2013/data.php

与 Vanilla 假设默认的每一对实体只有一个示例和一个标签不同，MIML 允许一对实体拥有多个示例和多个标签。这种将远程监督和 MIML 结合起来的解决方案，在一定程度上缓解了 WL 问题。

另外，也有学者使用基于大量特征训练来进行关系预测的主题模型（Topic Model）和模式相关性（Pattern Correlations）方法来降低噪声数据对模型的影响。这类方法称为基于特征的方法。

主题模型是在机器学习和自然语言处理等领域用来在一系列文档中发现抽象主题的一种统计模型，即从文档 d 中抽取出主题 t。当将主题模型应用到远程监督关系抽取时，则将包含实体对的句子视为文档 d，而句子所表达的关系视为主题 t。主题模型通过获取文本模式与关系之间的依赖来提高最终效果。目前 DSRE 使用的主题模型大都以隐含狄利克雷分布（latent Dirichlet allocation, LDA）[13]为基础，但各自选择的特征不同，从而导致了他们获取的模式不同。Yao 等人[14]在 2011 年提出了 LDA 的三种变形：Rel-LDA、Rel-LDA1 以及 Type-LDA，其中的 Rel-LDA 模型使用了句子中的三个特征：两个实体的名称以及他们之间的最短依存路径[9]。这些主题模型会以关系三元组（主体实体，客体实体，实体之间的依存路径）为输入，然后对这些三元组进行聚类处理，最后得到代表了不同关系的三元组的集合。

模式相关性更为直接地判断模式是否表达了目标关系，在不改变原始假设的情况下减少远程监督生成的错误标签的数量。Takamatsu 等人[14]在没有使用 At-Least-One 假设的前提下，提出了一个生成模型，可以用来预测每种模式是否通过隐藏变量表达某种关系，从而将频繁出现的模式上的错误标签移除，该模型的基本思想是: 如果文本 c 与关系 r 的参数对匹配，或与表达关系 r 的其他模式的参数对具有高度重叠，则 c 表达关系 r。黄蓓静等人[15]提出了基于句子模式聚类和模式评分对远程监督训练数据集进行降噪的方法，得到了噪声更少的数据集。

但以上基于特征的模型以及 Vanilla 模型在进行关系抽取时仅仅依赖于预先设计好的特征，例如句法依存树，词性标注等，这些特征通常从 NLP 工具中获取，如 2.2 节所示，这样就造成了 NLP 工具提取特征时产生的误差会传递到关系模型中，进一步降低模型的准确性。

2.3.2 基于深度学习的远程监督关系抽取

深度学习（Deep Learning，DL）是机器学习的分支，使用人工神经网络作为架构，对数据进行表征学习的一类算法。常见的深度学习框架包括卷积神经网络、循环神经网络、深度强化学习等，这些框架使从原始输入中提取高水平特征成为可能。目前，深度学习技术已经被成功应用在图像识别领域[16, 17]以及自然语言处理领域[18, 19]。

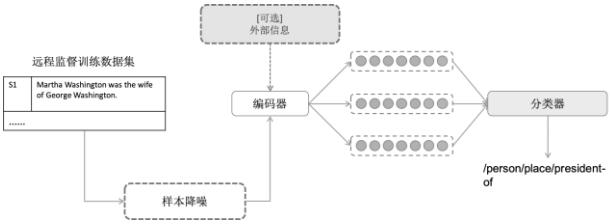

图 5 远程监督关系抽取通用工作流

在远程监督领域，越来越多的学者尝试摆脱特征工程，使用深度学习的方法进行关系抽取，在减少人工介入的情况下，提升关系抽取的性能。这些模型通用的工作流如图 5 所示，可以将其分为四个模块：样本降噪，外部信息融合，编码器以及分类器。首先通过远程监督得到关系抽取任务的数据集 S，可选地进行样本级别的降噪，然后对 S 中的句子 $\{s_1, s_2, \ldots, s_n\}$ 以及可选外部信息进行编码，将其从自然文本转化成计算机可以理解的语言，最后使用分类器进行分类，从而推断出每一对实体所拥有的关系。目前大部分远程监督关系抽取模型都基于该工作流进行设计，不同的模型会针对这四个模块中的一个或两个进行改进和优化，最终提升关系抽取性能。

图 6 远程监督关系抽取方法分类

在本文中，我们将远程监督关系抽取模型根据其优化重点分为四类：样本降噪，外部信息融合，编码器优化以及分类器优化，并将远程监督关系抽

取的方法按照图 6 进行分类整理，如表 2 所示，随后在每一小节中对相关的模型方法进行概述。

值得注意的是，一些生成学习、强化学习技术因为在多个阶段中都可以使用，属于较为通用的技术，在图中并没有明确将其归为哪一个阶段。

3 远程监督关系抽取的降噪

3.1 样本降噪

根据 Vanilla 假设构造出来的训练样本存在着大量的 WL 问题，优化编码层或者分类器的方法将降噪与远程监督关系抽取模型结合，降低噪声数据的影响，但这些方法忽略了一个实体对的所有句子都是假正例的情况。在这种情况下，一个独立而准确的句子级降噪策略是更好的选择。在进行关系抽取模型训练之前可以先对数据集进行清洗，在构造训练样本阶段就减少噪音数据，对更干净的数据集进行下一步的编码和分类，实验证明，这些更高质量的数据可以有效提高关系抽取任务的性能。这部分的研究工作主要分为 **At-Least-One 与多示例学习**和**样本清洗**。

3.1.1 At-Least-One 假设与多示例学习

Riedel 等人[11]认为当所使用的结构化数据（例如，知识图谱）并不是从语料集中构造出来的时候，两个实体出现在同一个句子是因为与同一个主题相关，但并不一定都表达了结构化数据中所标记的那个关系。他们通过人工标记的方法，统计基于 Vanilla 假设，使用纽约时报（New York Times, NYT）语料集后得到的数据集中的 3 个关系 contains、place_of_birth 以及 nationality 的错误标注情况，错误标注率分别为 20%、35%以及 38%。因此他们认为需要对 Vanilla 假设进行一定的宽松化处理，降低错误标注率，并提出了 At-Least-One 假设：

At-Least-One 假设. 假设两个实体之间具有关系 R，则在包含这两个实体的句子中，至少存在一个句子表达了关系 R.

At-Least-One 假设使远程监督得到的数据更加符合实际情况，但同时基于 At-Least-One 假设，在远程监督训练样本上的关系抽取任务变得更加复杂，因为我们无法确定哪一个句子真正表达了该关系。针对这个情况，Riedel 等人[11]将多示例学习（Multiple Instance Learning, MIL）的思想应用到模型中。多示例学习是 Dietterich 等人[12]在 1997 年提出的方法。在多示例学习中，训练集由一组具有分类标签的的包（bag）组成，其中每一个包拥有多个示例（instance），且每个包都有一个训练标记，而包中的示例是没有标记的，如果一个包中含有至少一正示例，则该包被标记为正，而如果一个包中所有示例都为负示例，则该包会被标记成负。与监督学习相比，多示例学习数据集中的样本标记是未知的；与无监督学习相比，多示例学习中每一个包的标记是已知的。这样的标记方法有效的提高了远程监督学习的容错性，更加符合实际应用场景。在 At-Least-One 假设中，所有包含同一个实体对的句子被分成一个组，形成包。

相比于 Vanilla 假设，At-Least-One 假设更加受研究学者欢迎，不少学者选择这个假设进行关系抽取任务，并根据这个假设设计关系抽取模型，如 Roth 等人[8]将分级主题模型（Hierarchical topic model）与 At-Least-One 结合在一起，在 KBP（Knowledge Base Population）数据集上得到比 MultiR 模型[20]与 MIML 模型[21]都高的 F1 值；也有学者选择使用 Riedel 等人[11]通过 At-Least-One 假设所构造出来的数据集 Riedel2010 作为实验数据，使其成为了远程监督领域最常用的数据集之一，如 Zeng 等人[22]在此数据集和多示例学习的基础上提出的分段卷积神经网络（Piecewise Convolutional Neural Networks，PCNN）模型等。

3.1.2 样本清洗

Takamatsu 等人[23]提出了一种减少错误标签数量的方法，使用否定模板来对样本进行清洗，具体流程如算法 1 所示。

首先他们借助实体的类别标签定义了句子的模板，如图 4 中的句子"Barack Obama was the president of the United States"，其模板为"[Person] president of [Country]"，而关系 *r* 否定模板（Negative Pattern）则是没有表达关系 r 的模板集合。然后对候选事实集合 *LD* 中的每一个事实（r, Pair, Sentence）都进行句子模板的抽取，得到相应的模板 *Pat*，如果 *Pat* 属于该关系的否定模板集，则将该句子移出相应的关系示例集合。经过这个算法，可以将部分错误标注的句子移除，提高数据集的质量。

表 2 远程监督关系抽取模型概览

分类	模型	作者	解决问题	知识图谱	语料集
样本降噪	At-Least-One	Riedel et al. 2010[11]	WL	Freebase[1]	NYT2010
	Generative model	Takamatsu et al. 2012[23]	WL	Freebase	Wikipedia[2]
	ADV	Wu et al. 2017[24]	WL	Freebase	NYT2010
	+DSGAN	Qin et al. 2018[25]	WL	Freebase	NYT2010
	AN	Han et al. 2018[26]	WL	Freebase	Wikipedia
	+RL	Qin et al. 2018[27]	WL	Freebase	NYT2010
外部信息融合	PCNN	Zeng et al. 2015[22]	WL	Freebase	NYT2010
	+D	Ji et al. 2017[28]	WL	Freebase	NYT2010
	Reside	Vashishth et al. 2018[29]	WL	Googe RE[3]	Text from Web
				Freebase	NYT2010
编码器优化方法	PCNN	Zeng et al. 2015[22]	WL	Freebase	NYT2010
	GloRE	Su et al. 2018[30]	WL	Freebase	NYT2010
	APCNN	Lin et al. 2016[31]	WL	Freebase	NYT2010
	CoType	Ren et al. 2017[32]	WL	Freebase	Wikipedia
				DBpedia[4]	Wiki-KBP[33]
				DBpedia	BioInfer[34]
	+STP+EWA+TL	Liu et al. 2018[35]	WL	Freebase	NYT2010
	HSAN	Zhou et al. 2018[36]	WL	Freebase	NYT2010
	BGWA、EA	Jat et al. 2018[37]	WL	Freebase	NYT2010
				Google RE	Text from Web
	Reside	Vashishth et al. 2018[29]	WL	Google RE	Text from Web
				Freebase	NYT2010
	+HATT	Han et al.2018[38]	LT	Freebase	NYT2010
	+KATT	Zhang el al.2019[39]	LT	Freebase	NYT2010
	C^2SA	Yuan et al. 2019[40]	WL	Freebase	NYT2010
	RAPCNN+BAG_ATT	Ye et al. 2019 [41]	WL	Freebase	NYT2010
分类器优化方法	MultiR	Hoffmann et al. 2011[20]	WL	Freebase	NYT2010
	MIML	Surdeanu et al. 2012[21]	WL	Freebase	NYT2010
				Wikidata[5]	KBP(2010,2011)
	MIML-RE	Jiang et al. 2016[42]	WL	Freebase	NYT2010
	+STP+EWA+TL	Liu et al. 2018[35]	WL	Freebase	NYT2010
	+TM	Luo et al. 2017[43]	WL	Freebase	NYT2010
				Wikidata	Wikipedia
	HRL	Takanobu et al. 2018[44]	WL	Freebase	NYT2010
	RLRE	Feng et al. 2018[45]	WL	Freebase	NYT2010
	+SL	Liu et al. 2017[46]	WL	Freebase	NYT2010
	EPNet	Sun et al. 2019[47]	WL	Freebase	NYT2010

[1] https://developers.google.com/freebase/

[2] 维基百科的信息框以及其他表格类数据。

[3] http://research.googleblog.com/2013/04/50000-lessons-on-how-to-read-relation.html

[4] DBPedia Spotlight:httpp://spotlight.dbpedia.org/

[5] https://www.wikidata.org/wiki/Wikidata:Main_Page

算法 1. 减少错误标签算法.

输入：通过远程监督得到的数据集LD，关系r的否定模板$NegPat(r)$

输出：清洗后的数据集LD'

```
FOR EACH r, Pair, Sentence in LD do:
    pattern Pat ←the pattern from (Pair, Sentence)
    IF Pat ∈ NegPat(r) then:
        remove (r, Pair, Sentence) from LD
    ENDIF
ENDFOR
RETURN LD
```

这个方法克服了传统基于模板的关系抽取方法的局限，在构建模板的过程中不需要人工的介入，实现了自动化的生成模板，然后进行关系分类。

随着深度学习的研究逐渐深入，对抗学习、强化学习等理论日渐成熟，研究学者们开始尝试将这些技术应用到样本的清洗阶段上，并且获得了不错的效果。

基于对抗学习。生成对抗网络（Generative Adversarial Networks, GAN）[48]是一种对抗式的，不依赖于任何先验假设的生成式模型。这种对抗指的是生成器（Generator）和判别器（Discriminator）的相互对抗。生成器（Generator）从隐含层变量中学习生成更“真实”的样本，以混淆判别器，判别器（Discriminator）则尽可能去判别该样本是否为生成的假样本，而当判别器无法正确判定数据是否由生成器生成时，生成器的训练就完成了。GAN 技术在刚提出时，被广泛使用在计算机视觉领域，主要被用于图像分类，后来也被应用在自然语言处理领域的文本分类任务上[49]。

2017 年，Wu 等人[24]将对抗学习的思想运用到远程监督的关系抽取任务中，在编码之前加入了对抗学习生成的噪声信息e_{adv}来训练分类器，在 Riedel2010 数据集和华盛顿大学（University of Washington，UW）数据集[50]上的实验证明加了对抗学习的模型会比原模型有一定的提升。

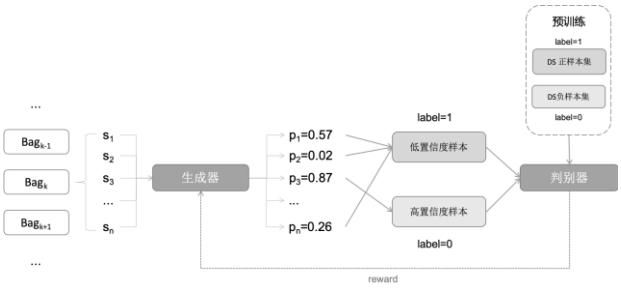

图 7 +DSGAN 模型工作流

在 2018 年，Qin 等人[51]在 Wu 等人[24]的基础上，改进了对抗学习在 DSRE 任务上的应用，利用对抗学习得到的生成器，对数据进行清洗，提供了一个更高质量的数据集给后续的关系抽取，并因此提出了+DSGAN 模型，其工作流如图 7 所示。

在+DSGAN 中，生成器对每一个输入的句子s_i都会计算该句子为有效示例的概率$p_G(s_i)$，通过相应的判别器计算得到的有效示例概率为$p_D(s_i)$。根据对抗学习的思想，生成器要使生成的数据s_j被判定为真实样本的概率$p_D(s_j)$更高，而判别器的目标是将生成器生成的样本的特征尽可能的识别出来。当判别器无法辨别样本是否来自生成器生成时，生成器的性能达到最好，这时使用生成器可以对某一样本为有效示例的概率进行计算，过滤掉低置信度的样本，从而提高样本的质量。最终的实验结果证明在 Reidel2010 数据集上，增加了 DSGAN 的经典模型（例如 PCNN）与没有添加 DSGAN 的原模型相比，关系抽取的 AUC 值普遍提升了 1%。

同样将对抗学习应用到 DSRE 任务中的 Han 等人[26]，他们认为从训练数据中采样生成样本能够更好的定位实际的噪声情况，因此提出了 AN 模型。在 AN 模型中，分为判别器和采样器，判别器是为了判断句子被标注正确的概率，而采样器则是从置信度低的样本集中选择出最具有迷惑性的句子去欺骗判别器，这两部分通过对抗学习进行训练，在训练过程中由于噪声数据无法降低采样器和判别器的损失值，因此在对抗训练期间噪声将逐渐被滤除。最后 AN 模型中的采样器可以有效的区分样本的置信度高低，而判别器可以更好的对实体对的关系进行分类，提高模型的性能。

基于深度强化学习。强化学习（Reinforcement Learning，RL），也称增强学习，是一类从与环境的交互中不断学习问题以及解决这类问题的方法，通过试错的方式来完成特定目标（比如取得最大奖励值）。RL 系统包括以下基本要素：状态s，动作a，策略π、状态转移概率P和奖励函数R，它们之间的联系为：智能体在环境状态s_t下根据策略π来决定下一步的动作a_t，然后接收到环境反馈的奖励r_t，并以转移概率p_t^a转移到下一个状态s_{t+1}。将深度学习与强化学习相结合后的技术为深度强化学习（Deep Reinforcement Learning，DRL）。

Qin 等人[27]提出+RL 模型，该模型使用深度强化学习来生成假正例的指示器，通过这个指示器可以识别出 FPI，并将其重新分配到负样本中，提高样本质量，并且最终提高关系抽取任务的性能，具体过程如图 8 所示。

图 8 RL 更新样本集的过程

+RL 模型假设当不正确标注的样本被过滤掉以后，关系分类的性能将有所提高，因此 Agent 可以根据关系分类任务的 F1 值判断这个远程监督的句子是否保留，因此在第 i 个 epoch，Reward 为$R_i = \alpha(F_1^i - F_1^{i-1})$，只有当 F1 值得到提升时，Reward 才为正。而在决策网络，+RL 模型使用了简单的 CNN 来进行关系分类，目的是判断当前的句子是否表达了目标关系类别，可以简化为二分类问题。在进行强化学习之前，作者使用一小部分的标注数据来训练决策网络，而在强化学习的每一次迭代中，Agent 都会从训练集中的正样本集P_t^{ori}中选择一个噪声样本集ψ_i，将其从正样本集中移除后可得到新的正样本集$P_t = P_t^{ori} - \psi_{i_1}$，同时将$\psi_i$重新分配到负样本集$N_t = N_t^{ori} + \psi_i$中，再使用这个数据集训练关系分类器，为了取得更好的 Reward，Agent 每一次都会移除包含更多 FPI 的样本集，最后每一个关系类都得到一个可作为 FPI 的指示器的 Agent，用于分辨出当前类别下的 FPI。

3.1.3 分析比较

样本降噪主要分为两类，一类是改变样本分布的假设，另一类则是样本清洗。早期的 Vanilla 假设间接导致了生成训练数据时也生成了噪声数据，因此对 Vanilla 假设进行有效的改进，可以获得新的容错性更高的数据集，以 At-Least-One 假设为例，该假设不强调每一个包含实体对（h，t）的句子都表达了 KG 中所标注的关系 *R*，只承认包中至少会有一句表达了关系 *R*，以包为训练基本单位，改变了原来以一个单句作为训练基本单位的情况。在理论上，At-Least-One 假设比 Vanilla 假设出错的概率会更低，即一个包中存在的示例都是假正例的情况会比某一个句子是假正例的情况更少。

而样本清洗的实现方式有两种，一种是直接作用于样本清洗阶段的方法，如+DSGAN 模型和+RL 模型等，这类模型与关系抽取的模型是独立的，可以作为工作流中一个可插拔的组件，与其他关系抽取模型配合使用，提高 DSRE 任务的效果；另一种则是与关系抽取模型结合在一起，在进行关系抽取的过程中不断对样本集执行增减操作，如 ADV 模型和 AN 模型，这类模型在关系抽取的训练过程中对样本集进行动态的增减，以不断的获得更低的损失值，提升 DSRE 任务的效果。

3.2 外部信息融合

深度学习方法可以自动对输入的低阶特征进行组合变换，得到高阶特征，这也是深度学习方法能够在一定程度上避免特征工程的原因之一。但在自然文本中，输入的字或者词都是离散稀疏的，不像图像一样是连续稠密的，因此变换得到的高阶特征并不是特别有效，学者们可以利用先验知识构造少数特征以提高模型的性能。

3.2.1 位置特征

位置特征（Position Feature）是指当前词离目标实体对的距离，这个距离可以帮助定位目标实体的位置。在 2014 年，Zeng 等人[52]将位置特征输入到传统关系抽取的模型训练过程中，并在 2015 年，将其应用到 PCNN 模型[22]中。位置特征指当前词到两个实体之间的相对距离，例如句子“Obama is the

44th and current president of United States”中，current 离实体对（Obama, the United States）的相对距离分别为 5 和-3，随后，这两个相对距离会被映射为两个 d 维的向量（d 一般为超参数），得到d_1和d_2，从而得到位置特征$PF=[d_1,d_2]$，最后与词特征（Word Feature, WF）进行拼接，最后得到编码器所需要的输入$[WF;PF]$。

3.2.2 实体描述信息

2015 年，Ji 等人[53]认为两个实体之间的关系 r 是主体实体（e_1）到客体实体（e_2）的映射，因此可以用$e_1+r\approx e_2$来对(e_1,e_2,r)进行建模。而在 2017 年 Ji 等人[28]提出的+D 模型，一方面参考这个映射关系，使用e_1-e_2来作为关系 r 的表征，另一方面他们认为实体的描述信息可以提供丰富的背景信息，因此从 Freebase 和维基页面中提取了实体的描述信息，将其加入模型训练过程中的损失函数中，即$L_e=\sum_{\{i=1\}}^{D}\| e_i-d_i \|_2^2$，其中$D=\{(e_i,d_i)|i=1,\ldots,|D|\}$。最后在 Riedel2010 的数据集的实验结果证明，实体的描述可以提供更多的实体信息，可以优化实体的表征，最终提高关系抽取的性能。

3.2.3 外部知识

此外，还可以使用其他知识图谱提供的信息作为特征输入，例如实体别名信息、关系别名信息、实体类别层级关系等等。在 2018 年，Vashishth 等人[29]在 Reside 模型中使用开放信息抽取系统（Open Information Extraction, Open IE）从句子中抽取关系的名称，结合从 PPDB 数据集[54]中获得关系的别名信息，根据关系名称和别名在嵌入空间中的相关距离，过滤掉无关的关系类别。

3.3 以编码器为中心的降噪方法

在编码器中，首先要将句子中的词语表示成计算机可以理解的语言，并且可以有效表示出样本的特征，即用向量表示自然语言中的词汇，然后将得到的词向量序列转化成句子矩阵。这个部分一般有两步，第一步称为词嵌入（Word Embedding），即将自然文本映射成向量；第二步我们称为编码（Encoding），即使用编码器对上一步获得的向量进行特征提取。在本文中，我们将以编码器为中心的降噪方法根据编码流程分为嵌入级优化方法和编码器级优化方法。

3.3.1 嵌入级优化方法

词嵌入（Word Embedding）是从单词等离散对象映射到向量的一个技术，也称词向量，是目前广泛使用的自然语言处理技术之一。词向量是一种既能表示词本身又可以考虑词间距离的表示方法，即相似的词在向量空间中处于相近的位置。研究学者比较常用到的是使用神经网络预训练好的词向量，如 GloVe[55]、Word2Vec[56]等。在 DSRE 任务中，研究人员也会根据任务的特殊性，对嵌入的计算进行优化。2017 年，Ren 等人[32]提出了对关系和实体的联合嵌入，提出了 CoType 模型。整个过程分为三部分工作：对关系类别的建模、对实体类别的建模、对实体-关系之间互信息的建模。其中针对这三部分工作，Ren 等人提出了指代级别特征共现、局部标签假设以及实体-关系互作用假设，以及三个假设对应的损失函数，最后将三个损失函数相加后得到最终的损失函数。实验证明，CoType 模型在开放域的 Riedel2010、Wiki-KBP 数据集以及生物医学领域的 BioInfer 数据集上进行关系抽取任务时，比 MultiR 模型，DS-Joint 联合抽取模型[57]表现更好。

表 3 关系表述与知识图谱中关系的共现分布的示例

关系类别	*[born in]*	*[died in]*
place_of_birth	1868	14
nationality	389	20
place_of_death	37	352

Riedel 等人[58]认为模型可以通过学习文本和知识库关系在同一连续潜在空间中的表征来完成联合推理，这个想法在完成知识图谱补全任务上得到了验证，获得了高准确率。然而 Su 等人[30]认为在远程监督背景下，这种使用基于单句的局部特征的方法，会更容易受 WL 问题的影响，因此提出了一种使用全局统计的思路：每一个句子文本中的关系表述（textual relation），可以用其在知识图谱关系中的共现分布来表示，并提出了 GloRE 模型。如表 3 所示，由文本表述[born in]与知识图谱中的关系共现分布可得知其很大可能是 place_of_birth 关系或者 nationality 关系，而不是 place_of_death 关系。因此虽然训练集中存在错误标签，但从全局统计的角度来看，错误标签占比较低，可以有效的突出正确

标签，提高模型的准确率。

实验证明，使用全局统计来代替局部统计的方式得到的关系嵌入，可以捕获更多的补充信息。作者用新的关系嵌入对 APCNN 模型进行扩展优化，P@1000 从 83.9％提高到 89.3％，错误率下降 33.5％。

3.3.2 编码器级优化方法

为了使计算机能够理解句子中的特征，我们需要对句子及其特征进行编码，对低阶特征进行变化组合得到高阶特征。随着 seq2seq 模型[59]的提出,编码器-解码器（Encoder-decoder）结构被广泛使用，在 DSRE 任务上，也有不少模型会在句子编码器（Sentences Encoder）上改进和创新，以求得到更好的句子表征，最终提高模型的表现。这些编码器大部分以卷积神经网络或者循环神经网络为基础，来实现对每一个示例的特征提取。

基于卷积神经网络。卷积神经网络（Convolution Neural Network，CNN）是一种具有局部连接、权重共享等特征的深层前馈神经网络，有一个或多个卷积层、池化层和最后的全连接层组成。其中卷积层的作用是提取一个局部区域的特征，而池化层的作用是进行特征选择，降低特征数量，并从而减少参数数量，池化的方法包括最大池化、均值池化、最小池化等。

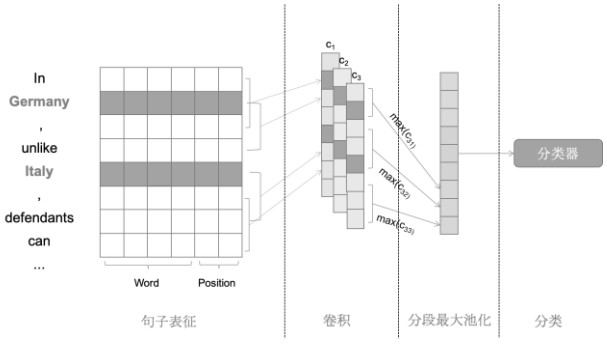

图 9 PCNN 模型示意图

Zeng 等人[52]在 2014 年为了获取整个句子的特征，使用了 CNN 进行特征的提取，并在最后得到句子表征$m = \{m_1, m_2, \ldots, m_{n_1}\}$。在 2015 年，Zeng 等人[22]对标准的 CNN 网络结构进行了改进，提出了使用分段卷积网络（Piecewise Convolutional Neural Networks，PCNN）来对句子进行编码，模型如图 9 所示。PCNN 编码器包括两个阶段：卷积（Convolution）和分段最大池化（Piecewise Max Pooling）。卷积层中使用 n 个卷积核对输入的序列 S进行卷积后，得到 n 个特征图（feature map）。

为了获取每一个 feature map 中最有意义的特征，一般会对卷积后得到的 feature map 进行最大池化处理（max pooling），然而，在关系抽取任务中，单纯的使用最大池化会使隐藏层大小快速下降以及抽取出来的特征粒度不够细，因此，PCNN 采用了分段最大池化：以两个实体为分割点，将一个句子分隔成三个部分，分别进行最大池化，最后将这三个结果进行拼接后进行非线性变换，最后使用 softmax 函数计算关系概率分布。

基于循环神经网络。循环神经网络（Recurrent Neural Networks，RNN）是一类具有短期记忆能力的神经网络，可以用来处理序列数据。然而当输入序列较长时，在参数学习过程中会存在梯度爆炸和消失问题，为了解决这个问题，人们对 RNN 进行改进，提出长短期记忆（Long Short-Term Memory，LSTM）[60]和门控循环单元（Gated Recurrent Unit，GRU）**[61]**。在自然语言处理领域，为了更好的捕捉到句子词汇的依赖关系，尤其是经常考虑到高阶依赖关系的关系抽取任务，往往会使用 RNN、LSTM 以及他们的变体等作为编码器模块对句子进行编码。

在 2016 年，Miwa 等人[62]使用 Bi-LSTM 对句子进行编码，同时使用树结构的 Bi-LSTM 来对句法依存树进行编码，提出了一种新颖的端到端神经模型来提取句子中的实体和关系。该模型的实体检测和关系抽取两个任务的参数共享，实现了联合抽取。最后分别在 ACE2005 和 ACE2004 两个数据集上获得了 F1 值的 12.1％和 5.7％的提升。类似的，2018 年，Liu 等人[35]也使用了双向门控循环单元（Bidirectional Gated Recurrent Unit，Bi-GRU）来对句法依存树进行编码，同时为了更好的捕捉到句子中的实体信息，加入了多层注意力对句子进行编码，包括实体层面的注意力，词级别的注意力，句子级的注意力。其中实体层面的注意力是为了在句子中突出目标实体，因此定义为：

$$a_{it}^{e} = \begin{cases} 1, & t = head, tail \\ 0, & others \end{cases}$$

仅当第 t 个词属于主体或客体实体时，$a_{it}^{e} = 1$。此外，参考 Zhou 等人[63]在 2016 年提出的词级别的注意力：

$$a_{it}^{w} = \frac{exp(h_{it}A^{w}r^{w})}{\sum_{t=1}^{T} exp(h_{it}A^{w}\, r^{w})}$$

然后结合这两个注意力以及 Bi-GRU 编码后的结果得到每一个句子的编码。

Yan 等人[64]认为 PCNN 模型打破了句子的一致性和内在关系，因此提出使用 PLSTM-CNN（Piecewise-LSTM Convolutional Neural Networks）来作为句子编码器。该编码器首先使用词嵌入拼接来表示句子，接着使用 PCNN 进行初步的编码，在 PCNN 编码后，使用 Bi-LSTM 对 PCNN 的输出进行二次编码后，得到最后的句子表征。同时，作者还提出了关系生成器来直接学习包的共享表征。他们将句子向量从原始语义空间转换为对应的关系表示空间。通过生成器转换后得的矢量可以看作对应关系中句子的另一种表征，可以对不相关的噪声语义进行过滤。作者的实验证明将 PLSTM-CNN 作为句子编码器时，模型的整体性能优于使用 PCNN 编码器的模型，在增加了关系生成器后，模型的平均 P@N 值要优于其他没使用关系生成器的模型。

基于图卷积神经网络。 当特征中存在类似图的特征时，例如句法依存树、短语树等特征，除了使用上文提到的树结构的 Bi-LSTM，也可以采用对图类特征进行编码的模块。为了获取图的空间信息，Marcheggiani 等人[65]在 2017 年提出了使用了图卷积网络（Graph Convolutional Network，GCN）进行句子的编码，因此，在 2018 年，Vashishth 等人[29]提出了模型 Reside：在选择 Bi-GRU 对文本表征（text representation）进行编码的同时，选择使用 GCN 对句法依存树进行编码，保留其图的特性。

最后获得的编码则由 GCN 编码后的序列$h_{i_{k+1}}^{gcn}$以及 Bi-GRU 后的序列h_i^{gru}拼接而成:

$$h_i^{concat} = \left[h_i^{gru}; h_{i_{k+1}}^{gcn}\right]$$

基于句子级别的注意力机制。 注意力机制（Attention Mechanism）可以通过自上而下的信息选择机制来过滤掉大量的无关信息，解决信息过载的问题，通常用作信息的筛选。注意力机制的计算可以分为两步，一是在所有输入信息上计算注意力分布，二是根据注意力分布来计算输入信息的加权平均。目前，注意力机制已经在语音识别、阅读理解、文本分类等多个任务上取得了很好的效果。

在 Zeng 等人[22]提出的 PCNN 模型存在着明显的不足之处：只选择了每个包中置信度最高的示例作为包的表征，这样导致包中的其他示例的信息被丢失，因为一个包内也可能存在多个真正例（True Positive Instance），而这些真正例可以提供更多的特征和信息，可以对包表征进行更大程度的优化。

针对这个情况，Lin 等人[31]在 PCNN 的基础上，增加了句子级别的注意力机制，提出了 APCNN 模型，该模型充分利用包中的其他示例的信息：以$s = \sum_i \alpha_i\, x_i$来作为包的表征，其中$\alpha_i$是权重比例，$x_i$ 为包中第 i 个句子的表征。Lin 等人定义了两种计算权重的方法，一种是直接取包中所有句子表征的均值，另外一种则是选择注意力的方法：

$$\alpha_i = \frac{exp(e_i)}{\sum_k exp(e_k)}$$

其中，$e_i = x_i A r$，r为关系 r 的标签的表征向量，A为加权对角矩阵，这样e_i在一定程度表示了相应示例与关系标签的相关性。实验结果证明了使用选择注意力机制计算权重可以更好的减轻包中噪声示例对包表征的影响，同时充分学习到真正例的特征，得到更高的模型性能。

在远程监督的背景下，APCNN 模型充分考虑了包中其他示例的信息，同时使用选择注意力机制来尽量降低噪声数据的影响，但这个方法依然存在可以改进的地方：一方面是使用包中所有示例信息，会造成计算量的增加；另一方面 APCNN 无法处理整一个包都是噪声数据的极端情况。不少学者以 APCNN 为基础进行模型的优化，并且都取得了性能的提升。

针对上述第一种情况，Zhou 等人[36]提出 HSAN 模型，该模型为了避免庞大的计算量，计算包中的句子与目标关系 r 的相关度，然后选择相关程度最高的 m 个句子来进行包表征的计算，同时考虑到句子中每个词的重要程度应该有所不同，作者还在句子表征的计算上增加了词级别的注意力（Word Attention，WA），最终获得选择的句子的表征$g_i = [g_{ic}; g_{ia}]$，其中g_{ic}为句子通过 PCNN 编码器后得到的表征，而g_{ia}则为句子通过双向长短期记忆网络（Bidirectional Long Short-Term Memory，

Bi-LSTM）编码后加入词级别注意力得到的表征。HSAN 模型能够在保证模型性能损失程度可接受的同时，使模型在每个包上训练时间从 91.56ms 下降到 17.55ms。

而为了解决 APCNN 对噪声包没有进行很好的处理的问题，2019 年，Yuan 等人[40]在 APCNN 的基础上，提出除了需要充分利用包内其他示例的信息意外，还需要关注类与类之间的信息，并根据这个思路提出了跨关系跨包注意力模型（Cross-relation Cross-bag Attention，C^{2}SA），其中，跨关系（Cross Relation）指的是关注关系类别之间的信息，即若句子 A 在关系R_1上的得分高，那很可能在与R_1对立的关系R_2上的得分会低，因此 Yuan 等人在计算包的表征时，对 APCNN 的注意力权重进行了修改，不同于 APCNN 模型只关注当前句子所标注的关系，C^{2}SA 在计算示例间的注意力权重时还关注了其他的关系，最后得到新的包的表征$b_{i,k}$。此外他们将所有标志为同一个关系的包组合起来，并将其组合得到一组包$\boldsymbol{B} = \{B_1, B_2, \ldots, B_{n_s}\}$命名为超包（super bag），同时将训练样本从包变成超包。超包$\boldsymbol{B}$的表征计算为$f = \sum_{i=1}^{n_s} \gamma_i \cdot b_{i,k}$，其中$\gamma_i$为包$B_i$所拥有的跨包注意力权重。

类似的，Ye 等人[41]也改进了包内注意力的计算，以及增加了跨越包的注意力，提出了可关系感知的模型 RAPCNN+BAG_ATT，以获得新的表征。与 Yuan 等人[40]不同的是，在计算包B_i中第 j 个句子的表征$x_{i,j}$与第 k 个关系的匹配度时，没有使用两者的相似度进行计算，使用了向量的内积进行计算：$e_{kj}^{i} = r_k \, x_{i,j}$。获得新的包的表征后，Ye 等人也将判定为相同关系的包组合起来一组包$G = \{g_1, g_2, \ldots, g_n\}$，成为新的训练样本单位，其表征的第 k 行表示为：$g_k = \sum_{i=1}^{n} \beta_{ik} b_k^i$，而$\beta_{ik}$为第 i 个包判定为第 k 个关系的置信度γ_{ik}的归一化结果，置信度可通过每个包与同组其他包的相似度来计算：$\gamma_{ik} = \sum_{i'=1,\ldots,n,i' \neq i} b_k^i b_k^{i'\mathrm{T}}$。

C^{2}SA 模型和 RAPCNN+BAG_ATT 模型都将标注了同一个关系的实体对都组合起来，在对组合后整体进行关系分类的训练，有效规避一整个实体对包都为噪声句子的情况，模型在学习过程中可以更低程度的被这些噪声所影响，更加准确的进行特征学习。

考虑到大部分使用句子级注意力的模型都是使用了一维（1-D）的注意力，Du 等人[66]认为 1-D 注意力没有关注到句子的不同的语义层面，提出了在模型中使用二维（2-D）词级别注意力，帮助模型从不同的语义层面来学习权重分布。模型主要分为三部分，第一部分包括输入层（input layer）、嵌入层（embedding layer）以及 BiLSTM 层；第二部分是为了解决目标实体对的表征学习，包括词级别的自注意力层、文本表征层以及平滑层；第三部分是包级别的表征学习，包括句子级别的注意模型、选择表征层以及输出层。其中 2-D 注意力主要用于第二部分的词级别的注意力层，定义为$A_{L_1} = softmax(W_{S2}^{L1} tanh(W_{S1}^{L1} H))$，其中 L1 指的是模型中的第一级注意力，即词级注意力，最终句子的表征M_{L1}为$M_{L1} = A_{L1} H^T$，同时考虑到如果每次注意力都取相似的权重的话，则M_{L1}中会存在大量的冗余信息，因此，他们参考 Lin 等人[67]提出的惩罚项，来限制A_{L1}，使A_{L1}要尽量接近于正交矩阵，则A_{L1}中每一行权重最大的地方尽量不出现在同一列，从而减少M_{L1}的冗余信息。

基于词级别的注意力机制。不同于以上使用句子级别的注意力机制的模型，Jat 等人[37]提出了基于词注意力的 Bi-GRU 模型（Bi-GRU based Word Attention Model，BGWA）模型如图 10 所示。

BGWA 模型对句子的词向量的计算使用与训练的词向量，然后对其使用 Bi-GRU 网络进行编码，然后再对每一个词加入词级别的注意力（图中的α_i），同时参考 PCNN 的分段池化的思想，对使用

词注意力后得到的序列使用分段最大池化，最后得到整个句子文本的向量表示。

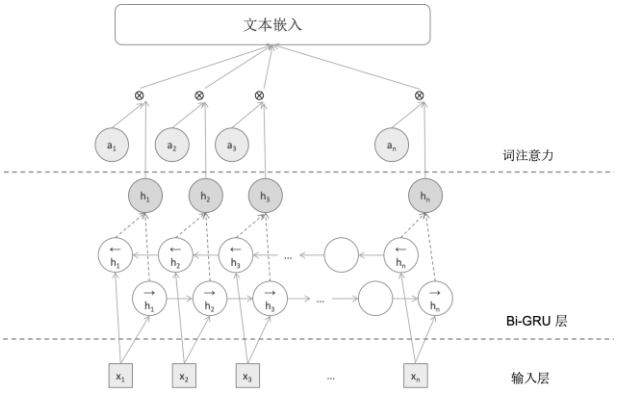

图 10 BGWA 模型示意图

基于实体级别的注意力机制。 考虑到实体的类别信息有助于缩小关系分类的范围，例如 Person 与 Person 之间不会存在 contain 等物体间的关系，Huang 等人[68]针对有监督的关系抽取任务提出了实体注意力模型（Entity Attention，EA）。Jat 等人[37]对 EA 模型进行了需要修改并将其使用到了远程监督关系抽取任务上。原 EA 模型是针对单个句子作为输入的，改进后的 EA 模型以一整个包的句子进行输入，使用 Zeng 等人[22]的方法从一个包中选择得分最高的句子作为有效示例。修改后的 EA 模型包括两个部分，如图 11 所示。

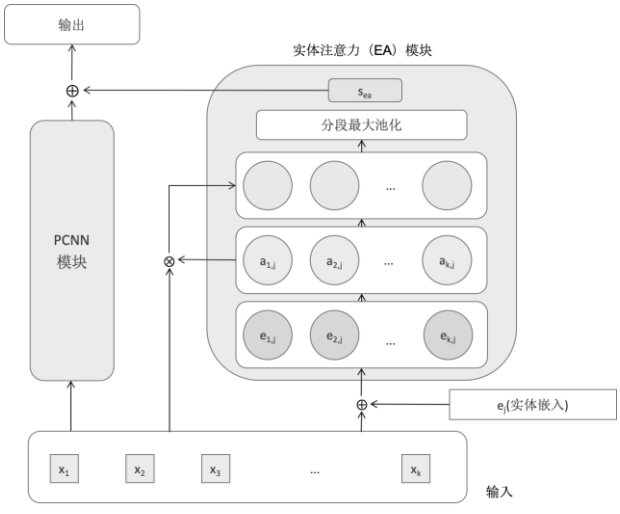

图 11 EA 模型示意图

改进后 EA 模型分为两个模块，其中 PCNN 模块用于对被选中的句子进行编码，而在实体注意力模块中，将句子中的每个词的词嵌入x_i分别与第 j 个实体嵌入e_j^{emb} 拼接后通过线性计算可得出每个词对第 j 个实体的注意力：

$$e_{i,j} = [x_i, e_j^{emb}] \times A_k \times r_k, i \in [1,k], j \in \{1,2\}$$

其中，i 意味着句子中的第 i 个词。对$e_{i,j}$进行 softmax 运算得到每个词的注意力得分$a_{i,j}$，与每个词的词嵌入进行元素积后得到加权的词嵌入，然后通过分段最大池化处理得到实体注意力模块的输出，最后将两个模块的的输出拼接后进行关系分类。

Jat 等人[37]还将 BGWA、改进后 EA 这两个模型进行融合，在 Reidel2010 数据集上以及 GDS 数据集上都达到了 state-of-art。

3.3.3 分析比较

编码层在工作流中担任需要将自然文本格式的原始输入以及引入的外部信息转化成向量的责任，因此编码层需要更加准确的将自然文本中传递的特征抽取出来，在向量中进行表示。

以编码器为中心的优化方法主要分为两大类，一类是嵌入级优化方法，这类方法可以优化嵌入，使嵌入具有更强的语义，但这类方法面向更加上游的任务，可能需要更多的语言学知识，如 CoType 模型，针对关系抽取任务中实体和关系的语言学特征，提出了三个假设，从而获得更好的嵌入表示。

因此更多的研究学者选择使用第二类方法：编码器级优化方法，这类方法主要对编码器的框架结构进行优化，以改进最后的分类效果，使用的深度学习技术包括了卷积神经网络、循环神经网络、注意力机制、图卷积网络。其中卷积神经网络具有较好的特征提取能力，由于可以并行计算，计算速度也较快，但卷积神经网络在结构上决定了其仅考虑当前输入，因此无法解决句子中的长距离依赖问题；而循环神经网络考虑当前输入以及上一步输入，具有记忆功能，因此能够有效解决长距离依赖问题，但特征提取能力较弱，计算速度也较慢。

而图卷积神经网络和注意力机制都可以与卷积神经网络、循环神经网络相结合，优化编码层。图卷积神经网络对句子中类似图的特征进行编码，保留图的特性，并最终提高句子表征的表达力。注意力机制则可以通过权重的分配对关键信息进行强调：通过对句子不同部分进行注意力的计算，最后得出更优的句子表征，如 BGWA 模型等；或者在使用多示例学习的基础上对同一个包中不同句

子进行注意力的设计，得出更全面的包表征，如 APCNN 模型等。

3.4 以分类器为中心的降噪方法

分类器的输入为经过编码层进行特征提取后的表征，在关系抽取任务中，分类器的输出往往为预测关系的概率分布，从而得到最终预测的关系。在本节中，我们将先介绍 DSRE 任务中的基础分类器，再根据对分类器优化的侧重点分为基于分类器训练的优化方法以及基于关系得分的优化方法。

3.4.1 基础分类器

在 Vanilla 模型中，分类器是一个多类逻辑回归分类器，直接从带有噪声的特征中学习权重。而 Riedel 等人[11]则使用因子图模型（Factor Graph）进行推断，预测实体间的关系，以及哪个句子表达了这个关系，如图 12 所示。其中变量Y代表两个实体之间的关系，变量$Z_i \in [0,1]$代表第 *i* 个指代（即第 *i* 个句子）是否表达了关系 *Y*。

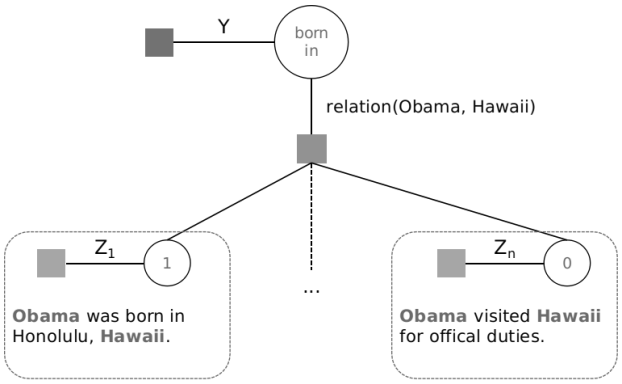

图 12 因子图模型示意图

在 PCNN 中，为了减少大量噪声数据的影响，选择使用 At-Least-One 假设（详见第 3.1.1 节）。假设$M = M_1, M_2, \ldots, M_T$表示训练数据中的 T 个包，每一个包有一个关系标签，$M_i = \{m_i^1, m_i^2, \ldots, m_i^n\}$表示第 *i* 个包中有 *n* 个示例。每一个包只选择一个示例来代表整个包，而 *o* 表示针对包M_i时模型的输出（未经过 softmax），o_r表示为第 *r* 个关系的得分。经过 softmax 就可以计算M_i包属于每一个类别的概率（m_i^j为该包被选择的示例）：

$$p(r|m_i^j;\theta) = \frac{e^{o_r}}{\sum_{k=1}^{n_1} e_{o_k}}$$

在多示例学习中，主要目的是得到每一个包的标签，并不会关注包里面的示例，因此，需要定义包级别的损失函数。根据 At-Least-One 的假设，每一个包中至少会有一个标注正确的示例。因此可以从每个包中找一个得分最高的示例来代表整一个包，因此，定义基于包的目标函数如下：

$$J(\theta) = \sum_{i=1}^{T} log\, p(y_i|m_i^j;\theta)$$

其中，$j* = argmax_j p(y_i|m_i^j;\theta)$，$1 \le j \le n$，即只选择得分最高的示例来代表整一个包。

Lin 等人[31]通过 APCNN 模型，优化了包S的表征s，在分类阶段，根据向量的相似度计算，将包的表征s与关系r的相似度o_r作为得分，最后使用 softmax 来归一化概率：

$$p(r|S,\theta) = \frac{\exp(o_r)}{\sum_{k=1}^{n_r} exp(o_k)}$$

其中，$o = Ms + d$，M为所有关系的嵌入矩阵。而在测试阶段，由于没有包的关系标签 *r* 未知，因此不能按照训练阶段的方法直接得到包的表征，Lin 等人对每一个关系r_i都计算了 Attention 权重，求出相应的包在此关系下的表征s_i以及关系得分o_i，选择o_i中关系r_i对应的值作为r_i的预测概率，最后将获得最大概率的关系作为预测的关系。这个模型最后在 Riedel2010 数据集上超越了原始的 PCNN 模型，获得了更好的结果。

为了更好提高分类器的性能，研究学者们在基础分类器的基础上进行了不同程度的改进，分为基于分类器训练的优化方法以及基于关系得分的优

化方法。

3.4.2 基于分类器训练的优化方法

基于强化学习。 Feng 等人[45]提出的 RLRE 模型，使用了强化学习的方法训练一个示例选择器（Instance Selector）来进行句子中高质量示例的挑选，将每个包中挑选出来的示例用来对分类器的训练。考虑到重叠关系的存在，他们在进行关系分类器训练的时候进行的是句子级别的关系分类。由于对于示例选择器没有一个显性的监督，即无法直接判断所选出来的句子质量高低，但可以将所选择的示例用以关系分类任务，以结果的概率分布作为 reward 来进行示例选择器的学习。

Takanobu 等人[44]也将强化学习使用到 DSRE 任务中，提出了 HRL 模型，该模型使用分层的 Agent 来进行关系实体的联合抽取：对于同一个文本，通过两层强化学习来实现整个关系抽取，其中高级别的强化学习 Agent 识别关系，如果所识别的关系不为 NA 时，则触发低级别的强化学习，进行实体标注子任务，当实体被识别时，子任务完成，高级别的强化学习 Agent 则继续扫描句子的其余部分，寻找其他关系。

基于迁移学习。为了进一步优化模型，减少模型学习时间，Liu 等人[35]将迁移学习应用到分类器训练的过程中，提出了+STP+EWA+TL 模型，考虑到实体类别的信息对关系类别有重要的提示作用，他们选择了实体类别标注任务作为预训练任务，得到共享模型参数θ_0以及主体实体和客体实体类别分类的私有参数$\theta_{subject}$、θ_{object}。最后由共享模型参数和关系抽取任务的私有参数确定关系抽取任务的损失函数。

最终实验证明，增加了句法依存树、以及多层注意力和迁移学习以后，在包中选择同等数量的示例的前提下，在 Riedel2010 数据集上的准确率比 APCNN 高了 7.7%。

基于动态标签。 Liu 等人[46]提出的+SL 模型中的软标签（Soft Label，SL）的获取过程也需要对数据进行预关系抽取。SL 的获得过程是一个动态获取过程，因此相同的包在不同的训练阶段可能会拥有不同的 SL，每一个实体对$s_i(< h_i, t_i >)$的软标签r_i可以通过以下公式计算：

$$r_i = arg\,max(o + max(o)A \odot L_i)$$

其中，L_i为$< h_i, t_i >$的关系类别的独热编码，A为远程监督标签的置信度矩阵，o则为关系得分。

类似的获取动态标签的还有 Sun 等人[47]，他们将强化学习和动态标签结合起来提出了 EPNet 模型，主要分为两个模块：负责进行关系抽取的抽取网络（Extraction Network，ENet）以及负责进行隐式标签获取的策略网络（Policy Network，PNet）。首先，使用 APCNN 模型得到句子和包的表征，然后对其进行关系分类，而 ENet 的分类结果将作为使用了强化学习的 PNet 的状态，然后在 PNet 中对这些状态进行相应的行为，计算出新的隐式标签，得到的隐式标签将作为 ENet 的输入，进行一轮的关系抽取后，将结果作为 PNet 的延迟反馈，并推动 PNet 的参数更新。以上两种获取动态标签的方法可以对错误的标签进行一定程度的校准，减少噪声对模型性能的影响。

3.4.3 基于关系得分的优化方法

基于噪声建模。2017 年，Luo 等人[43]为了对噪声建模，提出了+TM 模型，该模型可以构建每一个句子的转移矩阵，并最终计算出关系的分布，模型如图 14 所示。他们分别进行了句子级别的建模以及包级别的建模，当进行句子级别的建模时，每个句子都使用 PCNN 模型进行编码，最后将得到的向量输入到全连接层（Full Connection Layer），最后进行 softmax 计算出预测的关系分布 p。然后进行噪声的建模：先获取每个句子的嵌入x_n，然后使用 softmax 计算每个句子的转移矩阵 T。矩阵中的T_{ij}表达了被远程监督标记为关系 j 的句子的真实关系其实是 i 的条件概率，这样一来就可以获得每个句子的噪声模式。最后将 T 与前面的预测分布 p 相乘，获得该句子的关系分布 o：

$$o = T^T \cdot p$$

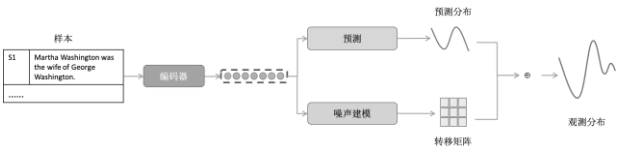

图 13 基于噪声建模进行关系抽取的工作流

包级别的建模与句子级别的建模类似，使用 APCNN 中的句子级别注意力得到包的表征s_i，最后

计算每个包的转移矩阵T_{ij}。

值得一提的是，在训练过程中，Luo 等人还借鉴了课程学习（Curriculum Learning）的思想，这是 Bengio 教授团队在 2009 年的 ICML 会议上提出的概念[69]，主要思想是模仿人类学习的特点，由简单到困难来学习课程，这样容易使模型找到更好的局部最优，同时加快训练的速度。

基于多示例多标签。At-Least-One 模型解决了多示例单标签的问题，但同时也假设了包中的每一个示例共享包的唯一标签，因此忽略了重叠关系的情况，即一对实体对可以存在多种关系的情况，但在实际场景中，重叠关系的情况是不可避免的，例如实体对（Jobs，Apple）会同时具有 Founded 和 CEO-of 的关系。在这种情况下，Hoffmann 等人[20]提出了基于概率的解决重叠关系的 MultiR 模型。类似的，Surdeanu 等人[21]将多示例多标签学习用于远程监督关系抽取任务中，提出了 MIML 模型，该模型允许同一个实体对在不同句子中的指代可以拥有不同的标签，如图 13 所示。

在 2016 年，Jiang 等人[42]为了解决重叠关系的问题，提出了 MIML-RE 模型：在 PCNN 的基础上，分类器选择了 sigmoid 函数来计算每个关系的概率，因此当关系的概率超过某个阈值时，就认为这个包表达了这个关系。

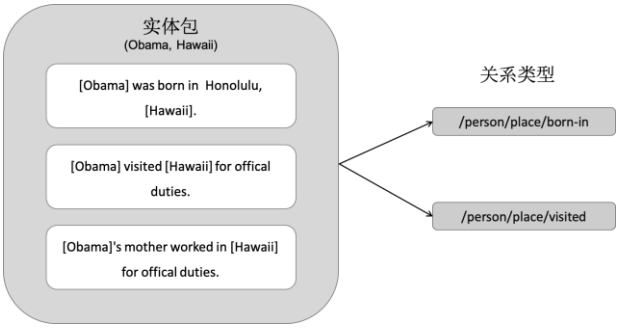

图 14 多示例多标签示例

3.4.4 分析比较

以分类器为中心的降噪方法是指以优化分类器为主要目的的方法，主要分为两个研究方向：基于分类器训练的优化方法以及基于关系得分的优化方法。

其中基于分类器训练的优化方法使用了强化学习、迁移学习和动态标签等。其中强化学习在训练分类器过程中引入反馈（reward），通过设计不同的行动（action）来实现噪声数据的识别或者直接进行关系抽取，提高了分类器的性能，但强化学习的方法往往受限于设计一个好的奖励函数来引导智能体采取正确的行为。基于迁移学习的方法可以有效的减少模型的训练时间，提高训练效率。基于动态标签的方法可以通过过滤掉错误的远程监督标签来提高关系抽取任务的性能。

基础分类器分为两个步骤：首先为对关系得分 o 的计算；然后为对关系得分 o 进行归一化。因此以分类器为中心的降噪方法的第二类方法则是基于关系得分的优化方法，通过对基础分类器的两个步骤进行不同程度的改进，其中基于噪声建模的 +TM 模型在计算关系得分时加入了噪声矩阵，提高了得分的准确度，而 MIML-RE 模型对归一化部分进行改进，使分类器可以更好的处理重叠关系的问题。

4 长尾问题解决方案

目前大部分远程监督关系抽取的改进方法都以解决 WL 问题为主要目的（见第三节），这些方法在示例数量多的关系上取得了准确率等性能的提升，然而并没有对那些长尾数据进行更加深入的研究。而无论在真实应用场景下，还是远程监督获得的数据集分布上看，长尾数据的问题是真实存在的，Riedel 等人[11]发现只有少数的纽约时报语料集的同一句话中提到了两个在 Freebase 中的实体。Gui 等人[70]通过分析 Freebase 中的 480 个关系，发现其中只有 87 个关系在纽约时报语料集中获得一个或以上示例，同时，由于长尾关系有用的示例数量少，更加可能发生一整个包都是噪声数据的情况，从而影响整个关系抽取任务的性能。

Gui 等人[70]认为处理长尾数据的挑战之一在于训练样例的缺乏，不利于构建有效的提取器。为了解决这个问题，他们参考解释学习（Explanation-Based Learning，EBL），提出了一种处理长尾数据的方法，该方法可以使用未标记的数据有效地学习关系提取规则并产生可解释的结果。与统计学习不同，EBL 不限于训练样本的理论界限，并且在有限的数据上表现良好。

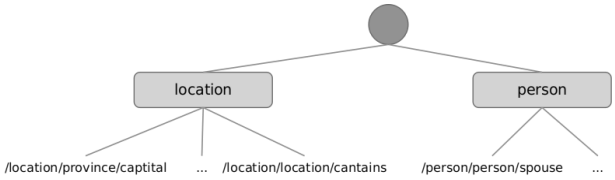

图 15 层次结构示意图

2018 年，Han 等人[38]在 APCNN 模型的基础上，将知识图谱中的层次结构，即图 15 所示的本体中的关系结构信息，应用到DSRE任务中，提出+HATT 模型。该模型使用从粗粒度到细粒度的注意力机制，来挑选出有效示例。他们认为在底层的关系（如图中的/person/person/spouse 等）可以提供该关系的更多特定的特征，例如句子中很可能出现 spouse 等关键词，而在更高层的关系中（如图中的 person））则可以捕捉到其子关系的共同特征。在底层关系中，容易受到数据稀疏的影响，性能并不稳定，可以使用细粒度的注意力做示例选择，而高层关系中，在示例选择上更具有鲁棒性，使用粗粒度的注意力。在训练过程中，由于利用到了其他类似关系的信息，不再是针对单个关系进行分类器的训练，因此长尾关系的示例对分类器的影响有所降低，对 DSRE 中的长尾问题提供了一个有效的解决方案。

2019 年，Zhang 等人[39]也将关系结构用于解决 DSRE 中的长尾问题上，提出了+KATT 模型，模型结构如图 16 所示。在示例编码层使用了 CNN 和 PCNN 分别进行编码实验，得到每一个包的表征 $S_{h,t}=\{s_1,s_2,\ldots,s_m\}$，同时，进行隐式的关系信息抽取：初始化关系的层次结构，使用 TransE[68]等知识表示方法得到的预训练知识图谱向量来表示结构中的每个节点，得到 $v_i^{implicit}$，其中父节点的向量为子节点向量的均值，然后使用 GCN 对关系的层次结构进行编码，学习其显式的关系信息，得到每个节点的显式嵌入 $v_i^{explicit}$，然后将 $v_i^{implicit}$ 和 $v_i^{explicit}$ 进行拼接，得到关系在层次结构中的向量表示 q_r。最后将包中第 k 个句子的表征 s_k 与 q_r 拼接后计算注意力权重，可得到包 $S_{h,t}$ 在层次结构中每一层的文本关系表征 $r_{h,t}^i, i\in\{0,1,\ldots,L-1\}$。最后考虑到对于不同的实体对而言，不同的层会有不同的贡献，因此+KATT 模型并没有直接将所有层的文本关系表征进行直接拼接，而在拼接时加上了注意力机制，来强调更重要的层:

$$r_{h,t}^i=\beta_i r_{h,t}^i$$

而 $r_{h,t}=Concat(r_{h,t}^0,\ldots,r_{h,t}^{L-1})$，最后将 $r_{h,t}$ 用于计算关系的条件概率。

按照第 3 节的分类方法可知，目前针对长尾问题的解决方案都使用了以编码层为中心的优化方法，并且都是基于注意力机制进行优化。这是考虑到长尾关系中的示例数量太少，甚至仅有的示例可能都是错的，如果对每一个关系都单独训练一个分类器，则针对长尾关系的分类器就没有足够的有效示例支撑，性能也随之无法保证，因此 HATT、KATT 模型都是在基于关系的层次结构对注意力机制进行改进，从知识图谱关系层次中获取其他相关关系的信息，为长尾关系的分类器的训练提供额外的支撑。

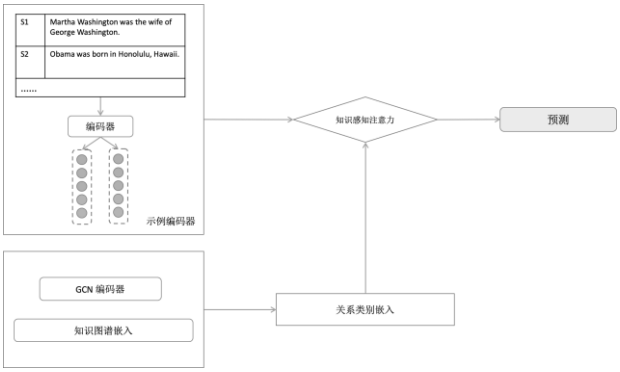

图 16 +KATT 模型示意图

5 数据集与评估

5.1 常用数据集

在远程监督领域，研究学者经常选择开源的大型语料集以知识图谱进行数据生成，其中比较常用的就维基页面、纽约时报（New York Times Corpus, NYT）[71]等语料集以及 Freebase、WikiData、DBPedia 等知识图谱，从而完成实验数据集的构建。

Riedel2010 数据集。目前比较常用的远程监督关系抽取数据集是 Riedel 等人[11]在 At-Least-One 假设下提出的 Riedel2010 数据集，该数据集由通过使用 Freebase 作为对齐所用知识图谱，对 NYT2010 的语料集进行远程监督标注得到，其中训练集中有 522,611 个句子，281,270 对实体，18,252 个关系事实；测试集中包含 172,228 个句子，96,678 对实体，1,950 个关系事实，总计 39,528 对实体以及 53 种关系。

Mintz2009 数据集。Mintz 等人[3]使用 FreeBase 知识图谱来标记维基页面的语料集所生成的

Mintz2009 数据集。

GIDS 数据集。 Jat 等人[37]使用人工标注的 Google 关系抽取语料集对网页信息进行标记得到的数据集，其中包含了 perGraduatedInstitution、perHasDegree、perPlaceOfBirth、perPlaceOfDeath 四种关系以及 NA 关系，示例分布情况如表 4 所示。

表 4 GIDS 数据集

关系类别	*示例数*	*实体对数*
perGraduatedInstitution	4456	2624
perHasDegree	2969	1434
perPlaceOfBirth	3356	2159
perPlaceOfDeath	3469	1948
NA	4574	2667

TIMERE 数据集。 Luo 等人[43]为了提高数据集的可靠性，选择使用与 WIkiData 知识图谱中时间相关的相关的三元组来标记维基百科语料集，其中包含了 278,141 个句子，12 种与时间有关的关系。同时，TIMERE 数据集可以通过时间信息的完整程度来对数据的可靠程度划分等级，即像包含年月日这种详细时间信息的数据可靠程度会比只包含部分时间信息的数据的可靠程度高，划分可靠等级以后，可以获得 184,579 个负样本，77,777 个正样本，正样本中包含了 22,214 个可靠样本，只包含年月的不可靠样本 2,094 个，只包含年的不可靠样本 53,469 个。

中文互动百科数据集。潘云等人[72] 使用在线资源互动百科构造人物关系知识库 HDKB，得到 18 种人物关系类型，随后使用 HDKB 对多个语料集（SogouC 语料集、sohu 新闻语料集和百度百科语料集）进行标记，其中不同的语料集标记的情况如表 5 所示。

表 5 中文互动百科数据集

语料集	*匹配成功的关系类别数*	*文本数*
SogouC	12	80000
Sohu 新闻	3	2560
百度百科	13	约 130000

5.2 评测指标与评估方式

5.2.1 评测指标

常见指标。 使用远程监督的目的是为了在减少人工参与的条件下，提供关系抽取模型训练所需的数据集，因此在进行远程监督的性能评估时，可以使用关系抽取的常用的评价指标进行评估，可以使用**准确率 P**，**召回率 R**，以及 **F1 值**三种评测指标。他们的计算公式如下所示：

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2PR}{P + R}$$

其中，*TP* 为真正例，*FN* 为假反例，*FP* 为假正例，*TN* 为真反例。通常为了直观形象的呈现实验结果，研究人员还会使用 **PR 曲线**（Precision-Recall curve）或者 **AUC**（Area Under Curve）**值**来将自己的模型的与其他模型进行对比。

P@N。 针对使用了多示例学习的模型，可以使用 P@N 来作为模型的评估指标，指的是在每个包中选择特定数量的示例进行实验时，将得分最高的前 N 个示例进行留出法进行评估而得出的准确率。常用的设置为在每个包中分别随机选出一个（One），两个（Two），全部（All）示例进行训练，分别评估这三种情况下的 P@100、P@200、P@300。

Hits@K。 为了更好对长尾关系的实验结果进行评估，Han 等人[38]提出的一个评估指标，指从测试集中抽取出一个子集，在该测试子集中，针对每一个实体对，其得分最高的前 K 个预测标签中存在其真实标签的概率（Han 等人选择的 K 值为 10、15、20）。

5.2.2 评估方式

由于远程监督是使用已有知识图谱来对数据集进行自动标注，因此得到的数据集并不是完全可信的，因此在进行实验时，往往面临着一个巨大的挑战：用来辅助进行训练集标注的知识图谱往往是 WikiData、DBpedia 等开放域知识图谱，其中含有的关系非常多，以及是训练数据总数也较多，可能同时涉及各个领域的知识，因此研究人们常用留出法（Hold-out）或者人工评估（Manual Evaluation）的方法进行效果的评估。

留出法。留出法将数据集分为训练集和测试集，其中测试集和训练集互斥。这种方法常常被用来对机器学习的模型进行评估，但由于远程监督使用的知识图谱并不一定会完全标注数据集中的所有句子，因此测试集中的数据也不一定会全部被标注。在这种情况下，使用留出法进行模型评估会低估模型的真实效果。

人工评估。 使用人工评估的方法，则可以比较准确的评估模型的真实效果，但在数据集很大的情况下，人工评估方法需要耗费大量的人力和时间，因此往往适用于小数据量的评估。

除了以上两种评估方式以外，由于数据存在长尾现象，而数据量过少时，噪声占比相对较大，甚至一个关系中的示例可能全是噪声数据，这样的数据用来实验无法有效的比较出模型的优劣，因此也有部分研究学者选择常见的关系集或者示例数更多的关系集合作为实验数据，如 Qin 等人[51]在实验中选择了拥有示例数量大于 1500 的关系进行实验。

6 远程监督关系抽取任务研究展望

随着信息抽取在 NLP 下游任务中广泛应用，关系抽取任务作为信息抽取的一环也备受关注，其中远程监督关系抽取任务已经成为当前的研究热点。然而，由于研究时间比较短，远程监督方法在关系抽取任务中的应用尚不算成熟，学术界的方法没有得到落地应用的验证，许多关键问题值得深入探索研究。本文总结以下研究方向。

（1） 基于联合抽取减少 DSRE 的误差传递。正如图 2 所示，Vanilla 模型在使用远程监督方法进行数据集的构建时，会先进行实体抽取，随后进行关系抽取。这两项任务执行时相互独立，两个模型的参数也相互独立，但在真实数据中，实体信息往往可以缩小可能关系的范围，实体类型信息对关系的信息也有明显的提示作用，如 person 和 location 这类实体之间不可能是 contains 关系等，因此实体信息对关系抽取任务是有帮助的。联合抽取在传统的关系抽取任务中已经慢慢成为了关注的热点，联合抽取可以减少实体识别到关系抽取这个串行流水线之间的误差，约束有效关系集的大小，例如 2017 年 Zheng 等人[73]提出使用一种新颖的标注方式进行联合关系抽取。目前也有学者将联合抽取的思想用于 DSRE 任务的句子建模的过程中（CoType 模型），并取得了初步研究成果。未来可以尝试基于联合抽取的远程监督关系抽取的端到端模型，提高模型的准确率。

（2）基于小样本学习进行长尾关系的关系抽取。正如 2.3.2 节中的关键问题中上所说，远程监督生成的数据中还存在长尾问题，在非通用关系上的示例数量较少，导致模型在这些关系上无法做出可信的预测。尽管第 4 节中的模型对长尾关系的处理效果有一定的提升，但总体效果还是不能令人满意的。研究学者会针对这种样本少的训练集使用小样本学习进行模型的训练。这种学习方法早年主要集中在图像领域的应用，近年来，在自然语言处理领域也慢慢开始出现相应的数据集和模型[74]。在关系抽取模型方面，Gao 等人[75]将这种小样本学习应用于带噪声的关系抽取任务中，提出了 HATT-Pro 模型，Ye 等人[76]改进了原型网络（Prototypical Network）中对示例的编码方式，提出了多层匹配聚合网络（Multi-Level Matching and Aggregation Network，MLMAN）。而在数据集方面，也有 Han 等人[77]提出的关系抽取领域的小样本数据集 FewRel。未来，随着元学习以及深度学习的发展，可以尝试将小样本学习与远程监督有机融合，解决数据集长尾关系中的 Few-shot 甚至 Zero-shot 问题。

（3）基于混合监督对远程监督完全落地进行过渡。使用远程监督的方法无法避免的会产生噪声数据，这部分噪声数据会对模型的训练和性能有很大的影响。目前远程监督关系抽取的性能还不能够完全投放到工业界使用，真正落地的应用系统并不多，Tran 等人[25]将远程监督关系抽取应用到了医学领域，对医学领域的常用关系 treatment 关系进行识别，但也只能针对医学领域示例数量最多的 treatment 关系有较好的准确率。考虑到监督学习关系抽取任务也有相应的标注好的数据集，这些数据可以以一定的权重与远程监督生成的数据进行融合，共同进行关系抽取任务。已经有学者进行初步的尝试[78-82]，阿里巴巴的神马知识图谱团队也在构造神马知识图谱的时候使用 Deepdive 框架[83]进行关系抽取任务_，其中数据标注阶段包含了远程监督和基于启发性规则的标注。这些混合监督的方法既可以在一定程度上减少构建数据集的成本，也可以在一定程度上提高关系抽取任务的性能。

（4）基于语言模型提高句子编码器的性能。语言模型是自然语言处理领域中一个基本却又重要的任务，是 NLP 中多项任务的基础，如机器翻译、语言识别等，可以用来判断一句话符不符合我们的表达习惯。在 DSRE 的编码层，往往需要借用语言模型对自然文本进行建模，而近年来也不断有新的语言模型提出，包括 ELMO[84]、BERT[85]等。更强的语言模型，可以更好的对文本中的特征进行表

_ https://zhuanlan.zhihu.com/p/34535849

达，而将这些技术综合应用到 DSRE 中还有待深入研究。为了将远程监督关系抽取模型投入实际使用，发挥实际的价值，基于语言模型对编码层进行改进也是当前可以深入研究的方向。

7 总结

远程监督方法大大减少了构建数据集的成本，是从无标注文本中实现自动化关系抽取的关键一环，受到了学术界的广泛关注。深度学习技术的发展和知识图谱相关技术的发展，为远程监督关系抽取任务带来了机遇，但远程监督方法相应带来的噪声数据又给这个任务带来了挑战。本文对远程监督带来的错误标注问题以及长尾问题进行了阐述分析，并以基于深度学习的方法为重点，将解决方案分为样本降噪，外部信息融合，编码器优化方法以及分类器优化方法四类，按照分类梳理已有的研究成果，总结远程监督关系抽取任务的常用数据集与评测指标，尝试建立一个较为完整的领域研究视图，希望能对相关领域研究者提供帮助。

参 考 文 献

[1] Craven M., Kumlien J. Constructing biological knowledge bases by extracting information from text sources//Proceedings of the ISMB, Heidelberg, Germany, 1999: 77-86.

[2] Wu F., Weld D.S. Autonomously semantifying wikipedia//Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal, 2007: 41-50.

[3] Mintz M., Bills S., Snow R., Jurafsky D. Distant supervision for relation extraction without labeled data//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, Morristown, NJ, USA, 2009: 1003-1011.

[4] Go A., Bhayani R., Huang L. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009, 1(12): 2009.

[5] Plank B., Agić Ž. Distant supervision from disparate sources for low-resource part-of-speech tagging. arXiv:1808.09733, 2018.

[6] Qin L., Liu Y., Che W., Wen H., Liu T., End-to-end task-oriented dialogue system with distantly supervised knowledge base retriever. Chinese computational linguistics and natural language processing based on naturally annotated big data. Springer. 2018: 238-249.

[7] Lee S., Song Y., Choi M., Kim H. Bagging-based active learning model for named entity recognition with distant supervision//Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp), Hong Kong, China, 2016: 321-324.

[8] Roth B., Barth T., Wiegand M., Klakow D. A survey of noise reduction methods for distant supervision//Proceedings of the 2013 workshop on Automated knowledge base construction, San Francisco, USA, 2013: 73-78.

[9] Smirnova A., Cudré-Mauroux P. Relation extraction using distant supervision: A survey. ACM Computing Surveys (CSUR), 2018, 51(5): 106.

[10] Dumitrache A., Aroyo L., Welty C. False positive and cross-relation signals in distant supervision data. arXiv:1711.05186, 2017.

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

[12] Dietterich T.G., Lathrop R.H., Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence, 1997, 89(1-2): 31-71.

[13] Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation. Journal of machine learning research, 2003, 3(Jan): 993-1022.

[14] Yao L., Haghighi A., Riedel S., Mccallum A. Structured relation discovery using generative models//Proceedings of the empirical methods in natural language processing, Edinburgh, UK, 2011: 1456-1466.

[15] Huang Pei-Jing, He Liang, Yang Jing, Research on noise reduction in distant supervised personal relation

extraction. Computer Applications and Software, 2017, 34(7):11-18. (in Chinese)
(黄蓓静, 贺樑, 杨静. 远程监督人物关系抽取中的去噪研究. 计算机应用与软件, 2017, 34(7): 11-18.)

[16] Liang Y., Huang H., Cai Z., Hao Z., Tan K.C. Deep infrared pedestrian classification based on automatic image matting. Applied Soft Computing, 2019, 77(484-496.

[17] Krizhevsky A., Sutskever I., Hinton G.E. Imagenet classification with deep convolutional neural networks//Proceedings of the Advances in neural information processing systems, Nevada, USA, 2012: 1097-1105.

[18] Hinton G., Deng L., Yu D., Dahl G., Mohamed A.-r., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Kingsbury B. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal processing magazine, 2012, 29(6):82-97.

[19] Bengio Y., Ducharme R., Vincent P., Jauvin C. A neural probabilistic language model. Journal of machine learning research, 2003, 3(Feb): 1137-1155.

[20] Hoffmann R., Zhang C., Ling X., Zettlemoyer L., Weld D.S. Knowledge-based weak supervision for information extraction of overlapping relations//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, Oregon, USA, 2011: 541-550.

[21] Surdeanu M., Tibshirani J., Nallapati R., Manning C.D. Multi-instance multi-label learning for relation extraction//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, Jeju Island, Korea, 2012: 455-465.

[22] Zeng D., Liu K., Chen Y., Zhao J. Distant supervision for relation extraction via piecewise convolutional neural networks//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015: 1753-1762.

[23] Takamatsu S., Sato I., Nakagawa H. Reducing wrong labels in distant supervision for relation extraction//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, Jeju Island, Korea, 2012: 721-729.

[24] Wu Y., Bamman D., Russell S. Adversarial training for relation extraction//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017: 1778-1783.

[25] Tran T., Kavuluru R. Distant supervision for treatment relation extraction by leveraging mesh subheadings. Artificial Intelligence in Medicine, 2019, 98:18-26.

[26] Han X., Liu Z., Sun M. Denoising distant supervision for relation extraction via instance-level adversarial training. arXiv:1805.10959, 2018.

[27] Qin P., Xu W., Wang W.Y. Robust distant supervision relation extraction via deep reinforcement learning. arXiv:1805.09927, 2018.

[28] Ji G.L., Liu K., He S.Z., Zhao J. Distant supervision for relation extraction with sentence-level attention and entity descriptions. Thirty-First Aaai Conference on Artificial Intelligence, 2017, 3060-3066.

[29] Vashishth S., Joshi R., Prayaga S.S., Bhattacharyya C., Talukdar P. Reside: Improving distantly-supervised neural relation extraction using side information. arXiv:1812.04361, 2018.

[30] Su Y., Liu H., Yavuz S., Gür I., Sun H., Yan X. Global relation embedding for relation extraction//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, USA, 2018: 820-830.

[31] Lin Y., Shen S., Liu Z., Luan H., Sun M. Neural relation extraction with selective attention over instances//Proceedings of the meeting of the association for computational linguistics, Berlin, Germany, 2016: 2124-2133.

[32] Ren X., Wu Z., He W., Qu M., Voss C.R., Ji H., Abdelzaher T., Han J. Cotype: Joint extraction of typed entities and relations with knowledge bases. the web conference, 2017, 1015-1024.

[33] Ling X., Weld D.S. Fine-grained entity recognition//Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto,

Ontario, Canada, 2012: 94-100.

[34] Pyysalo S., Ginter F., Heimonen J., Bjorne J., Boberg J., Jarvinen J., Salakoski T. Bioinfer: A corpus for information extraction in the biomedical domain. BMC Bioinformatics, 2007, 8(1): 50-50.

[35] Liu T., Zhang X., Zhou W., Jia W. Neural relation extraction via inner-sentence noise reduction and transfer learning. arXiv :1808.06738, 2018.

[36] Zhou P., Xu J., Qi Z., Bao H., Chen Z., Xu B. Distant supervision for relation extraction with hierarchical selective attention. Neural Networks, 2018, 108:240-247.

[37] Jat S., Khandelwal S., Talukdar P.P. Improving distantly supervised relation extraction using word and entity based attention. arXiv:1804.06987, 2018.

[38] Han X., Yu P., Liu Z., Sun M., Li P. Hierarchical relation extraction with coarse-to-fine grained attention//Proceedings of the empirical methods in natural language processing, Brussels, Belgium, 2018: 2236-2245.

[39] Zhang N., Deng S., Sun Z., Wang G., Chen X., Zhang W., Chen H. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks//Proceedings of the north american chapter of the association for computational linguistics, Minneapolis, Minnesota, USA, 2019: 3016-3025.

[40] Yuan Y., Liu L., Tang S., Zhang Z., Zhuang Y., Pu S., Wu F., Ren X. Cross-relation cross-bag attention for distantly-supervised relation extraction//Proceedings of the national conference on artificial intelligence, Honolulu, Hawaii, USA, 2019: 419-426.

[41] Ye Z., Ling Z. Distant supervision relation extraction with intra-bag and inter-bag attentions//Proceedings of the north american chapter of the association for computational linguistics, Minneapolis, Minnesota,USA, 2019: 2810-2819.

[42] Jiang X., Wang Q., Li P., Wang B. Relation extraction with multi-instance multi-label convolutional neural networks//Proceedings of the international conference on computational linguistics, Osaka, Japan, 2016: 1471-1480.

[43] Luo B., Feng Y., Wang Z., Zhu Z., Huang S., Yan R., Zhao D. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017: 430-439.

[44] Takanobu R., Zhang T., Liu J., Huang M. A hierarchical framework for relation extraction with reinforcement learning//Proceedings of the national conference on artificial intelligence, Honolulu, Hawaii, USA, 2019: 7072-7079.

[45] Feng J., Huang M., Zhao L., Yang Y., Zhu X. Reinforcement learning for relation classification from noisy data//Proceedings of the national conference on artificial intelligence, New Orleans, Louisiana, USA, 2018: 5779-5786.

[46] Liu T., Wang K., Chang B., Sui Z. A soft-label method for noise-tolerant distantly supervised relation extraction//Proceedings of the empirical methods in natural language processing, Copenhagen, Denmark, 2017: 1790-1795.

[47] Sun T., Zhang C., Ji Y., Hu Z. Reinforcement learning for distantly supervised relation extraction. IEEE Access, 2019, 7:98023-98033.

[48] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative adversarial nets//Proceedings of the Advances in neural information processing systems, Cambridge, MA, USA, 2014: 2672-2680.

[49] Miyato T., Dai A.M., Goodfellow I. Adversarial training methods for semi-supervised text classification. arXiv:1605.07725, 2016.

[50] Liu A., Soderland S., Bragg J., Lin C.H., Ling X., Weld D.S. Effective crowd annotation for relation extraction//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, USA, 2016: 897-906.

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

[52] Zeng D., Liu K., Lai S., Zhou G., Zhao J. Relation classification via convolutional deep neural network. //Proceedings of COLING 2014, the 25th

International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 2014, 2335-2344.

[53] Ji G., He S., Xu L., Liu K., Zhao J. Knowledge graph embedding via dynamic mapping matrix//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), Beijing, China, 2015: 687-696.

[54] Pavlick E., Rastogi P., Ganitkevitch J., Van Durme B., Callisonburch C. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification//Proceedings of the international joint conference on natural language processing, Beijing, China, 2015: 425-430.

[55] Pennington J., Socher R., Manning C. Glove: Global vectors for word representation//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 2014: 1532-1543.

[56] Mikolov T., Chen K., Corrado G.S., Dean J. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.

[57] Li Q., Ji H. Incremental joint extraction of entity mentions and relations//Proceedings of the meeting of the association for computational linguistics, Baltimore, Maryland, 2014: 402-412.

[58] Riedel S., Yao L., Mccallum A., Marlin B.M. Relation extraction with matrix factorization and universal schemas//Proceedings of the north american chapter of the association for computational linguistics, Atlanta, Georgia, USA, 2013: 74-84.

[59] Sutskever I., Vinyals O., Le Q.V. Sequence to sequence learning with neural networks//Proceedings of the neural information processing systems, Montreal, Canada, 2014: 3104-3112.

[60] Hochreiter S., Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780.

[61] Cho K., Van Merrienboer B., Bahdanau D., Bengio Y. On the properties of neural machine translation: Encoder--decoder approaches//Proceedings of the empirical methods in natural language processing, Doha, Qatar, 2014: 103-111.

[62] Miwa M., Bansal M. End-to-end relation extraction using lstms on sequences and tree structures//Proceedings of the meeting of the association for computational linguistics, Berlin, Germany, 2016: 1105-1116.

[63] Zhou P., Shi W., Tian J., Qi Z., Li B., Hao H., Xu B. Attention-based bidirectional long short-term memory networks for relation classification//Proceedings of the meeting of the association for computational linguistics, Berlin, Germany, 2016: 207-212.

[64] Yan D., Hu B. Shared representation generator for relation extraction with piecewise-lstm convolutional neural networks. IEEE Access, 2019, 7:31672-31680.

[65] Marcheggiani D., Titov I. Encoding sentences with graph convolutional networks for semantic role labeling//Proceedings of the empirical methods in natural language processing, Copenhagen, Denmark, 2017: 1506-1515.

[66] Du J., Han J., Way A., Wan D. Multi-level structured self-attentions for distantly supervised relation extraction//Proceedings of the empirical methods in natural language processing, Brussels, Belgium, 2018: 2216-2225.

[67] Lin Z., Feng M., Santos C.N.D., Yu M., Xiang B., Zhou B., Bengio Y. A structured self-attentive sentence embedding//Proceedings of the international conference on learning representations, Toulon, France, 2017.

[68] Shen Y., Huang X. Attention-based convolutional neural network for semantic relation extraction//Proceedings of the international conference on computational linguistics, Osaka, Japan, 2016: 2526-2536.

[69] Bengio Y., Louradour J., Collobert R., Weston J. Curriculum learning//Proceedings of the international conference on machine learning, Montreal, Canada, 2009: 41-48.

[70] Gui Y., Qian L., Man Z., Gao Z. Exploring long tail data in distantly supervised relation extraction. //Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 514-522.

[71] Sandhaus E. The new york times annotated corpus.

Linguistic Data Consortium, Philadelphia, 2008, 6(12): e26752.

[72] Pan Yun, Bulebulihan·Yishabay, Yang Jing, Yin Min, Distant Supervised Personal Relation Extraction Using Chinese Online Resource, Journal of Chinese Computer Systems, 2015, 36(4): 701-706.(in Chinese)
（潘云, 布勒布丽汗·伊沙巴依, 杨静, 尹敏, 利用中文在线资源的远程监督人物关系抽取. 小型微型计算机系统, 2015, 36(4): 701-706.）

[73] Zheng S., Wang F., Bao H., Hao Y., Zhou P., Xu B. Joint extraction of entities and relations based on a novel tagging scheme//Proceedings of the meeting of the association for computational linguistics, Vancouver, Canada, 2017: 1227-1236.

[74] Xiong W., Yu M., Chang S., Guo X., Wang W.Y. One-shot relational learning for knowledge graphs//Proceedings of the empirical methods in natural language processing, Brussels, Belgium, 2018: 1980-1990.

[75] Gao T., Han X., Liu Z., Sun M. Hybrid attention-based prototypical networks for noisy few-shot relation classification//Proceedings of the national conference on artificial intelligence, Honolulu, Hawaii, USA, 2019: 6407-6414.

[76] Ye Z., Ling Z. Multi-level matching and aggregation network for few-shot relation classification//Proceedings of the meeting of the association for computational linguistics, Florence, Italy, 2019: 2872-2881.

[77] Han X., Zhu H., Yu P., Wang Z., Yao Y., Liu Z., Sun M. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation//Proceedings of the empirical methods in natural language processing, Brussels, Belgium, 2018: 4803-4809.

[78] Yu Xiao-Kang, Chen Ling, Guo Jing, Cai Ya-Ya, Wu Yong, Wang Jing-Chang, Relation Extraction Method Combining Clause Level Distant Supervision and Semi-supervised Ensemble Learning. Pattern Recognition and Artificial Intelligence, 2017, 30(1) : 54-63.(in Chinese)
（余小康, 陈岭, 郭敬, 蔡雅雅, 吴勇, 王敬昌. 结合从句级远程监督与半监督集成学习的关系抽取方法. 模式识别与人工智能, 2017, 30(1): 54-63.）

[79] Nguyen T.T., Moschitti A. Joint distant and direct supervision for relation extraction//Proceedings of the international joint conference on natural language processing, Chiang Mai, Thailand, 2011: 732-740.

[80] Pershina M., Min B., Xu W., Grishman R. Infusion of labeled data into distant supervision for relation extraction//Proceedings of the meeting of the association for computational linguistics, Baltimore, Maryland, 2014: 732-738.

[81] Angeli G., Tibshirani J., Wu J., Manning C.D. Combining distant and partial supervision for relation extraction//Proceedings of the empirical methods in natural language processing, Doha, Qatar, 2014: 1556-1567.

[82] Beltagy I., Lo K., Ammar W. Combining distant and direct supervision for neural relation extraction//Proceedings of the north american chapter of the association for computational linguistics, Minneapolis, Minnesota, USA, 2019: 1858-1867.

[83] Zhang C. Deepdive: A data management system for automatic knowledge base construction. University of Wisconsin-Madison, Madison, Wisconsin, 2015.

[84] Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. Deep contextualized word representations//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, USA, 2018: 2227-2237.

[85] Devlin J., Chang M., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the north american chapter of the association for computational linguistics, Minneapolis, Minnesota, USA, 2019: 4171-4186.

YANG Sui-Zhu, born in 1995，M.S. candidate. Her research interests include knowledge graph and distant supervision.

LIU Yan-Xia, born in 1979, Ph.D., associate professor. Her research interests include knowledge graph and neural networks.

ZHANG Kai-Wen, born in 1997, M.S. candidate. His research interest is knowledge graph.

HONG Yin, born in 1995, M.S. candidate. His research interest is end-to-end relation extraction.

HANG Han, born in 1980, Ph.D., professor. His research interests include intelligent algorithms and evolutionary computing.

Background

With the rapid development of knowledge graph, as an important subtask of the construction process, relation extraction is earning more and more popularity, which is also one of the crucial parts of information extraction. However, the supervised relation extraction methods require human-annotated datasets, which is high-cost and time-consuming. With such situation, the task called distantly-supervised relation extraction (DSRE) was proposed, being able to construct the dataset for relation extraction task with existing knowledge graphs and corpus, laying the foundation for automatic relation extraction. Nevertheless, the workflow of DSRE is not perfect, for the reason that it is accompanied by two major problem, wrong label (WL) and long tail (LT) distribution. The problem hinders the improvement of relation extraction, widely concerned by academic and business. At the same time, deep learning is earning more and more attention because of its amazing performance in many fields, e.g. text classifier. Therefore, compared with the tradition machine learning methods, many researchers tend to apply deep learning to DSRE in order to reduce the impact of WL and LT. Fortunately the authors finally improve the performance of DSRE with some deep learning technologies, including convolutional neural network, attention mechanism, generative adversarial network and so on.

This paper summarizes the effort made by the authors not only in academic but also in business. At the first, we introduce the background of DSRE and the vanilla assumption. Then we analyze the major problem and give a brief review of the feature-based methods. Moreover, we classify the existing solutions according to different modules, and discuss their advantages and disadvantages. What's more, we introduce the common datasets in detail, as well as summarize the evaluation methods and metrics. At the end, we look forward the trends of DSRE.

This work is supported by National Natural Science Foundation of China (No.61876208), Guangdong Province Key Area R&D Program (No.2018B010109003), Guangzhou science and technology project (No.201802010007, No.201804010276).