第3章

关系抽取

在互联网迅猛发展的情况下，绝大多数信息都可以形成电子文本，无处不在的信息，其数量正呈现爆炸式的增长，数据已深入生活的方方面面。如何快速高效地从开放领域的文本中抽取出有效信息，成为摆在人们面前的重要问题。

信息抽取包含3个关键技术：实体抽取、关系抽取、事件抽取。其中实体抽取是关系抽取和事件抽取的基础，旨在从文本中识别出人名、地名、机构名、日期、数额等实体信息。为了深入理解自然语言文本信息，要在实体识别的基础上，抽取出这些实体之间存在的语义关系。这项抽取实体间语义关系的任务，即关系抽取。实体间的关系可被形式化描述为关系三元组<Entity1，Relation，Entity2>，其中Entity1和Entity2是实体类型，Relation是关系描述。关系抽取即从自然语言文本中抽取出关系三元组<Entity1，Relation，Entity2>，从而提取文本信息。

对于命名实体的识别，相关工作已经趋于成熟，并应用于多个数据相关领域。但是，命名实体识别获取的信息不具有联络性，单一、离散且缺乏意义，这对于知识库构建和语义理解是远远不够的，于是，采用实体关系抽取技术使实体之间形成关系网络。由于文本的非结构性与不规则性，导致机器难以处理，需要通过特征表示的方式将其转换为机器可读的结构化数据，而命名实体识别可以将实体从无结构性的文本中独立出来，作为信息抽取过程中的一个关键特征。

由此可见，命名实体识别是信息抽取中最为表层的一步。研究者们希望可以探索这些独立的命名实体之间所具有的联系，例如，通过命名实体识别，发现机构名称在同一句子中出现的概率较大，希望找到机构之间的关联，例如，股东关系或包含关系等。进一步地，需要研究发现命名实体之间的关联，将离散的数据信息进行关联与整合，连接成关系网络。需要研究如何提取文本中可能存在的隐含着关系信息的实体、实体属性与文本结构，以训练具有实体关系分类能力的分类模型。关系抽取目标是研究正确的文本及文本关系信息，保证在不影响信息容量的前提下，提取结果具有较高的准确率和召回率(Recall)。作为数据信息挖掘中至关重要的一步，关系抽取技术已经成为领域内大量研究学者的研究重点[1]。

其研究成果主要应用在文本摘要、自动问答、机器翻译、语义网标注、知识图谱等。随着

近年来对信息抽取的兴起，实体关系抽取问题进一步得到广泛关注和深入研究，一些研究成果及时出现在近几年人工智能、自然语言处理等相关领域的国际会议上，如 ACL、EMNLP、ICLR、AAAI、KDD、NAACL、ECML-PKDD 等。目前，Google 和百度均在研究构建知识图谱技术，而抽取实体之间的语义关系是这个过程中关键的一环，对多个命名实体大规模建立信息联系，可促进知识库与知识图谱的构建。在海量信息处理中，通过关系抽取技术，可以通过从偏向自然语言的无结构信息中提取结构化的关系元组，协助机器处理自然语言并提高效率。另外，进一步挖掘文本信息的语义关系，可深入理解用户的搜索目的，用于协助搜索引擎提供更为精确的查询结果。所以，关系抽取不仅具有理论价值，更在多个场景中具有相当可观的应用价值。

经典的实体关系抽取方法主要分为有监督、半监督、无监督、远程监督 4 类。有监督的实体关系抽取主要分为基于特征和基于核函数的方法。有监督方法需要手工标注大量的训练数据，浪费时间精力，因此，人们继而提出了基于半监督、无监督、远程监督的关系抽取方法来解决人工标注语料问题。

经典方法存在特征提取误差传播问题，极大地影响实体关系抽取效果。随着近些年深度学习的崛起，学者们逐渐将深度学习应用到实体关系抽取任务中。基于数据集标注量级的差异，深度学习的实体关系抽取任务分为有监督、半监督、远程监督 3 类[1]。

3.1 实体关系抽取定义

实体关系抽取作为信息抽取的重要任务，是指在实体识别的基础上，从非结构化文本中抽取出预先定义的实体关系。实体对的关系可被形式化描述为关系三元组$<e_1,r,e_2>$，其中，e_1 和 e_2 是实体，r 属于目标关系集$R\{r_1,r_2,r_3,\cdots,r_i\}$。关系抽取的任务是从自然语言文本中抽取出关系三元组$<e_1,r,e_2>$，从而提取文本信息。

在关系抽取领域，实体是现实世界中的一个对象或对象集合。它分为三大类：命名实体(Named Entity，NE)、代词实体(Pronoun Entity，PE)和名词性实体(Nominal Entity，NoE)，例如，“China”和“Trump”是命名实体，“her”和“we”是代词实体，“the country”和“the man”是名词性实体。一般而言，我们提到的实体是命名实体，在文本预处理过程中都是对文本进行命名实体识别，其余的实体都通过指代消解或共指消解来识别。

在自然语言处理中，人们定义了一种介于两个或多个实体间的关系，即实体关系。其中，介于两个实体间的关系叫二元关系，3 个及 3 个以上实体间的关系叫作多元实体关系。从文本语义上来看，实体关系抽取可以分为明确性实体关系抽取和隐含性实体关系抽取。顾名思义，明确性实体关系抽取是文本的直接语义关系就是所要抽取的实体关系，隐含性实体关系抽取就是文本所表达的直接语义关系不是所要抽取的实体关系。其中，实体关系可以用数学表达方式表达如下：

$$R(e_1,e_2,\cdots,e_n)=c \tag{3.1}$$

其中，$e_1,e_2,\cdots,e_n$ 分别是关系实例中的实体，R 表示关系实例(Relation)，c 表示关系实例的关系类别(Category)，如 per：children(父母-孩子关系)，org：location(组织-位置关系)等。

在抽取实体关系时，需要预先给定一些关系类别，然后将文本指定到一个关系类别中。

实体关系抽取的应用非常广泛，在生物信息学领域，可以提取出蛋白质与疾病的关系，从而找到病因。它也是自动问答系统的技术支撑，它可以从问题中抽取出实体，从而在知识库中找到它们的关系。例如，关系实例“谁是苹果公司的 CEO?”，关系抽取就可以在系统信息库中找到结构化关系“CEO of(Apple Inc.)”，因此找到答案“库克”。实体关系抽取的广泛应用大大减少了所需耗费的人力和物力，同时也节约了大量时间成本。

鉴于实体关系抽取的巨大应用价值，越来越多的专家和学者开始研究实体关系抽取技术，希望找到可以提高关系抽取效率的方法，这进一步推动了实体关系抽取技术的发展。目前，实体关系抽取方法根据抽取原理可以分为基于规则的方法和基于机器学习的方法。

最初人们使用的关系抽取方法是基于规则的方法，该方法需要那些通晓语言学知识的专家根据抽取内容制定特殊的能够充分有效地描述文本内容的人工规则，这些规则包含了一些词汇、语法信息和语义信息等特征，然后在语料库中，系统需要根据这些已制定的规则查找符合规则的文本，最后就可以得到实体的语义类型。基于规则的方法需要专门的人力来制定规则，以使系统能够在特定领域抽取到可靠的关系。这种方法的实现非常烦琐，不但需要特定领域的专业人员，而且还需要耗费大量的人力、物力去完成整个抽取过程，所以付出的代价非常大，而且效果也不是非常理想。此外，制定的规则都是针对特定领域进行的，当移植到其他领域抽取关系时，得到的结果非常差，因此该方法不适用于大规模复杂数据库。

鉴于基于规则方法的上述缺点，基于机器学习的方法应运而生，基于机器学习的方法主要是基于数学领域中的统计学原理来实现，该方法领域限制不大，可移植性强，效率也很高，耗费的资源也较少，因此渐渐替代了基于规则的方法，成为目前应用最广泛和研究价值最大的关系抽取方法。

基于机器学习的关系抽取流程如图 3-1 所示，主要包括训练过程和预测过程，数据集分为训练集和测试集。训练过程的目的是用机器学习算法根据训练集得到一个模型，预测过程的目的是根据模型预测测试集文本的类别，实现对测试文本集的关系抽取。

基于机器学习的文本关系抽取主要包括 4 部分：文本的预处理、文本分析、关系表示和关系抽取模型。文本预处理的目的是将语料库中含有噪声的数据变为可被计算机处理的纯文本数据，这是因为互联网数据复杂多变，没有统一的格式，这些数据可能是 HTML、XML 等格式的数据，所以在进行关系抽取前需进行数据清洗，把语料数据中的网络标签去除掉，变为纯文本格式，以有利于文本特征的抽取；文本分析是从文本中选取出每个文本的特征，即文本的特征表示。从而将无结构的纯文本变为计算机可以识别处理结构化文本，主要的处理过程有分词、词性标注、命名实体识别、句法分析或依存分析等；关系表示是对关系实例的一种模式化表示，就是文本的模式表示，这是计算机处理自然语言的基础。由于如何能够清晰有效地表达出文本的语义信息和结构信息对后续的关系模式识别有重要影响，所以关系模式表示在关系抽取中起着重要的作用。关系抽取模型主要是基于关系表示的分类模型，通过预定义的关系类型，训练基于各种分类原理的分类器，最终实现对测试文本的关系预测。

基于机器学习的关系抽取方法按照有没有已标注好的文本集时的类别预测原理可以分为有监督学习、半监督学习和远程监督学习。本节将重点介绍这几类机器学习算法。

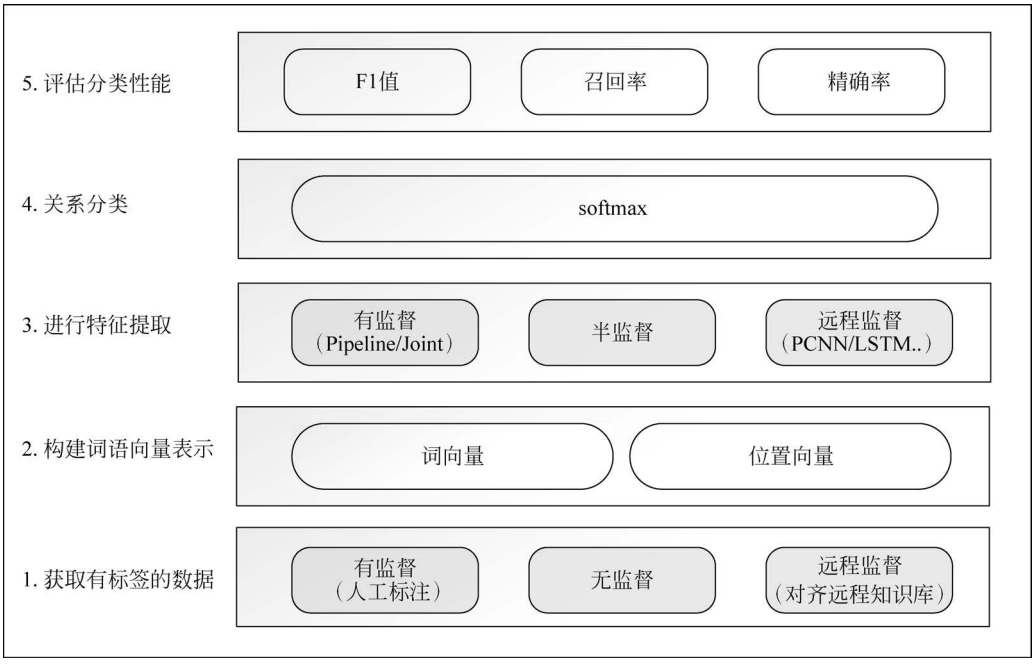

图 3-1 基于机器学习的关系抽取流程框架

3.2 实体关系抽取框架

针对实体关系抽取任务，基于深度学习的抽取过程如下。

（1）获取有标签的数据：有监督方法通过人工标记获取有标签数据集，远程监督方法通过自动对齐远程知识库获取有标签数据集。

（2）构建词语向量表示：将有标签句子分词，将每个词语编码成计算机可以接受的词向量，并求出每个词语与句子中实体对的相对位置，作为这个词语的位置向量，将词向量与位置向量组合作为这个词语的最终向量表示。

（3）进行特征提取：将句子中每一个词语的向量表示输入神经网络中，利用神经网络模型提取句子特征，进而训练一个特征提取器。

（4）关系分类：测试时根据预先定义好的关系种类，将特征提取出的向量放入非线性层进行分类，提取最终的实体对关系。

（5）评估分类性能：最后，对关系分类结果进行评估[2]。

3.3 评测方法

关系抽取领域有 3 项基本评价指标：精确率(Precision)、召回率和 F 值(F-score)。

（1）精确率是从查准率的角度对实体关系抽取效果进行评估，其计算公式为：

$$\text{Precision}_R = \frac{\text{被正确抽取的属于关系 } R \text{ 的实体对个数}}{\text{所有被抽取为关系 } R \text{ 的实体对个数}} \tag{3.2}$$

（2）召回率是从查全率的角度对抽取效果进行评估，其计算公式为：

$$\text{Recall}_R = \frac{\text{被正确抽取的属于关系 } R \text{ 的实体对个数}}{\text{实际应被抽取的属于关系 } R \text{ 的实体对个数}} \tag{3.3}$$

（3）F 值。对于关系抽取来说，精确率和召回率是相互影响的，二者存在互补关系，因此，F 值综合了精确率和召回率的信息，其计算公式为：

$$F_\beta = \frac{(\beta^2 + 1) \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

其中，β 是一个调节精确率与召回率比重的参数，实际测试中，一般认为精确率与召回率同等重要，因此，β 值一般设置成 1，即 F_1 值是 F 值的特殊形式。因此，式(3.4)可以表示为：

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.5}$$

3.4 有监督实体关系抽取方法

在有监督实体关系抽取中，解决实体关系抽取的方法可以分为流水线学习和联合学习两种：流水线学习方法是指在实体识别已经完成的基础上直接进行实体之间关系的抽取；联合学习方法主要是基于神经网络的端到端模型，同时完成实体的识别和实体间关系的抽取[3]。下面对流水线方法（基于 RNN 模型的实体关系抽取方法、基于 CNN 模型的实体关系抽取方法）进行介绍。

基于流水线的方法进行关系抽取的主要流程可以描述为：针对已经标注好目标实体对的句子进行关系抽取，最后把存在实体关系的三元组作为预测结果输出。一些基于流水线方法的关系抽取模型被陆续提出，其中，采用基于 RNN、CNN、LSTM 及其改进模型的网络结构，因其高精度获得了学术界的大量关注。

1. 基于 CNN 模型的实体关系抽取方法

CNN 的基本结构包括两层：其一为特征提取层，每个神经元的输入与前一层的局部接收域相连，并提取该局部的特征；其二是特征映射层，网络的每个计算层由多个特征映射组成，每个特征映射是一个平面，平面上所有神经元的权值相等，减少了网络中自由参数的个数。由于同一特征映射面上的神经元权值相同，所以 CNN 网络可以并行学习。图 3-2 描述了 CNN 用于关系分类的神经网络的体系结构。网络对输入句子提取多个级别的特征向量，它主要包括以下 3 个组件：词向量表示、特征提取和输出。图的右部分显示了句子级特征向量构建过程：每个词语向量由词特征（WF）和位置特征（PF）共同组成，将词语向量放入卷积层提取句子级特征[4]。图 3-2 的左上部分为提取词汇级和句子级特征的过程，然后直接连接以形成最终的句子特征向量。最后如图 3-2 的左下部分，通过隐含层和 softmax 层得到最终的分类结果[3]。

2. 基于 RNN 模型的实体关系抽取方法

RNN 在处理单元之间既有内部的反馈连接又有前馈连接，可以利用其内部的记忆来处理任意时序的序列信息，具有学习任意长度的各种短语和句子的组合向量表示的能力，已成功应用在多种 NLP 任务中。与 RNN 相比，前馈网络更适合处理序列化输入，但 RNN 也存在着以下两个缺点。

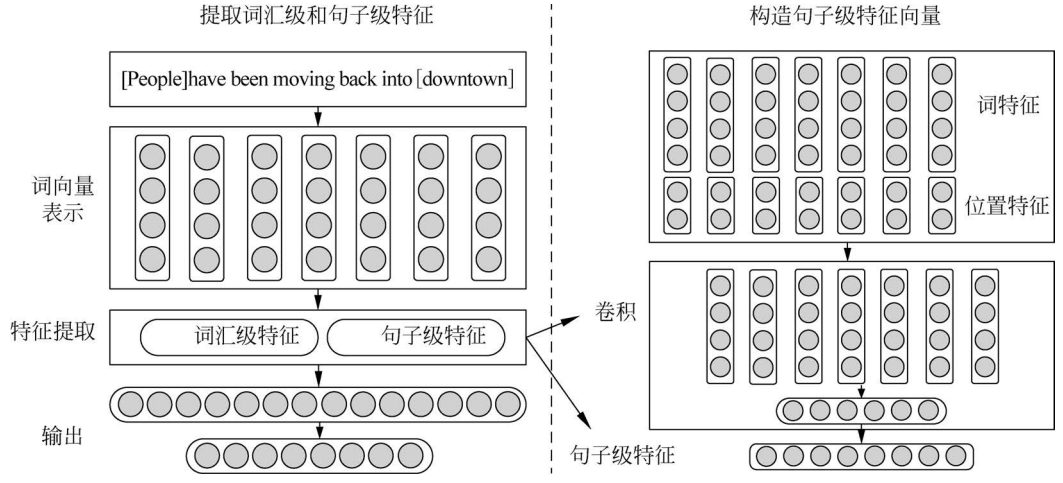

图 3-2 基于 CDNN 的关系抽取框架

(1) 在网络训练时,RNN 容易出现梯度消失、梯度爆炸的问题,因此,传统 RNN 在实际中很难处理长期依赖,这一点在 LSTM 网络中有所改进。

(2) 由于 RNN 的内部结构复杂,网络训练周期较长,而 CNN 结构相对简单,主要包括前置的卷积层和后置的全连接层,训练更快速。

由于梯度消失、梯度爆炸的问题,传统的 RNN 在实际中很难处理长期依赖,后面时间的节点对于前面时间的节点感知力下降。而 LSTM 网络通过 3 个门控操作及细胞状态解决了这些问题,能够从语料中学习到长期依赖关系。Yan 等提出了基于 LSTM 的融合句法依存分析树的最短路径以及词向量特征、词性特征、WordNet 特征、句法类型特征来进行关系抽取,该论文的模型如图 3-3 所示[5]。首先,如图 3-3 的左下部分,首先将句子解析为依赖树,并提取最短依赖路径(SDP)作为网络的输入,沿着 SDP,使用 4 种不同类型的信息(称为通道),包括单词、词性标签、语法关系和 WordNet 上位词[6];在每个通道中(图 3-3 右部分是每个通道的细节图),词语被映射成向量,捕获输入的基本含义,两个递归神经网络分别沿着 SDP 的左右子路径获取信息,网络中的 LSTM 单元用于有效信息的传播;之后,如图 3-3 左上部分,最大池化层从每个路径中的 LSTM 节点收集信息,来自不同通道的池化层连接在一起,然后输入到隐含层;最后,使用 softmax 输出层用于关系分类。

基于深度学习的有监督领域关系抽取方法与经典方法的对比如下:基于有监督学习的经典方法严重依赖于词性标注、句法解析等自然语言处理标注工具中提供的分类特征,而自然语言处理标注工具中往往存在大量错误,这些错误会在关系抽取系统中不断传播放大,最终影响关系抽取的效果。而基于深度学习的有监督方法可以在神经网络模型中自动学习特征,将低层特征进行组合,形成更加抽象的高层特征,用来寻找数据的分布式特征表示,能够避免人工特征选择等步骤,减少并改善特征抽取过程中的误差积累问题。

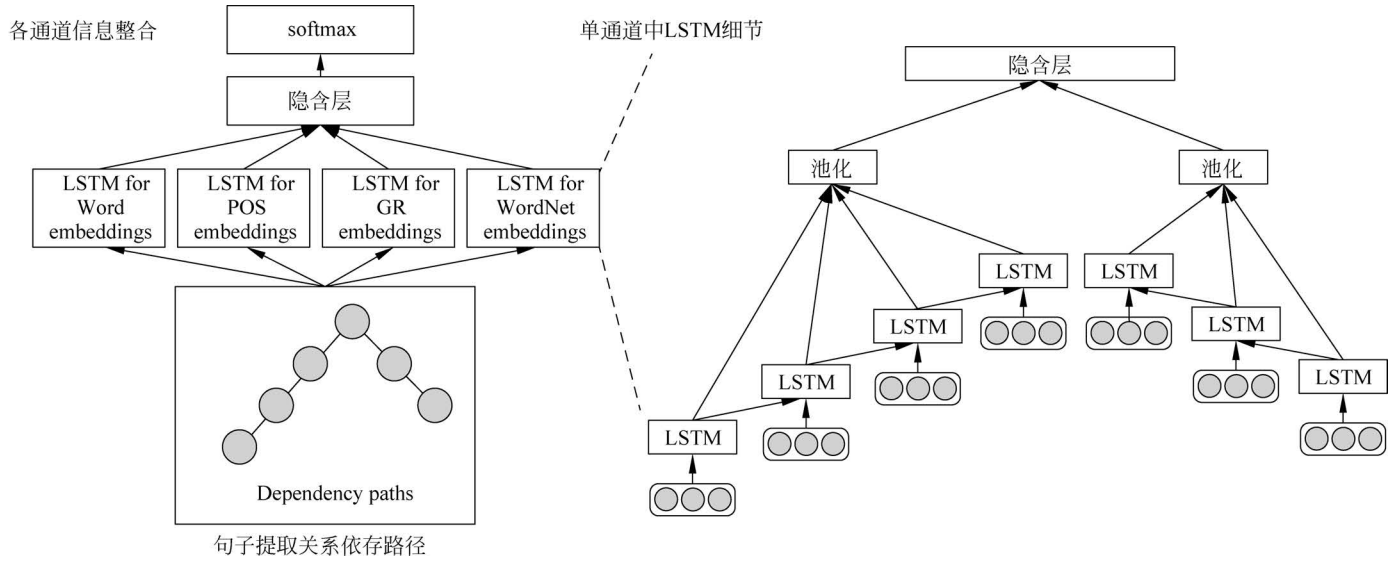

图 3-3　基于 LSTM 及最短依存路径的关系抽取方法

3.5 半监督实体关系抽取方法

半监督学习算法也称弱监督学习、半指导学习，它中和了监督算法和无监督算法的优点，只需要采用少量的标记语料，在分类模型的训练过程中加入无标记语料。典型的方法是自举(Bootstrapping)方法，该方法由 Sergey Brin 等提出，Bootstrapping 方法是指，在迭代的初始阶段，投入少量的种子集合，循序渐进，逐步将集合发展到体量具有一定规模[7]。同时，在迭代过程中，若实体对之间具有某种关系，则关系内容必须唯一。

目前，半监督学习方法仍处于探索和发展阶段，虽然克服了监督性学习在人工标注方面的问题，但自然语言语法复杂、句式多样且语义丰富，且半监督学习对关系种子集合具有较强依赖性。因此，致力于提出一个合理的获取语料集合的方法和高效的关系分类模型。

有监督学习方法具有较高的精确率，但有些过度依赖人工标记的语料，无法对大规模文本数据进行预测和分类；无监督学习方法能够处理大规模无结构文本数据，可移植性强，但无法确定抽取出的关系类别。此时，半监督学习方法应运而生。半监督学习使用少量已标注数据集作为初始种子集，通过一种循环学习机制去标注大量未标注数据。半监督实体关系抽取不但能够减少人工标记语料的数量，而且能够处理大量未标注语料集，所以受到了众多学者的推崇。目前，半监督关系抽取算法中常用的主要有协同训练方法、标注传播方法和 Bootstrapping 算法[8]。

1. 自举方法

自举方法是一种相对基础，抽取效果较好的半监督学习方法，典型的雪球(Snowball)就是以此思想为基础研究的关系抽取系统，其流程示意图如图 3-4 所示[14]。

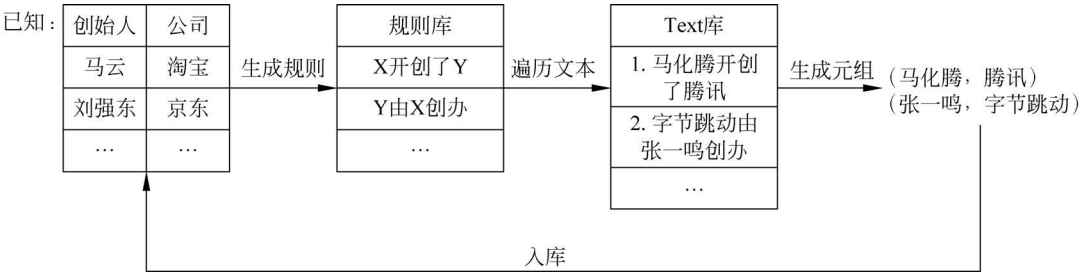

图 3-4 自举算法流程图

自举方法的训练大体可分为两个阶段。在第一阶段，需要提供少量具有代表性的种子集合，该种子集合可以是关系元组，其中的关系类型已经过人工标记[9]。然后将已标记元组结合语料训练一个监督性模型，产生一个可预测关系类型的分类器。至此，第一阶段的监督性训练结束。在第二阶段，利用已训练好的分类器对未经过关系标记的输入语料进行预测，可以得到一组新的关系元组，筛选其中置信度较高的部分，作为新的标记语料输入分类模型，对分类模型进一步训练与修正。重复该阶段直到达到停止条件，停止条件可以为人工定义的迭代次数或者运行直到没有新的关系元组生成，最终得到利用无标记样本强化的关系分类模型[8]。

自举是一种迭代式方法，其步骤如下。

(1) 给定一些初始种子(seed tuple)，即，具有某些关系的实体组，例如，<姚明 夫妻 叶莉>，<搜狗 CEO 王小川>。

(2) 从语料库中找到包含实体组的句子，根据这些句子总结出相应的模式，如 X 的妻子是 Y、X 的 CEO 是 Y。

(3) 根据新老 pattern 抽取更多的 tuple，再次总结 pattern，不断地进行迭代。

其中，具有代表性的自举方法有：DIPRE(Dual Iterative Pattern Relation Extraction)、Snowball。下面以 DIPRE 为例进行说明。以< Author，Book >这样一个关系举例阐述 DIPRE 方法，主要步骤如下。

(1) 从原始语料库中找到，包含种子实体组的语句上下文，并总结相应的 pattern(模式)。pattern 的具体表现形式为包含 5 个元素的元组，< order，urlprefix，prefix，middle，suffix >，简单地说，这 5 个元素为：当用 pattern 从语料库中寻找相应关系实体组时，若一个网页 URL 匹配 urlprefix * (* 为通配符)，且 order(顺序)为 True(order 为布尔值、其他为字符串类型)，该网页存在某句话能匹配正则式：* prefix，author，middle，title，suffix * ，则当前实体组< author，title >满足这个 pattern(注：若 order 为 False，则正则式中 author 与 title 的位置需要互换)。

(2) 根据总结的 pattern 从语料库中寻找相应的实体组。

(3) 根据实体组生成新的 pattern。

2. 协同训练

协同训练(Co-Training)目前是半监督机器学习方法研究中一个十分热门的方向，在实体关系抽取研究中也具有广泛的应用。协同训练要求训练语料能够分成完全独立并充分冗余的两个视图，将两组语料分别划分为用于训练初始分类模型的标记语料和用于加深训练过程的无标记语料，这样，可以通过两组标记语料训练得到两个分类模型。两个分类模型分别对相应的测试语料进行预测，将预测结果置信度较高的训练集交换到另一个分类器中继续训练，如此迭代直到所有的未标注测试数据集全部加入或者迭代次数达到阈值。目前已出现多种置信度评估的方法，以达到模型协同训练，优势互补的目的。图 3-5 展示了协同训练的核心思想。

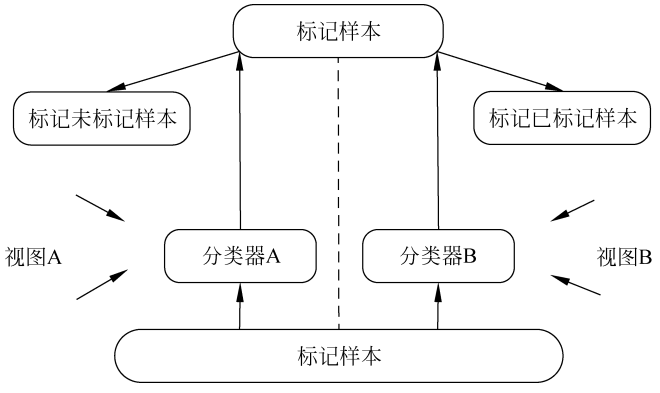

图 3-5　协同训练算法流程

协同训练算法之所以与自举不同，是因为它训练了具有差异的分类器，使得训练过程中两分类模型逐步趋向标准化，而这个问题的关键就是保证分类模型的完全独立，其根源是保证训练语料能够分解为两个视图，两者需要满足以下条件。

（1）视图可以充分表达问题，能够训练良好的分类模型。

（2）两个视图需要完全独立两个条件缺一不可，由于视图的拆分相对困难，所以研究者往往将其作为重点，良好的协同训练算法可以大大提升实体关系分类的性能。

有文献指出，在满足以上条件的前提下，协同训练中新增样本的信息度等同于随机抽样中样本的信息度。

3. 标注传播

标注传播算法使用图的思想实现标注样本对无标注样本的分类联合概率预测，采用相似度函数判断样本之间的权重，标注样本根据权重对相邻无标注样本传播分类概率[10]。标记数据与无标记数据通过相似度判断聚合成一个规模巨大的连通图，图的顶点为标记数据和无标记数据，标记数据会保持一个稳定的关于关系类型的概率分布，无标记数据需要标记数据根据边的权重来决定传播的难易度，顶点之间边的权重为两者的相似度，这需要预定义公式计算得到。在标注传播的过程中，标记数据顶点根据边的权重将关系概率信息传播给具有高相似度的相邻顶点上，使标记数据顶点对周围有连通性的顶点进行指导，相似度越高，指导性越强。标记传播方法对数据的训练过程可以平滑过渡到无标记训练数据，新标记的数据可以对其他数据继续训练，最终得到一个稳定的连通结构，保持了数据间的联络性。标注传播可以突破传统算法的一些局限性，它不受预料形态的影响，只要数据关系模式是相似的，标记结果就可以正常传递，已有研究者将其应用于实体关系抽取。但由于计算过程具有传播性，它同样具有在大规模语料条件下计算时间过长，消耗内存巨大的缺点，且如果不同数据类型所对应的语料数量过于不平衡，其系统的性能就会大大下降。因此标注传播在实体关系抽取领域的应用还在探索阶段。

3.6 远程监督实体关系抽取方法

在面对大量无标签数据时，有监督的关系抽取消耗大量人力，显得力不从心。与有监督实体关系抽取相比，远程监督方法缺少人工标注数据集，因此，远程监督方法比有监督多一步远程对齐知识库给无标签数据打标的过程[11]。而构建关系抽取模型的部分，与有监督领域的流水线方法差别不大。因此，远程监督实体关系抽取应运而生，通过数据自动对齐远程知识库来解决开放域中大量无标签数据自动标注的问题。远程监督标注数据时主要有两个问题：噪声和特征提取误差传播。噪声问题是由于远程监督的强假设条件，导致大量数据的关系被错误标记，使得训练数据存在大量噪声；而特征提取中的误差传播问题是由于传统方法主要是利用 NLP 工具进行数据集的特征提取，因此会引入大量的传播误差。自从深度学习的崛起并且在有监督领域取得良好的关系抽取效果后，用深度学习提取特征的思路来替代特征工程的想法越来越清晰：用词向量、位置向量来表示句子中的实体和其他词语；用深度模型对句子建模，构建句子向量；最后进行关系分类[12]。

下面基于图 3-6 介绍目前应用最广的基于 LSTM 的实体关系抽取方法。

（1）LSTM 网络抽取实体对方向性信息：HE 等首先将句子的最短依存路径（SDP）分

割成两个子路径作为 LSTM 结构的输入，自动地抽取特征，以此来抽取实体对的方向性信息。

（2）CNN 网络提取句子整体信息：尽管 SDP 对关系抽取非常有效，但是这并不能捕捉到句子的全部特征。针对此问题，作者将全部句子放进 CNN 网络，进而抽取句子的表示。

（3）特征融合：将 LSTM 隐含层单元以及 CNN 的非线性单元相融合，通过 softmax 层来标注实体对的对应关系。

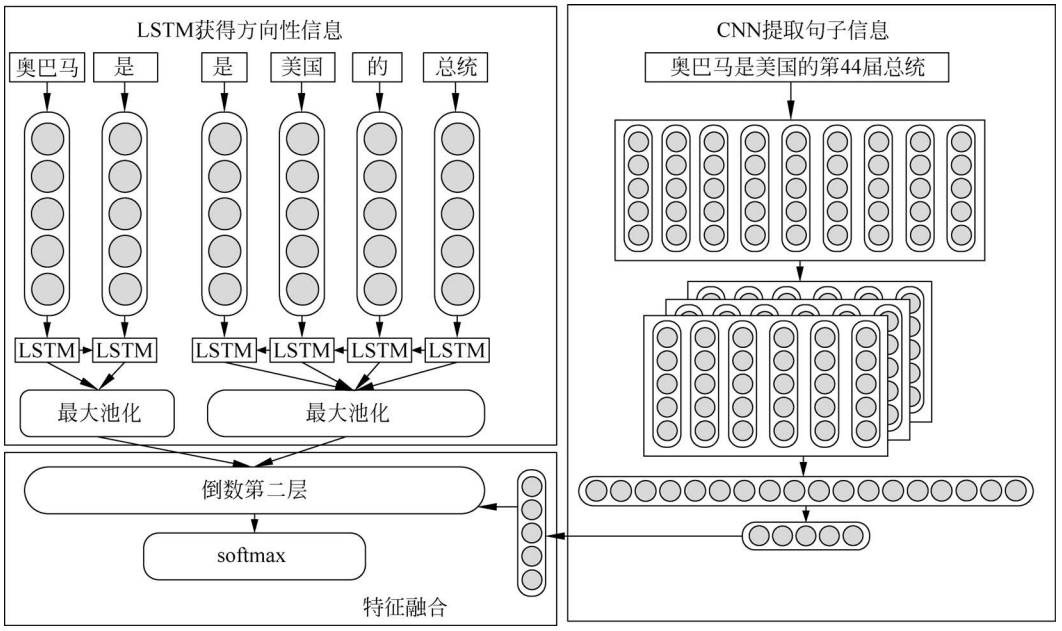

图 3-6　基于 LSTM 的远程监督实体关系抽取框架

1. 基于深度学习的远程监督关系抽取方法与经典方法的对比

经典的远程监督方法是在解决远程监督中强假设条件造成大量错误标签的问题，而深度学习方法主要是在解决特征提取中的误差传播问题[13]。

远程监督的提出，是因为在开放域中存在大量无规则非结构化数据，人工标注虽能使标注的精确率较高，但是时间和人力消耗巨大，在面对大量数据集时显得不切实际。因此，远程监督实现一种数据集自动对齐远程知识库进行关系提取的方法，可自动标注数据。但由于其强假设条件造成了大量错误标签问题，之后，经典的远程监督的改进都是在改进处理错误标签的算法[14]。

深度学习的提出，是为了解决数据特征构造过程依赖于 NER 等 NLP 工具，中间过程出错会造成错误传播的问题。现今基于深度学习的远程监督实体关系抽取框架已包含经典方法中对错误标签的处理，因此可以认为目前的远程监督关系抽取框架是基于传统方法的扩展优化。

2. 基于深度学习的远程监督关系抽取方法与有监督方法的对比

有监督的实体关系抽取依靠人工标注的方法得到数据集，数据集精确率较高，训练出的关系抽取模型效果较好，具有很好的实验价值。但人工标注数据集的方法需要耗费大量人力成本，且标注数据的数量有限、扩展性差、领域性强，导致构造的关系抽取模型对人工标注

的数据具有依赖性，不利于模型的跨领域泛化能力，领域迁移性较差[4]。

远程监督在面对大量无标签数据时，相较于有监督实体关系抽取具有明显优势。人力标注大量无标签数据显得不切实际，因此远程监督采用对齐远程知识库的方式自动标注数据，极大地减少了人力的损耗且领域迁移性较强。但远程监督自动标注得到的数据精确率较低，因此在训练模型时，错误标签的误差会逐层传播，最终影响整个模型的效果。因此，现今的远程监督实体关系抽取模型的效果普遍比有监督模型抽取效果差。

3. 关系抽取存在的问题与挑战

1）数据规模问题

人工精准地标注句子级别的数据代价十分高昂，需要耗费大量的时间和人力。在实际场景中，面向数以千计的关系、数以千万计的实体对以及数以亿计的句子，依靠人工标注训练数据几乎是不可能完成的任务。

2）学习能力问题

在实际情况下，实体间关系和实体对的出现频率往往服从长尾分布，存在大量的样例较少的关系或实体对。神经网络模型的效果需要依赖大规模标注数据来保证，存在“举十反一”的问题。如何提高深度模型的学习能力，实现“举一反三”，是关系抽取需要解决的问题。

3）复杂语境问题

现有模型主要从单个句子中抽取实体间关系，要求句子必须同时包含两个实体。实际上，大量的实体间关系往往表现在一篇文档的多个句子中，甚至在多个文档中。如何在更复杂的语境下进行关系抽取，也是关系抽取面临的问题。开放关系问题：现有任务设定一般假设有预先定义好的封闭关系集合，将任务转换为关系分类问题。这样的话，文本中蕴含的实体间的新型关系无法被有效获取。如何利用深度学习模型自动发现实体间的新型关系，实现开放关系抽取，仍然是一个开放问题。

参考文献

[1] 李冬梅，张扬，李东远，等. 实体关系抽取方法研究综述[J]. 计算机研究与发展，2020，57(07)：1424-1448.

[2] 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，2019，30(06)：1793-1818.

[3] 黄勋，游宏梁，于洋. 关系抽取技术研究综述[J]. 现代图书情报技术，2013(11)：30-39.

[4] 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.

[5] Yan Y，Okazaki N，Matsuo Y，et al. Unsupervised relation extraction by mining Wikipedia texts using information from the web[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP，2009.

[6] Jabbari A，Sauvage O，Zeine H，et al. A French corpus and annotation schema for named entity recognition and relation extraction of financial news [C]//Proceedings of The 12th Language Resources and Evaluation Conference，2020.

[7] Brin S. Extracting patterns and relations from the world wide web[C]//International Workshop on the World Wide Web and Databases，1998.

[8] Elsahar H，Demidova E，Gottschalk S，et al. Unsupervised open relation extraction[C]//European

Semantic Web Conference,2017.

[9] Zhao S,Hu M,Cai Z,et al. Modeling dense cross-modal interactions for joint entity-relation extraction [C]//International Joint Conference on Artificial Intelligence,2020.

[10] Zheng S,Hao Y,Lu D,et al. Joint entity and relation extraction based on a hybrid neural network [J]. Neuro Computing,2017,257: 59-66.

[11] Miwa M,Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,2016.

[12] Bekoulis G,Deleu J,Demeester T,et al. Joint entity recognition and relation extraction as a multi-head selection problem[J]. Expert Systems with Applications,2018,114: 34-45.

[13] Jiang X,Wang Q,Li P,et al. Relation extraction with multi-instance multi-label convolutional neural networks [C]//the 26th International Conference on Computational Linguistics: Technical Papers,2016.

[14] Zhu J,Nie Z,Liu X,et al. Statsnowball: A statistical approach to extracting entity relationships [C]//Proceedings of the 18th International Conference on World Wide Web,2009.