第5章

基于远程监督的RE

5.1 引言

在信息爆炸的时代，海量的文本数据不断涌现，其中蕴含着人类知识的宝藏。然而，要从这些数据中提取出有用的知识却是一项艰巨而烦琐的任务。RE作为NLP领域的重要任务之一，旨在从文本中自动识别和提取实体之间的语义关系，为知识图谱构建、问答系统和文本理解等任务提供技术支持。

传统的RE方法主要依赖于人工标注的训练数据，即人工为文本中的实体对手动标注相应的关系。然而，这种方法存在着高昂的标注成本和标注数据的稀缺性问题，限制了其在大规模应用中的可行性和效果。为了解决这一问题，研究者们提出了远程监督信息抽取的方法，该方法利用知识库中的实体关系作为监督信号，从未标注的文本数据中提取实体关系。远程监督信息抽取为信息抽取任务的研究和应用提供了一种更高效、更经济的解决方案。

5.2 问题引入

基于远程监督的RE方法具有许多优势。首先，它可以充分利用现有的知识图谱或数据库中的丰富事实，无须进行大规模的人工标注，从而大大降低了标注成本和时间。其次，远程监督方法可以快速适应不同的领域和语境，因为它建立在现有的知识之上，不受特定领域训练数据的限制。此外，远程监督方法还能够处理大规模的文本数据，从中挖掘出更多的关系实例，提高RE的覆盖范围和准确性。

然而，尽管远程监督信息抽取方法具有一定的优势，但仍然存在一些挑战和问题值得研究和探索。首先，知识库中的实体关系可能存在噪声和不准确性，这会影响到远程监督信号的质量，从而对信息抽取的性能产生负面影响。其次，远程监督信息抽取方法在处理未标注文本时，往往会面临实体消歧、指代消解和模糊性等语义问题，如

何有效地解决这些问题也是一个亟待解决的难题。此外，由于知识库的有限性，远程监督信息抽取方法在处理新领域或少样本情况下的效果可能不尽如人意。因此，如何提高远程监督信息抽取方法的稳健性和适应性，以适应不同领域和数据条件下的信息抽取需求，是当前研究的热点之一。针对这些问题，本章将深入研究远程监督信息抽取方法，并提出一些创新性的解决方案，期待在信息抽取领域得到更好的性能和效果。

5.3　基于对抗学习的远程监督 RE

5.3.1　引言

RE 旨在通过对文本中包含的实体之间的语义关系进行分类，从纯文本中提取关系事实。已经投入了许多工作来进行 RE 研究，早期的工作基于手工特征，或者基于神经网络的最近工作。这些模型都遵循监督学习的方法，这种方法是有效的，但在实践中高质量标注数据的需求是一个主要瓶颈。

手动标注大规模的训练数据需要耗费时间和人力。因此，Mintz 等提出了远程监督方法，通过对齐知识图谱和文本来自动生成训练句子。远程监督假设如果知识图谱中存在两个实体之间的关系，则包含这两个实体的所有句子都会被标注为具有该关系。远程监督是一种自动获取训练数据的有效方法，但它不可避免地会遇到错误标注的问题。

为了解决错误标注问题，Riedel 等提出了 MIL（多实例学习）。Lin 等进一步提出了一种神经注意力机制，用于降低噪声实例的权重。这些方法在 RE 中取得了显著的改进，但仍然远远不够令人满意。原因是大多数去噪方法只是以非监督的方式计算每个句子的软权重，这只能在信息丰富和噪声实例之间做出粗略的区分。此外，这些方法无法很好地处理那些具有不足句子的实体对。

为了更好地区分有用信息和噪声实例，受到对抗学习思想的启发，应用对抗训练机制来增强 RE 的性能。对抗训练的思想已经在 RE 中得到探索，通过对句子嵌入进行扰动生成对抗性例子，但这些例子不一定对应于现实世界的句子。相反，通过从现有的训练数据中进行采样生成对抗性例子，这样可能更能准确地定位现实世界中的噪声。

基于对抗学习的方法包含两个模块：鉴别器和采样器。该方法将远程监督数据分成两部分，有自信的部分和不自信的部分。鉴别器用于判断哪些句子更有可能被正确标注，将自信的数据作为正例，不自信的数据作为负例。采样器模块用于从不自信的数据中选择最具困惑性的句子，以尽可能地欺骗鉴别器。此外，在几个训练轮次中，还会动态地从不自信的集合中选择最具有用信息和自信度的实例加入自信集合，以丰

富鉴别器的训练实例。

鉴别器和采样器进行对抗训练。在训练过程中，采样器的行为将教导鉴别器专注于改进那些最具困惑性的实例。由于噪声实例无法降低采样器和鉴别器的损失函数，因此噪声会在对抗训练过程中逐渐被过滤掉。最终，采样器可以有效区分不自信数据中的有用信息丰富实例，鉴别器可以很好地对文本中的实体之间的关系进行分类。与前述的 MIL 去噪方法相比，本方法在更细的粒度上实现了更有效的噪声检测。

在“中国少数民族古籍总目提要”数据集上进行实验。实验结果表明，本方法的对抗去噪方法有效地降低了噪声，并显著优于其他基准方法。

5.3.2 相关工作

1. RE

RE 是 NLP 中的重要任务，旨在从文本语料库中提取关系事实。在 RE 的研究领域，已经提出了一些抽取方法，取得了一定的研究成果，特别是在监督式 RE 中。Mintz 等将纯文本与知识图谱对齐，假设所有提及两个实体的句子都能描述它们在知识图谱中的关系，提出了一种远程监督的 RE 模型。

然而，远程监督不可避免地伴随着错误标注的问题，针对这一问题，Riedel 等和 Hoffmann 等应用 MIL 机制进行 RE，考虑每个实例的可靠性，并将包含相同实体对的多个句子组合在一起以减轻噪声问题的影响。

近年来，神经模型在 RE 中得到了广泛应用。这些神经网络模型能够在不需要显式语言分析的情况下准确地捕捉文本中实体间的关系。基于这些神经架构和 MIL 机制，Lin 等提出了句子级别的注意力机制来减少错误标注对 RE 结果的影响。总体来说，这些 MIL 模型通常对有用信息丰富和噪声实例进行软权重调整。有些研究进一步采用外部信息来提高去噪性能，如 Liu 等通过手动设置标签置信度来去除实体对级别的噪声。

随着 RE 研究的不断深入，人们陆续提出了更复杂的 RE 机制，如强化学习等，被用来从噪声数据中选择正例句子。然而，这些复杂的机制通常需要很长的时间进行微调，并且在实践中的收敛性也存在一定需要改进的地方，针对这些问题，提出了一种新颖的基于对抗网络的细粒度去噪方法，通过对抗训练来进行 RE。该方法简单而有效，适用于多种神经网络架构，并能扩展到大规模数据。

2. 对抗训练

Szegedy 等提出通过向原始数据添加噪声形式的微小扰动来生成对抗性例子，这些扰动噪声对人类来说通常无法区分，但会导致模型做出错误的预测。Goodfellow 等分析了对抗性例子，并提出了用于图像分类任务的对抗训练。随后，Goodfellow 等

提出了一个成熟的对抗训练框架，并使用该框架训练生成模型。

对抗训练在 NLP 中也得到了探索。Miyato 等提出了通过向词嵌入添加扰动进行文本分类的对抗训练。扰动添加的思想进一步应用于其他 NLP 任务，包括语言模型和 RE。与通过向实例嵌入添加扰动生成伪对抗性的例子不同，对抗训练通过从真实世界的噪声数据中采样对抗性例子进行对抗训练。基于对抗学习的方法中的对抗性例子可以更好地对应 RE 的实际情境。因此，本方法更有利于解决远程监督中的错误标注问题，在实验中将会展示这一点。

5.3.3　方法

本节介绍了用于去噪 RE 的实例对抗训练模型。该模型将整个训练数据分为两部分，即自信实例集合 I_c 和不自信实例集合 I_u。采用句子编码器来嵌入表示句子语义。对抗训练框架由采样器和鉴别器组成，分别对应噪声过滤器和关系分类器。

1. 框架

实例对抗训练模型的整体框架包括鉴别器 D 和采样器 S，其中 S 从不自信集合 I_u 中采样对抗性例子，而鉴别器 D 通过学习判断给定实例是来自 I_c 还是 I_u。

假设每个实例 $s \in I_c$ 都暴露出其标记关系 r_s 的隐含语义。相反，不自信实例 $s \in I_u$ 在对抗训练过程中不能被信任地正确标记。因此，将 D 实现为一个函数 $D(s, r_s)$，用于判断给定实例 s 是否暴露出其标记关系 r_s 的隐含语义：如果是，那么该实例来自 I_c；如果不是，则该实例来自 I_u。

训练过程是一个最小最大博弈，可以形式化如下：

$$\phi = \min_{p_u} \max_{D} (E_{s \sim p_c}[\log(D(s, r_s))] + E_{s \sim p_u}[\log(1 - D(s, r_s))]) \quad (5.1)$$

其中，p_c 是自信数据的分布；采样器 S 根据概率分布 p_u 从不自信数据中采样对抗性例子。经过充分的训练，S 倾向于从 I_u 中采样那些信息丰富的实例，而不是噪声实例，而 D 成为对噪声数据具有良好稳健性的关系分类器。

2. 采样器

采样器模块旨在从不自信集合 I_u 中选择最具困惑性的句子，通过优化概率分布 p_u 尽可能地欺骗鉴别器。因此，需要计算不自信集合 I_u 中每个实例的困惑分数。

给定一个实例 s，可以使用神经网络句子编码器将其语义信息表示为嵌入向量 $\boldsymbol{y}$。在这里，可以根据句子嵌入向量 $\boldsymbol{y}$ 计算困惑分数如下：

$$\boldsymbol{C}(s) = W \cdot \boldsymbol{y} \quad (5.2)$$

其中，W 是一个分隔超平面。

进一步定义 $P_u(s)$为在 I_u 上的困惑概率：

$$P_u(s)=\frac{\exp(\boldsymbol{C}(s))}{\sum_{s\in I_u}\exp(\boldsymbol{C}(s))} \tag{5.3}$$

如式(5.3)所示，在不自信实例集合中，将那些具有高 $D(s,r_s)$分数的实例视为困惑实例，它们会欺骗鉴别器 D 做出错误的决策。一个优化的采样器会给这些最具困惑性的实例分配较大的困惑分数。因此，将优化采样器模块的损失函数形式化如下：

$$L_S=-\sum_{s\in I_u}P_u(s)\log(D(s,r_s)) \tag{5.4}$$

在优化采样器时，将组件 $P_u(s)$视为更新的参数。

需要注意的是，当一个实例被标记为 $r_s=\text{NA}$ 时，表示该实例的关系不可用，可能是不确定或没有关系。由于这些实例总是被错误地预测为其他关系，为了让鉴别器抑制这种趋势，特别将 $D(s,\text{NA})$定义为该实例在所有可行关系上的平均分数：

$$D(s,\text{NA})=\frac{1}{|R|-1}\sum_{r\in R,r\neq\text{NA}}D(s,r) \tag{5.5}$$

其中，R 表示关系的集合。

3. 鉴别器

鉴别器负责判断给定实例 s 的标记关系 r_s 是否正确。鉴别器基于实例的嵌入向量 $\boldsymbol{y}$ 与其标记关系 r_s 之间的语义相关性来实现。相关性使用 Sigmoid 函数计算，如下所示：

$$D(s,r_s)=\sigma(r_s\cdot\boldsymbol{y}) \tag{5.6}$$

其中，优化后的鉴别器将对 I_c(自信实例)中的实例分配高分，对 I_u(不自信实例)中的实例分配低分。优化鉴别器的损失函数如下所示：

$$L_D=-\sum_{s\in I_c}\frac{1}{|I_c|}\log(D(s,r_s))-\sum_{s\in I_u}P_u(s)\log(1-D(s,r_s)) \tag{5.7}$$

其中，优化鉴别器将组件 $D(s,r_s)$视为更新的参数。

在实践中，由于计算量过大，数据集通常无法频繁遍历。为了训练效率的便利性，可以简单地以近似概率分布对子集进行采样，提出了一种新的优化损失函数：

$$\widetilde{L}_D=-\sum_{s\in\hat{I}_c}\frac{1}{|\hat{I}_c|}\log(D(s,r_s))-\sum_{s\in\widetilde{I}_u}Q_u(s)\log(1-D(s,r_s)) \tag{5.8}$$

其中，$\hat{I}_c$ 和 $\widetilde{I}_u$ 分别是从 I_c 和 I_u 中进行采样的子集，而 $Q_u(s)$是对方程中 $P_u(s)$的近似，其定义如下所示：

$$Q_u(s)=\frac{\exp(\boldsymbol{C}(s)^\alpha)}{\sum_{s\in\widetilde{I}_u}\exp(\boldsymbol{C}(s)^\alpha)} \tag{5.9}$$

其中，α 是一个控制困惑概率分布锐度的超参数。为了一致性，进一步将方程中的 L_S 近似为

$$\widetilde{L}_S = -\sum_{s\in\widetilde{I}_u} Q_u(s)\log(D(s,r_s)) \tag{5.10}$$

其中，式(5.8)中的 $\widetilde{L}_D$ 和式(5.10)中的 $\widetilde{L}_S$ 用于优化对抗性训练模型。

4. 实例编码器

给定包含两个实体的实例 $\boldsymbol{s}$，采用多种神经网络结构将句子编码为连续的低维嵌入向量 $\boldsymbol{y}$，这些向量能够捕捉两个实体之间标记关系的隐含语义。

(1) 输入层。

输入层的目标是将离散的语言符号(即单词)映射为连续的输入嵌入向量。对于包含 n 个单词$\{w_1,w_2,\cdots,w_n\}$的实例 $\boldsymbol{s}$，使用 Skip-Gram 将所有单词嵌入 k_w 维空间$\{w_1,w_2,\cdots,w_n\}$中。对于每个单词 w_i，还将其与两个实体的相对距离嵌入为两个 k_p 维向量，然后将它们连接为一个统一的位置嵌入 p_i。最终，得到编码层的 k_i 维输入嵌入向量如下：

$$\boldsymbol{s}=\{x_1,x_2,\cdots,x_n\}=\{[w_1:p_1],[w_2:p_2],\cdots,[w_n:p_n]\} \tag{5.11}$$

(2) 编码层。

在编码层，选择了 4 种典型的体系结构，包括 CNN、分段卷积神经网络(Piecewose Convolutional Neural Network，PCNN)、RNN 和双向循环神经网络(Bidirectional Recurrent Neural Network，BiRNN)，将实例的输入嵌入进一步编码为句子嵌入。

CNN 将大小为 m 的卷积核滑动到输入序列$\{x_1,x_2,\cdots,x_n\}$上，得到 k_h 维的隐藏嵌入向量：

$$\boldsymbol{h}_i=\mathrm{CNN}(x_{i-\frac{m-1}{2}},x_{i-\frac{m-2}{2}},\cdots,x_{i+\frac{m-1}{2}}) \tag{5.12}$$

然后，对式(5.12)中这些隐藏嵌入向量进行最大池化，输出最终的实例嵌入向量 $\boldsymbol{y}$，如下所示：

$$[\boldsymbol{y}]_j=\max\{[h_1]_j,[h_2]_j,\cdots,[h_n]_j\} \tag{5.13}$$

PCNN 是对 CNN 的扩展，也采用大小为 m 的卷积核获取隐藏嵌入向量。随后，PCNN 将式(5.13)中隐藏嵌入向量划分为 3 个段落，分别为$\{h_1,h_2,\cdots,h_{e_1}\}$，$\{h_{e_1+1},h_{e_1+2},\cdots,h_{e_2}\}$和$\{h_{e_2+1},h_{e_2+2},\cdots,h_n\}$，其中 e_1 和 e_2 是实体位置。PCNN 对每个段落应用分段最大池化：

$$[y_1]_j=\max\{h_1,h_2,\cdots,h_{e_1}\} \tag{5.14}$$

$$[y_2]_j=\max\{h_{e_1+1},h_{e_2+2},\cdots,h_{e_2}\} \tag{5.15}$$

$$[y_3]_j = \max\{h_{e_2+1}, h_{e_2+2}, \cdots, h_n\} \tag{5.16}$$

通过式(5.14)、式(5.15)、式(5.16)中连接所有池化结果，PCNN 最终输出一个 $3 \cdot k_h$ 维的实例嵌入向量 $\boldsymbol{y}$，如下所示：

$$\boldsymbol{y} = [y_1; y_2; y_3] \tag{5.17}$$

如式(5.17)所示，RNN 是用于建模顺序数据的，它使其隐藏状态随时间步骤的变化与输入嵌入向量相对应：

$$\boldsymbol{h}_i = \mathrm{RNN}(x_i, \boldsymbol{h}_{i-1}) \tag{5.18}$$

其中，RNN(·)是循环单元，$\boldsymbol{h}_i \in \mathbf{R}^{k_h}$ 是时间步骤 i 的隐藏嵌入向量。本节中，选择 GRU 作为循环单元。将最后一个时间步骤的隐藏嵌入向量作为实例嵌入向量，即 $\boldsymbol{y} = \boldsymbol{h}_n$。

Bi-RNN 旨在融合句子序列的两侧信息。Bi-RNN 分为前向和后向方向，如下所示：

$$\overrightarrow{\boldsymbol{h}}_i = \mathrm{RNN}_{\mathrm{f}}(x_i, \overrightarrow{\boldsymbol{h}}_{i-1}) \tag{5.19}$$

$$\overleftarrow{\boldsymbol{h}}_i = \mathrm{RNN}_{\mathrm{b}}(x_i, \overleftarrow{\boldsymbol{h}}_{i+1}) \tag{5.20}$$

式(5.19)、式(5.20)中 $\overrightarrow{\boldsymbol{h}}_i$ 和 $\overleftarrow{\boldsymbol{h}}_i$ 分别是前向和后向 RNN 在位置 i 的隐藏状态。将前向和后向 RNN 的隐藏状态连接起来作为实例嵌入向量 $\boldsymbol{y}$：

$$\boldsymbol{y} = [\overrightarrow{\boldsymbol{h}}_n; \overleftarrow{\boldsymbol{h}}_1] \tag{5.21}$$

5. 初始化和实现细节

下面将介绍对抗训练模型的初始化和实现细节。将优化函数定义为

$$L = \widetilde{L}_D + \lambda \widetilde{L}_S \tag{5.22}$$

其中，λ 是一个调和因子。在实践中，对抗训练中的两个模块都使用随机梯度下降(Stochastic Gradient Descent，SGD)进行交替优化。

由于本模型框架比典型的生成对抗网络(Generative Adversarial Network，GAN)简单得多，不需要校准损失函数之间的交替比例，因此可以简单地使用 1∶1 的比例。这使得本模型能够有效地学习大规模数据。此外，还可以将 λ 整合到采样器 $\widetilde{L}_S$ 的学习率中，以避免调整超参数 λ。

对抗训练开始时，在整个训练数据上预训练一个关系分类器。关系分类器将整个数据分为一个小的有自信实例集合和一个大的无自信实例集合。在对抗训练过程中，每隔一段训练周期，从无自信实例集合中选择一些由采样器推荐且被鉴别器识别出来的实例，以丰富有自信实例集合。

5.3.4　实验设置

在本节中，通过实验来展示本实例对抗训练方法的有效性。首先介绍数据集和参数设置。然后，将本方法与传统的神经网络方法和基于特征的方法进行性能比较，用于RE。为了进一步验证本方法能够更好地区分那些有信息量的实例和噪声实例，还对那些只有少数句子的实体对进行评估。

1. 参数设置

在本模型中，从{0.5，0.1，0.05，0.01}中选择学习率 α_d 和 α_s，分别用于训练鉴别器和采样器。对于其他参数，参照Zeng(2014)、Lin(2016)和Wu(2011)提出的方法，简单地使用其中使用的设置，以便与基准模型的结果进行公平比较。表5.1显示了实验中使用的所有参数。在训练过程中，每10个训练周期就在无自信集中选择信息量最大、自信度最高的实例来丰富自信集。

表5.1　参数设置

参　数　名	参　数　值
鉴别器学习率(α_d)	0.1
采样器学习率(α_s)	0.01
CNN隐藏层维度(k_h)	230
RNN隐藏层维度(k_h)	150
CNN定位尺寸(k_p)	5
RNN定位尺寸(k_p)	3
词语尺寸(k_ω)	50
卷积核大小(m)	3
随机失活率(p)	0.5

2. 整体评估结果

遵循Mintz等的方法进行保留集评估。通过将测试集中的实体对与不同关系组合，构建候选三元组，并根据它们对应的句子表示对这些三元组进行排序。将知识图谱(Knowledge Graph，KG)中的三元组视为正确的，其他三元组视为不正确的，根据它们的精确率-召回率的结果评估不同模型的性能，如表5.2所示。

表5.2　不同召回率的各种模型的精确率　　单位：%

方　法		不同召回率下的精确率			精确率的平均值
		召回率=0.1	召回率=0.2	召回率=0.3	
CNN+	ATT	67.4	52.5	45.8	55.8
	AN	75.3	66.3	54.3	65.3

续表

方　法		不同召回率下的精确率			精确率的平均值
		召回率=0.1	召回率=0.2	召回率=0.3	
RNN+	ATT	63.4	55.9	48.4	55.0
	AN	75.2	64.3	55.5	65.8
PCNN+	ATT	69.5	60.4	51.6	60.6
	ADV	71.6	58.7	51.9	60.1
	AN	80.3	70.3	60.2	70.3
BiRNN+	ATT	66.2	58.8	52.6	64.4
	ADV	72.2	64.8	55.6	65.3
	AN	79.8	67.1	54.3	66.1

其中给出了各种神经网络架构(包括 CNN、RNN、PCNN 和 BiRNN)与各种去噪方法的结果：+ATT 是基于实例的选择性注意方法；+ADV 是通过向实例嵌入添加小的对抗扰动来进行去噪的方法；+AN 是基于对抗学习的远程监督的 RE 提出的对抗训练方法。还将基于对抗学习的方法与 Mintz、MultiR 和实例多标签(Multi-Instance Multi-Label，MIML)这些基于特征的模型进行比较。基线模型的结果均来自于相关论文或开源代码中报告的数据。从中可以观察到以下几点。

(1) 神经网络模型在整个召回率范围内明显优于所有基于特征的模型。当召回率逐渐增长时，基于特征的模型的性能迅速下降。然而，所有神经模型仍然保持稳定且具有竞争力的精确率。这表明，在噪声环境中，人工设计的特征无法很好地工作，NLP 工具带来的不可避免的错误将进一步影响性能。相比之下，神经网络模型自动学习的实例嵌入可以有效地从噪声数据中捕捉到隐含的关系语义，用于 RE。

(2) 对于 CNN(CNN 和 PCNN)和 RNN(RNN 和 BiRNN)，采用对抗训练的模型优于采用句级注意力的模型。句级注意力通过计算每个句子的软权重来减少噪声，仅对有信息的实例和噪声实例进行粗粒度的区分。相比之下，采用对抗训练方法训练的神经模型会生成或采样含有噪声的对抗性示例，并迫使关系分类器克服它们。因此，采用对抗训练的模型可以在更细粒度上提供有效的去噪效果。总体而言，采用本对抗训练方法的模型在采用对抗训练的模型中取得了最好的结果。这表明，与通过添加扰动生成伪对抗性示例相比，通过从真实实例中采样对抗性示例的方法可以更好地区分有信息的实例和噪声实例。

(3) 为了更好地比较各种去噪方法，此处将评估结果进行了展示。由于此处更关注排名靠前的结果的性能，在这里展示了当召回率为 0.1、0.2、0.3 时的精确率分数以及它们的平均值。可以发现，复杂的神经模型(PCNN、BiRNN)在使用相同的去噪方法时表现得比简单的神经网络(CNN、RNN)更好。采用对抗训练的方法显著改善了 CNN 和 RNN 的性能，而本方法(AN)的表现始终比对抗训练的基线方法(ADV)好得

多。改变去噪方法带来的改进比改变神经模型带来的改进更为显著。这表明错误标注问题是阻碍远程监督 RE 模型有效工作的关键因素。

3. 对抗训练的效果

为了进一步验证本对抗训练方法的有效性，在一个更具挑战性的场景中评估了本方法和传统的 MIL 去噪方法在实体对只有少量句子时的 RE 性能。

对于每个实体对，随机选择一个句子、两个句子和所有句子来构建 3 个实验设置。在保留集评估中报告了 P@100、P@200、P@300 和它们的平均值。由于 PCNN 在上述比较中是最好的神经模型，简单地使用 PCNN 来将本方法（AN）与最近的最先进的去噪方法（ATT）及其简单版本＋ONE 和＋AVG 进行比较。从结果中可以观察到以下几点。

（1）本方法在与 ATT 方法及其简单版本的比较中取得了一致且显著的改进，尤其是当每个实体对只对应一个或两个句子时。原因在于，包括 ATT 在内的大多数 MIL 去噪方法通常假设至少有一个提到给定实体对的实例可以表达它们的关系，并且总是为实体对选择至少一条有信息的句子。然而，这个假设并不总是成立，尤其是当实体对只对应少量句子时，很可能没有一个实例可以表达给定实体对的关系。相比之下，本对抗训练方法不受这个假设的限制。通过对实例进行个别处理，本方法在每个实体对的实例数量较少时仍然有效。

（2）当考虑更多实例时，所有模型的结果都有所改善。PCNN＋ATT 和 PCNN＋AN 比那些简单方法实现了更大的改进。远程监督数据的增长为训练 RE 模型提供了更多信息，但也带来了可能影响性能的更多噪声。本方法在数据增长时仍然保持着对 ATT 方法的优越性。这表明本方法可以提供更稳健可靠的方案来去噪远程监督数据。

4. 案例研究

对于数据集文本中频繁出现的关系“流传于”，此处使用采样器分别选取了正例和负例实例。对于每个句子，以粗体突出显示实体。从图 5.1 中可以发现：前面的正例明显对应于关系“流传于”，而那些负例则未能反映这种关系。这些例子表明本采样器能够有效区分有信息和有噪声的实例。本次实验的抽取结构如图 5.1 所示。

本节提出了一种通过实例级对抗训练的去噪远程监督 RE 方法。通过将整个数据集分为自信集和不自信集，本方法对采样器和鉴别器进行对抗训练。采样器的目标是从不自信集中选择最具困惑性的实例，而鉴别器的目标是区分来自自信集或不自信集的实例。在实验中，将此方法应用于不同的 RE 神经网络架构。实验结果表明，本方法在更细粒度上有效降低了噪声，并显著优于现有的基线方法。本方法还对那些实

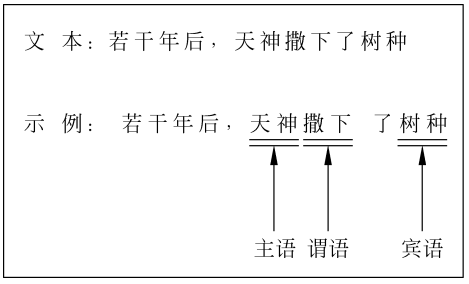

图 5.1 本次实验的抽取结构(见彩插)

例较少的长尾实体具有稳健性。

5.4 基于深度强化学习的远程监督 RE

5.4.1 引言

关系抽取是信息抽取和自然语言理解的核心任务。关系抽取的目标是预测句子中实体的关系。在下游应用中,关系抽取是构建知识图谱的关键模块,是结构化搜索、情感分析、问答和摘要等许多 NLP 应用的重要组成部分。

RE 算法早期开发遇到的一个主要问题是数据稀疏问题——人类注释者要通过数百万个句子的大型语料库来提供大量的标记训练是极其困难的。因此,远程监督关系抽取变得流行,因为它使用知识库中的实体对,并选择来自未标记的数据中的一组噪声实例。

近年来,人们提出了神经网络方法来在这些噪声条件下训练关系抽取器。为了抑制噪声,最近的研究提出使用注意力机制在一组噪声句子上放置软权重,并选择样本。然而,仅选择一个示例或仅使用软注意力权重并不是最佳策略,为了提高稳健性,需要一种系统的解决方案来利用更多实例,同时消除假正例(FP),并将它们放入正确的位置——负例集中。

在本节中研究了使用动态选择策略进行稳定监督的可能性。具体地说,设计了一个深度强化学习代理,其目标是学习根据关系分类器的性能变化选择是删除还是保留远程监督的候选实例。直观地说,代理希望删除 FP,并重建一组经过清理的远程监督实例,以根据分类准确性最大化重新分配。这个方法是与分类器无关的,它可以应用于任何现有的远程监督模型。根据经验,该方法在基于深度神经网络的模型中带来了一致的性能提升,在广泛使用的“纽约时报”数据集上取得了出色的性能。主要的贡献有以下 3 方面。

(1) 这是一种基于深度强化学习的远程监督模型。

(2) 该方法与分类器无关。

(3) 该方法可以提高神经相关的提取器的性能。

5.4.2　相关工作

Mintz 是第一个将依赖路径和特征聚合相结合进行远程监督研究的学者。但是，此方法会引入大量 FP，因为同一实体对可能具有多个关系。为了缓解及解决这个问题，Surdeanu 进一步提出了一个多实例多标签学习框架来提高性能。请注意，这些早期方法并没有明确消除噪声实例，而是希望模型能够抑制噪声。

近年来，随着神经网络技术的进步，引入了深度学习方法，并希望在隐藏层中模拟噪声的远程监督过程。然而，该方法只为每个实体对选择一个最合理的实例，不可避免地错过了许多有价值的训练。最近，Lin 提出了一种注意力机制，从一组嘈杂的实例中选择合理的实例。然而，软注意力权重分配可能不是最佳解决方案，因为 FP 应该完全删除并放在 F 集中。Ji 通过结合外部知识来丰富实体对，从而提高注意力权重的准确性。尽管上述这些方法可以选择高质量的实例，但它们忽略了 FP 的情况：一个实体对的所有句子都属于 FP。在这项工作中，将采取一种激进的方法来解决这个问题——尽可能多地利用远距离标记的资源，同时学习一个独立的 FP 指标来消除 FP，并将它们放在 P 集中。

5.4.3　实验过程

本节采用策略梯度的 RL(强化学习)方法来生成一系列关系指标，并通过将 FP 样本移动到 F 集，来到达重新分配训练数据集的目的。因此，本节实验旨在证明 RL 代理具有这种能力。

1. 数据和评估指标

在常用数据集上对该方法进行评估。该数据集首次在 Riedel 中提出。此数据集是通过将 Freebase 中的实体对与“纽约时报”语料库校准而生成的。“纽约时报”语料库的实体中提及其被以斯坦福大学命名的实体识别认可。2005—2006 年的句子用作训练语料库，2007 年的句子用作测试语料库。有 52 个实际关系和一个特殊关系 NA，它表明头部和尾部实体之间没有关系。NA 的句子来自存在于实际关系的同一句子中但不出现在 Freebase 中的实体对。

与之前的工作类似，采用保留评估来评估该模型可以提供对分类能力的近似测量，而无须花费昂贵的人工评估。与训练集的生成类似，测试集中的实体对也是从 Freebase 中选择的，这将用从“纽约时报”语料库中发现的句子进行预测。

2. 策略梯度

RL 代理的操作空间仅包含两个操作。因此，可以将代理建模为二元分类器。此研究采用单一窗口 CNN 作为网络。详细的超参数设置如表 5.3 所示。

表 5.3 超参数设置

超 参 数	值
窗口大小	3
卷积核大小	100
批量大小	64
调节器	100

至于词嵌入，直接使用 Lin 发布的词嵌入文件，它只保留了“纽约时报”语料库中出现次数超过 100 次的单词。此外，对位置嵌入具有相同的维度设置，最大相对距离设置为±30(“－”和“＋”代表整体的左侧和右侧)。强化学习的学习率为 2e－5。对于每种关系类型，固定的数值 γ_t、γ_v 是根据预先训练的代理。当一种关系类型有太多远距离监督的正例句时(例如，/location/location/contains 有 75 768 个句子)，对大小为 7500 个句子的子集进行采样来训练代理。对于删除句子的平均向量，在预训练过程和再训练过程的第一个状态中，将其设置为全零向量。

3. 关系分类器

关系分类器中使用了一个简单的 CNN 模型，因为简单的模型对于训练集的质量都很敏感。正例训练集中 P_t^{ori} 和 P_v^{ori} 的比例是 2∶1，它们都是从 Riedel 数据集的训练集中直接提取的。对应的 N_t^{ori} 和 N_v^{ori} 是从 Riedel 负例数据集中随机选取的，其大小是相对应的正例集的两倍。

4. FP 样本的影响

Zeng 和 Lin 研究了解决远程监督关系抽取错误标记问题的稳健模型。Zeng 将至少一个多实例学习与深度神经网络相结合，仅提取一个主动句来预测实体对之间的关系；Lin 将一个实体对的所有句子组合起来，并为其分配软注意力权重，以这种方式生成该实体对的综合关系表示。然而，FP 现象还包括一个实体对的所有句子都是 F 的情况，这是因为语料库与知识库不完全对齐。通过手动检查，这种现象在 Riedel 数据集和 Freebase 之间也很常见。显然，对于这种情况，上述两种方法都无能为力。

本节所提出的强化学习方法就是为了解决这个问题。采用 RL 代理通过将假阳性样本移动到负样本集中来重新分配 Riedel 数据集，然后使用 Zeng 和 Lin 预测此数据集上的关系，并将性能与原始 Riedel 数据集上的性能进行比较。在 RL 代理的帮助

下，相同的模型可以通过更合理的训练数据集实现明显的改进。为了给出更直观的比较，计算了每条 P-R 曲线的 AUC（曲线下面积）值，该值反映了这些曲线下的面积大小。这些可比较的结果也表明了基于策略的强化学习方法的有效性。而且，从 t 检验评估结果可以看出，所有 p 值均小于 5e－2，因此该改进的效果是显著的。

5. 实验结果

本节讨论了具有远程监督的深度强化学习框架，其中每个实体仅使用一个实例，结合软注意力权重实现远程监督。与已有同类框架相比，其优点在于通过引入基于梯度学习策略的重定位 FP 样本，未标记样本得到了更合理的利用。其目标是教导强化代理优化选择/重新分配策略，以最大程度提升关系分类性能。基于深度强化学习远程监督的关系抽取的特点在于框架不依赖关系分类器的特定形式，换句话说，它是一种即插即用的技术，可以应用于多种关系抽取的任务中。

本次实验的抽取结构如图 5.2 所示。

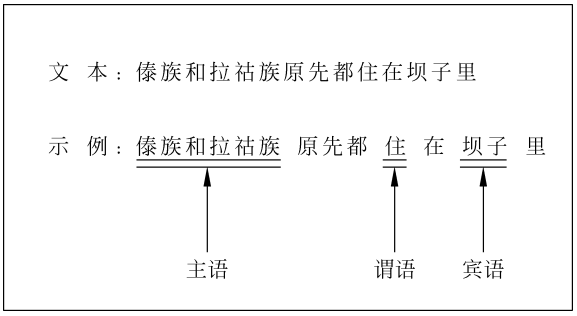

图 5.2　本次实验的抽取结构（见彩插）

5.5　基于句子级别注意力机制的远程监督 RE

5.5.1　引言

关系抽取是对给定句子中两个实体之间的关系进行分类的问题。远程监督是一种常用的开发关系抽取器的技术。此研究发现，在远程监督关系抽取设置中，大多数句子都很长，可能受益于词注意，以获得更好的句子表示。基于目前的发展，共在 3 方面提出了改进。首先，提出了两种新的词注意模型——基于双向门控循环单元的词注意（Bidirectional Gated Recurrent Unit Word Attention，BGWA）模型和以实体为中心的注意（Entity-centered Attention，EA）模型用于远程监督关系抽取，并且采用加权投票方法结合多个互补模型的组合模型，改进了关系抽取。其次，引入了一种新的用于关系抽取的远程监督数据集 GDS（谷歌远程监督）。GDS 消除了在所有以前远距离

监督基准数据集中存在的测试数据噪声，使可信的自动评估成为可能。通过在多个真实数据集上的大量实验，证明了所述方法的有效性。

对句子中两个实体之间的语义关系进行分类，称为关系抽取。对非结构化文本的理解是各种自然语言理解任务的重要步骤，如知识库构建、问答等。监督方法在关系抽取任务上取得了成功。然而，获得监督学习所需的大量训练数据是昂贵的，因此在Web规模的关系抽取任务中存在限制。

为了克服这一挑战，Mintz等提出了一种远程监督的关系抽取方法，通过采用文本语料库和知识库的交集，帮助自动生成新的训练数据。远程监督假设指出，对于参与关系的一对实体，任何在文本语料库中提到该实体对的句子都是关系事实的积极例子。这个假设是从一个实体对之间的多个关系标签的多个句子中输出证据，因此，将远距离监督数据集中的关系抽取问题作为一个MIML(多实例多标签)问题。然而，DS(远程监督)假设太强，由于知识库中的事实缺失，可能会引入假阴性样本等噪声。于是，提出了一个关系抽取模型和一个新的数据集来改进关系抽取。将“实例”定义为包含一个实体对的句子，而“实例集”定义为包含相同实体对的一组句子。

由Zeng等观察到的Riedel远程监督数据集是一个流行的DS基准数据集，其中有40个或更多的单词。此研究发现，并不是所有出现在这些长句子中的单词都有助于表达给定的关系。在这项工作中制定了各种词注意力机制，以帮助关系抽取模型关注给定句子中正确的上下文。

MIML假设指出，在一个对应于一个实体对的实例集合中，该集合中至少有一个句子应该表达分配给该集合的真实关系。然而，此研究发现，这在目前可用的基准数据集中并不总是正确的。特别是当前的数据集在测试集中有噪声时，例如，如果一个事实在知识库中缺失，它可能被标记为假，导致训练和测试集中出现假阴性标签。测试集中的噪声阻碍了模型的正确比较，并可能有利于过拟合的模型。为解决这一挑战，建立了GDS数据集，这是一个新的用于远程监督关系抽取的数据集。GDS来源于谷歌关系抽取语料库。这个新的数据集解决了远程监督评估的一个重要缺点，并使其在此设置下的自动评估更加可靠。

5.5.2 相关工作

在MIML设置中提出了远距离监督数据集的关系抽取。在这一领域的后续工作中有很大一部分是为了放松原始DS模型所做的强烈假设。在过去的几年中，深度学习模型已经减少了算法对手动删除依赖签名的特性。Zeng等介绍了使用基于CNN的模型进行关系抽取的方法，提出了一种PCNN(分段卷积神经网络)模型，利用分段最大池化方法保留了句子的结构特征，显著改善了精确回忆曲线。然而，PCNN方法在实例集中只使用了一个句子来预测关系标签和反向传播。通过引入一种注意力机

制，从实例集中选择一组句子进行关系标签预测，旨在利用排序模型中的句间信息进行关系抽取。所做的假设是，对于一个特定的实体对，每一个单独提到的信息可能不足以表达所讨论的关系，所以可能需要使用多次提到的信息来决定性地做出预测。

相关研究提出了通过增加背景知识来减少训练数据中噪声的方法。其中一篇文献提出了一个文本和知识库(Knowledge Base，KB)实体的联合嵌入模型，其中 KB 的已知部分作为监督信号的一部分。另一篇文献提出了使用间接监督，如关系标签之间的一致性、关系和参数之间的一致性，以及使用马尔可夫逻辑网络的邻居实例之间的一致性，降低噪声的影响。另外的文献针对在多任务设置中使用实例集间耦合抽取关系进行研究，从而提高性能。

注意模型通过反向传播来学习一个特征在监督任务中的重要性。神经网络中的注意力机制已经成功地应用于各种问题，如机器翻译、图像字幕和远距离监督关系抽取。

5.5.3　方法

1. 背景

Zeng 等提出了 PCNN 模型，这是一个成功的远距离监督关系抽取模型。关系抽取任务的成功与否取决于是否能从包含实体对的句子中提取出正确的结构特征。神经网络，如 CNN，已经被提出，以缓解为给定任务手动设计特征的需要。由于 CNN 的输出依赖于句子中标记的数量，因此经常采用最大池化操作来消除这种依赖关系。然而，使用一个单一的最大池错过了一些对关系抽取任务有用的结构特征。PCNN 模型将包含两个实体的句子 si 的卷积滤波器输出 ci 分为 3 部分，分别为 ci1、ci2、ci3，句子上下文分别在第一个实体的左边、两个实体之间以及第二个实体的右边，并对这 3 部分中的每个部分执行最大池化。因此，利用实体位置信息，在最大池操作后保留句子的结构特征。

2. BGWA 模型

通过句子实例，“柳宗元，唐代文学家，出生于河东郡，在今山西省运城市一带。”，考虑一下在实体对(柳宗元，河东郡)之间表达(人，城市)关系的句子。在句子中，短语“出生于”有助于识别句子中的正确关系。可以想象，识别这些关键短语或单词将有助于提高关系抽取性能。在此基础上提出的第一个模型——BGWA 模型使用一种对单词的注意力机制来识别这些关键短语。根据以往的经验，之前还没有关于在远程监督环境中使用词注意的工作。

BGWA 模型使用双向门控循环单元(Bidirectional Gated Recurrent Unit,Bi-GRU)来编码句子上下文。它是 RNN 的一种变体,旨在捕捉单词中的长期依赖关系。一个 Bi-GRU 在句子中同时向前和向后运行,以捕捉单词上下文的两边。

在此模型中,一个句子中只有少数单词(通常与实体有关)对关系的表示具有决定性作用。BGWA 模型的主要步骤如下。

(1) 词嵌入:将输入句子中的每个单词表示为一个向量,通常使用预训练的词嵌入(如 Word2Vec 或 GloVe)进行初始化。

(2) 双向 GRU 层:使用 Bi-GRU 对词嵌入进行编码。Bi-GRU 由两个独立的 GRU 组成,分别从左到右(正向)和从右到左(反向)处理句子。Bi-GRU 可以捕捉单词在不同上下文中的信息,从而获得更丰富的句子表示。

(3) 词注意力层:在这一步中,模型对各个单词的编码进行加权,从而突出与关系表示相关的关键词。注意力权重通过一个注意力机制计算得出,该机制学习如何根据单词的上下文信息分配权重。具体来说,注意力权重是通过一个全连接层和一个 Softmax 层计算得出的。

(4) 关系分类:将加权后的词向量相加,得到一个包含关系信息的句子向量。这个向量被输入一个全连接层和一个 Softmax 层,以预测句子中实体间的关系类型。

BGWA 模型的优点在于,它可以自动学习句子中与关系表示相关的关键词,而不需要手动设计特征。此外,通过使用 Bi-GRU,模型能够捕捉单词在不同上下文中的信息,从而获得更丰富的句子表示。然而,该模型仍可能受限于较大的词汇表和长句子,导致计算效率较低。在实际应用中,可以尝试使用更高效的模型架构(如 Transformer)和 PLM(如 BERT、GPT)来进一步提高关系抽取的性能。

3. EA 模型

“柳宗元,唐代文学家,出生于河东郡,在今山西省运城市一带。”有 4 个涉及的实体(柳宗元,河东郡,山西省,运城市),在句子中,对于实体柳宗元,文学家这个词有助于确定这个实体是一个人。这些额外的信息有助于通过只观察人际关系和城市之间的关系来缩小关系的可能性。于是在 2016 年沈雅田提出了一个实体注意力模型,以一个句子作为模型的输入。Sharmistha Jat 等对此模型进行修改和调整,以适应远程监督设置,并提出了 EA 模型。

EA 模型的目标是从输入句子中学习一个向量表示,该表示主要关注实体对周围的上下文信息,从而有助于预测它们之间的关系。

EA 模型的主要步骤如下。第一步是词嵌入,将输入句子中的每个单词表示为一个向量,通常使用预训练的词嵌入(如 Word2Vec 或 GloVe)进行初始化。第二步是上下文编码,对词嵌入进行上下文编码,通常使用 RNN、LSTM、GRU 或 Transformer

等神经网络架构。这一步骤的目标是捕捉句子中单词的上下文信息。第三步是实体注意力层，模型使用注意力机制为输入句子的实体对分配权重。注意力权重基于实体对和上下文编码之间的相互关系，通常通过一个全连接层和一个 Softmax 层计算得出。计算权重后，对实体对进行加权求和，以得到一个关注实体对上下文的句子表示。最后一步是关系分类，将实体注意力句子表示输入一个全连接层和一个 Softmax 层，以预测句子中实体间的关系类型。

EA 模型的主要优势在于它专注于实体对周围的上下文信息，从而有助于捕捉它们之间的关系。然而，EA 模型可能仍受限于计算效率和长句子处理问题。在实际应用中，可以尝试使用其他高效模型架构（如 Transformer）及 PLM（如 BERT、GPT）来进一步提高 RE 任务的性能。

4. 集成模型

BGWA、EA 和 PCNN 具有互补的优势。PCNN 利用 CNN 从句子中提取高级语义特征，然后使用分段最大池化层来选择最有效的特征。基于实体的注意有助于突出显示句子中出现的每个实体的重要关系词，从而称赞基于 PCNN 的特征。除了以实体为中心的词之外，并非句子中的所有词对关系抽取都同样重要。BGWA 模型通过选择与句子中的关系相关的单词来解决这一问题。

将 BGWA、EA 和 PCNN 三个模型结合在一起形成集成模型可以提高关系抽取任务的性能。集成模型通过结合多个模型的优点，降低单个模型的不足。

将这三个模型集成的一种方法如下。首先进行预处理，将输入句子中的每个单词表示为一个向量，通常使用预训练的词嵌入（如 Word2Vec 或 GloVe）进行初始化。其次进行上下文编码，对词嵌入进行上下文编码。这里可以使用 Bi-GRU，以捕捉句子中单词的上下文信息。再次到实体注意力层，应用 EA 模型的实体注意力层，为输入句子的实体对分配权重，得到一个关于实体对上下文的句子表示。然后到词注意力层，应用 BGWA 模型的词注意力层为各个单词的编码分配权重，突出与关系表示相关的关键词。将加权后的词向量相加，得到一个包含关系信息的句子向量。下一步到分段卷积层，利用 PCNN 模型的分段卷积层对上下文编码进行局部特征抽取。将实体对之间的单词分为三个段，分别应用卷积和池化操作，将这三个段的表示拼接在一起，形成一个向量。再下一步进行特征融合，将实体注意力句子表示、词注意力句子向量和分段卷积向量连接在一起，形成一个融合了三个模型特征的向量。最后进行关系分类，将融合特征向量输入一个全连接层和一个 Softmax 层，以预测句子中实体间的关系类型。

通过将 BGWA、EA 和 PCNN 三个模型集成在一起，可以利用它们各自的优点，例如 BGWA 的词注意力、EA 的实体注意力和 PCNN 的局部特征抽取。这种集成方

法有望在信息抽取任务中实现更高的性能。然而，这种方法可能会增加模型的复杂性和计算成本。在实际应用中，可以根据具体需求和资源进行调整。

5. GDS

存在几个使用 DS 进行关系抽取的基准数据集，DS 用于在所有这些数据集中创建训练集和测试集，从而引入噪声。虽然在远程监督中的训练噪声是预期的，但测试数据中的噪声是很麻烦的，因为它可能会导致不正确的评估。远程监督假设增加了两种噪声：由于缺少 KB 事实而具有不正确标签的(a)样本，以及没有实例支持 KB 事实的(b)样本。

为了克服这些挑战，开发了 GDS，这是一个新的使用远程监督的关系抽取数据集，其目的是降低 DS 设置中的噪声，确保标记关系是正确的，并且对于 GDS 中的每个实例集，该集合中至少有一个句子表示分配给该集合的关系。

GDS 是一种基于远程监督的关系抽取数据集，由斯坦福大学 NLP 组(Stanford NLP)开发。与传统的关系抽取数据集不同，GDS 利用知识图谱中的实体关系来远程监督数据集的标注，以提高标注的准确性和数据集的质量。

具体而言，GDS 使用开源的知识图谱(开放信息抽取，OpenIE)来抽取出实体之间的关系。然后，利用这些关系作为模板，自动标注大规模文本数据集中的实体关系。这种远程监督的方法可以解决传统关系抽取数据集中标注不全、标注错误等问题，同时也能够利用大规模的未标注文本数据，提高数据集的规模和覆盖面。

GDS 的数据集包括两部分：GDSv1 和 GDSv2。GDSv1 包括来自维基百科、新闻、百度百科等网站的文本数据，共计约 12 万个句子。GDSv2 则包括来自维基百科、新闻、论坛等网站的文本数据，共计约 300 万个句子。在 GDSv2 中，除了常见的关系类型，如“出生于”“拥有”等，还包括一些新的关系类型，如“受到”“出演”等。

GDS 的数据集已经被广泛应用于关系抽取、实体链接等 NLP 任务中，成为了该领域的重要基准数据集之一。同时，GDS 的远程监督方法也为其他关系抽取数据集的构建提供了参考。

5.5.4 实验结果

本次实验在“中国少数民族古籍总目提要”数据集上进行，将数据集按 7∶3 的比例分成训练集与测试集，在训练集上进行训练，在测试集上进行预测，用 BGWA 模型、EA 模型和集成模型的远程监督的信息抽取。

本次实验的抽取结构如图 5.3 所示。

BGWA 模型、EA 模型和集成模型中的性能评价指标，如精确率、召回率、$F1$ 评分、算法消耗时间如表 5.4 所示。

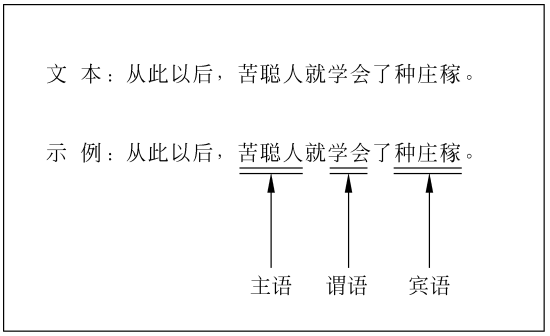

图 5.3　本次实验的抽取结构(见彩插)

表 5.4　不同模型的评价指标

模　　型	评 价 指 标			
	精确率	召回率	*F*1 评分	算法消耗时间/s
BGWA 模型	0.92	0.87	0.90	120
EA 模型	0.88	0.91	0.90	180
集成模型	0.95	0.94	0.92	100

从表 5.4 中可以看出，BGWA 模型的算法消耗时间相对较短，优于 EA 模型。集成模型从精确率、召回率、*F*1 评分和算法消耗时间上整体均优于 BGWA 模型和 EA 模型。

5.5.5　比较

1. 详细介绍三种模型

1) BGWA 模型

BGWA 模型是一种基于 GCN 和注意力机制的关系抽取模型。在 BGWA 模型中，将文本数据转换为图结构，其中每个实体对应一个节点，实体之间的关系对应边。然后利用 GCN 对图数据进行卷积操作，得到节点的隐藏层表示。为了进一步提高模型的性能，BGWA 模型还引入了注意力机制，对不同节点的表示进行加权，得到最终的关系表示。BGWA 模型具有较好的可解释性和稳健性，在一些关系抽取任务中取得了较好的效果。

2) EA 模型

EA 模型是一种基于注意力机制和实体对齐的关系抽取模型。在 EA 模型中，将文本数据转换为实体对齐的形式，其中每个实体对应一个向量，实体之间的关系对应向量的组合。然后利用注意力机制对不同实体的表示进行加权，得到最终的关系表示。EA 模型引入的实体对齐机制，对不同实体的表示进行对齐，可以更好地捕捉实

体之间的语义关系。EA 模型具有较好的可解释性和泛化能力，在一些关系抽取任务中取得了较好的效果。

3）BGWA、EA、PCNN 三者集成的模型

BGWA、EA、PCNN 三者集成的模型是一种基于多模态融合的关系抽取模型。在该模型中，将文本数据同时转换为图结构和实体对齐的形式，然后利用 BGWA 模型和 EA 模型分别对图数据和实体向量进行处理，得到两种不同的关系表示。为了进一步提高模型性能，还引入了 PCNN 模型，对文本数据进行卷积和池化操作，得到文本的关系表示。最后，将三种不同的关系表示进行融合，得到最终的关系表示。BGWA、EA、PCNN 三者集成的模型具有较好的泛化能力和稳健性，在一些关系抽取任务中取得了较好的效果。

2. 比较分析

从上述介绍中可以看出，BGWA 模型和 EA 模型都是基于 GNN 的关系抽取模型，分别采用了不同的方法来抽取实体之间的关系。BGWA 模型采用了图 CNN 和注意力机制，而 EA 模型采用了注意力机制和实体对齐机制。两种模型都具有较好的可解释性和稳健性，在不同的关系抽取任务中都有较好的表现。

BGWA、EA、PCNN 三者集成的模型则是基于多模态融合的关系抽取模型，将图数据、实体向量和文本数据同时考虑，以提高模型的泛化能力和稳健性。该模型在一些关系抽取任务中取得了较好的效果，但相对于单一模态的模型，其复杂度较高。

综合来看，选择哪种模型需要根据具体应用场景和需求来决定。BGWA 模型和 EA 模型适合处理图数据和实体向量，而 BGWA、EA、PCNN 三者集成的模型则适合处理多模态数据。

5.6 基于实体级别注意力机制的远程监督 RE

5.6.1 引言

远程监督方法在关系抽取中是一种常用的技术，但它也存在一些问题。远程监督方法使用知识库或标记数据中的实体关系标签来指导关系抽取。然而，这些标签可能存在噪声，因为知识库可能包含错误或不完整的信息。这会导致模型学习到错误的关系模式，从而影响关系抽取的准确性。另外远程监督方法假设知识库中的实体与文本中的实体是对齐的，即它们表示相同的实体。然而，在实际应用中，这种对齐并不总是完美的。如果对齐存在错误，那么通过远程监督得到的标签可能与实际的关系不匹配，导致训练出的模型表现不佳。在远程监督方法中，从已有的知识库启发式地与文

本对齐，将对齐结果视为标记数据。然而启发式对齐并不准确，可能会标记错误。

对于第一个问题，可以把远程监督关系抽取转换为一个多实例问题，其中考虑了实例标签的不确定性；对于第二个问题，直接不使用特征工程，而是使用具有分段最大池的卷积体系结构来自动学习相关特征。

5.6.2　相关工作

关系抽取是 NLP 中最重要的课题之一。许多用于关系抽取的方法被提出，譬如引导法、无监督关系发现和监督分类等。监督方法是关系抽取中最常用，也是效果相对较好的方法。在监督方法中，关系抽取被认为是一个多分类问题，并可能受到缺乏标记数据的影响。为了解决这个问题，Mintz 等采用了 Freebase 进行远程监督。另外，为了解决训练数据生成的算法有时会面临错误标签问题，Surdeanu 等提出了用于多实例学习的宽松远程监督假设。在多实例学习中，可以考虑实例标签的不确定性。多实例学习的重点是区分不同的实例集合。

这些方法已经被证明对于关系抽取是有效的。然而，它们的性能在很大程度上取决于设计的特征的质量。大多数现有的研究集中于提取特征以识别两个实体之间的关系。先前的方法通常可以分为两类：基于特征的方法和基于核函数的方法。然而基于特征的方法在将结构化表示转换为特征向量时需要选择合适的特征集。基于核函数的方法为利用输入分类线索的表示提供了一种自然的替代方法，例如句法解析树，能够使用大量特征而无须显式地提取它们。已经有几种核函数被提出，例如卷积树核函数、子序列核函数和依赖树核函数。然而，使用现有的 NLP 工具很难设计高质量的特征。随着最近对神经网络研究的不断深入，这里将多实例学习融入 CNN 用以完成这样的任务。其核心思想是将文本段落划分为不同的部分，并对每个部分进行卷积操作，从而捕捉局部上下文信息。

5.6.3　融入多实例学习的基于分段 CNN 的 RE

1. 分段 CNN

此模型是一种用于关系抽取的 CNN 模型，其目标是从文本中识别和抽取实体之间的关系。此模型的核心思想是将文本段落划分为不同的部分，并对每个部分进行卷积操作，从而捕捉局部上下文信息。具体来说，模型将文本段落划分为 3 部分：实体 1 之前的文本、实体 1 和实体 2 之间的文本，以及实体 2 之后的文本。如图 5.4 所示，把一个句子按两个实体切分为前、中、后 3 部分的词语，然后将一般的最大池化层相应地划分为 3 段最大池化层，从而获取句子的结构信息。

图 5.4 分段最大池化层模型

分段 CNN 的整体结构步骤如下。

(1) 文本特征输入表示。使用词嵌入和位置特征嵌入,把句子中每个词的这两种特征拼接起来。词嵌入使用的是预训练的 Word2Vec 词向量,用 Skip-Gram 模型来训练。位置特征是某个词与两个实体的相对距离,位置特征嵌入就是把两个相对距离转换为向量,再拼接起来。

(2) 卷积操作。对划分的 3 部分进行卷积操作。使用一维 CNN 对每个部分进行卷积操作,以捕捉局部上下文信息。卷积操作将局部窗口中的单词表示映射为固定长度的特征向量。

(3) 分段最大池化操作。设计了分段最大池化层代替一般的最大池化层,提取更丰富的文本结构特征。在每个部分的卷积结果上进行池化操作。使用池化操作来获取每个部分中最显著的特征向量。常用的池化操作是 max-pooling,选择每个部分中具有最大特征值的特征向量。一般的最大池化层直接从多个特征中选出一个最重要的特征,实际上是对卷积层的输出进行降维,但问题是维度降低过快,无法获取实体对在句子中所拥有的结构信息。

(4) 特征融合。将 3 部分的池化结果进行拼接,形成整个文本段落的表示。

(5) 关系分类。使用全连接层和 Softmax 激活函数对文本段落的表示进行分类,预测实体之间的关系类别。

此模型通过局部卷积和池化操作,能够有效地捕捉文本中实体之间的上下文信息,从而提高关系抽取的性能。它在关系抽取任务中取得了一定的成果,并被广泛应用于 NLP 领域。

2. 多实例学习

一般神经网络模型是句子经过神经网络的 Softmax 层后,得到概率分布,然后与关系标签的 one-hot 向量相比较,计算交叉熵损失,最后进行反向传播。这里多实例学习的做法则有些差别,目标函数仍然是交叉熵损失函数,但是基于实体对级别去计算损失,而不是基于句子级别。计算交叉熵损失分为如下两步。

(1) 对于每个实体对,会有 7 个包含该实体对的句子,每个句子经过 Softmax 层都可以得到一个概率分布,进而得到预测的关系标签和概率值。为了消除错误标注样

本的影响，从这些句子中仅挑出一个概率值最大的句子和它的预测结果作为这个实体对的预测结果，用于计算交叉熵损失，公式如下所示：

$$j^{*}=\operatorname{argmax} p(y_i \mid m_i^j;\theta) \quad 1 \leqslant j \leqslant q_i \tag{5.23}$$

（2）如果一个神经网络批大小包含有 T 个实体对，那么用第一步挑选出来的 T 个句子计算交叉熵损失，公式如下所示：

$$J(\theta)=\sum_{i=1}^{T}\log p(y_i \mid m_i^j;\theta) \tag{5.24}$$

得到交叉熵损失后，用梯度下降法求出梯度，并进行误差反向传播。

5.6.4　实验结果

这里实现的目标是从“中国少数民族古籍总目提要”数据集中提取人物之间的关系。这个数据集收集了大量的中国少数民族古籍，每篇文章都包含多个句子。将每篇文章的所有实例组成一个包，并为每个包分配一个标签，表示人物之间的关系。然后对于每个句子中的人物实体使用预训练的词嵌入模型将其转换为词向量。将每个词向量作为输入，结合上下文信息，形成句子的特征表示。然后将每个句子划分为实体之前、实体之间和实体之后的 3 部分。对于每个部分，使用卷积层和池化层来提取局部上下文特征。卷积层可以捕捉到词语之间的关系，而池化层可以提取每个部分的最显著特征。对于每个包，将其中的实例输入分段 CNN 中进行特征提取和关系分类。使用包级别的标签来训练分类器，以预测人物之间的关系。这样，就可以利用包内实例之间的关系信息进行关系抽取。本次实验的抽取示例如图 5.5 所示。

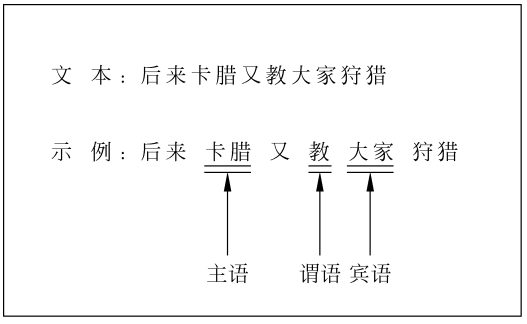

图 5.5　本次实验的抽取结构（见彩插）

随后进行模型训练和评估，使用带有包级别标签的训练数据对分段卷积神经网模型进行训练，优化模型参数。使用验证集对模型进行调优和选择超参数。其实验结果如表 5.5 所示，Mintz 代表了由 Mintz 等提出的传统的基于远程监督的模型；MultiR 是由 Hoffmann 等提出的一种多实例学习方法；PCNNs＋MIL 表示此模型的方法。表 5.5 中给出了提取的前 100、前 200 和前 500 个实例的手动计算精度。结果表明，

该方法的精度也比传统的估算方法高。

表 5.5 各个模型手动计算精度

模型	不同实例数量的手动计算精度		
	前 100 个实例的计算精度	前 200 个实例的计算精度	前 500 个实例的计算精度
Mintz	0.72	0.68	0.56
MultiR	0.79	0.72	0.58
MIML	0.82	0.74	0.62
PCNNs+MIL	0.84	0.81	0.71

5.7 基于图卷积的远程监督 RE

5.7.1 引言

随着近年来深度学习的发展,传统关系抽取方法逐渐被基于深度学习的关系抽取方法所取代。关系抽取作为 NLP 领域的重要任务之一,对于信息检索、知识图谱构建以及问题回答等应用具有关键性意义。然而,由于文本中的关系信息通常需要通过人工标注的训练数据进行学习,数据获取和标注成本高昂,限制了关系抽取方法的可扩展性和适用范围。为了解决这一问题,远程监督方法被引入,远程监督的关系抽取方法通过将知识库中的关系实例与非结构化文本自动对齐来训练抽取器。由于知识库中的事实可能存在错误或不完整性,远程监督方法会引入一定程度的标注错误。这些标注错误进而会导致模型在含有噪声的数据上进行训练,从而降低了关系抽取的性能。远程监督模型通常忽略大型知识库包含的现成的辅助信息。为解决这一问题,提出了一种改进的远程监督神经网络关系抽取方法,通过引入额外的辅助信息来增强模型的性能。这些辅助信息可以是关系别名、实体类型等领域知识。此方法设计了一种远程监控神经网络关系抽取方法,它利用大型知识库中的附加边信息来改进关系抽取,使用实体类型和关系别名信息在预测关系时施加软约束。此方法使用图卷积网络从文本中对语法信息进行编码,并在有限的边信息可用时提高性能。

5.7.2 相关工作

远程监督的性能在很大程度上依赖于手工设计特征的质量。另外在神经网络关系的抽取模型研究中,Zeng 等于 2014 年提出了一种基于 CNN 的端到端方法,用于自动捕捉相关词汇和句子级特征,后续又通过使用分段最大池化进一步改进了性能。另外,Nagarajan 等采用了注意力机制来从多个有效句子中进行学习。此方法中也利用了注意力机制,用于学习句子和包的表示。研究表明,依存树特征在关系抽取任务中

发挥重要作用。Mintz 等的研究发现，通过利用依存树特征可以更准确地捕捉实体之间的关系。He 等则通过基于 tree-GRU 模型的方法进一步利用依存树特征取得了令人满意的结果。除了依存树特征，近年来，GCN 在关系抽取中也受到了广泛关注。Defferrard 等提出的 GCN 被证明在建模句法信息方面非常有效。利用 GCN 可以捕捉句子中的上下文依赖关系，从而更好地理解实体之间的关系。因此本方法在关系抽取中采用了 GCN，以利用依存树特征和句法信息，提高关系抽取任务的性能。另外从知识图谱中获取了实体类型和关系别名的辅助信息，并合理地利用它们。

5.7.3　利用辅助信息进行远程监督神经 RE

模型的整体结构如图 5.6 所示。模型首先通过连接来自 Bi-GRU 和 Syntactic GCN 的嵌入（用⊕表示）来对每个标记进行编码，然后应用词注意力机制。接着将句子嵌入与来自侧信息获取部分的关系别名信息连接起来计算句子的注意力。最终，包含实体类型信息的包表示被送入 Softmax 分类器进行关系预测。详细的展开过程将在后面的三部分中详述。

1. 句法句子编码

句法句子编码是一种将句子的句法结构信息编码为向量表示的技术。在 NLP 领域，句法结构指句子中单词之间的语法关系，如句子的依赖关系和短语结构。传统的词袋模型或序列模型通常只考虑了单词的线性顺序，而忽略了句法结构的信息。然而，句法结构对于句子的语义理解和语言推理非常重要。这一结构的目标是将句子的句法结构信息纳入句子的表示中，从而增强模型对句子语义的建模能力。为了实现这一结构，本节方法在连接的位置和单词嵌入上使用 Bi-GRU 对每个标记的本地上下文进行编码。尽管 Bi-GRU 能够捕获本地上下文，但它无法捕获可以通过依赖项边缘捕获的远程依赖项。为了捕获远程依赖，对于给定的句子，使用 Stanford CoreNLP 生成其依赖树。然后在依赖关系图上运行 GCN，通过获得的编码句法信息，以及嵌入的侧面信息来改进神经网络关系的抽取，并将其编码附加到每个令牌的表示中。最后，使用对标记的关注来抑制不相关的标记并获得整个句子的嵌入。

具体过程如下：每个句子都有 m 个标记 $\{w_1, w_2, \cdots, w_m\}$，首先使用 k 维的 GloVe 嵌入来表示每个标记。为了将标记与目标实体的相对位置结合起来，使用 p 维的位置嵌入将组合的标记嵌入叠加在一起，得到句子表示 $h \in \mathbf{R}^{m \times (k+2p)}$，然后，使用 Bi-GRU 除 h，得到新的句子表示 $h^{\text{gru}} \in \mathbf{R}^{m \times d_{\text{gru}}}$，其中 d_{gru} 是隐藏的状态维数。

虽然 Bi-GRU 能够捕获本地上下文，但它无法捕获可以通过依赖关系边缘捕获的远程依赖关系。这里使用句法 GCN 来编码这些信息。对于给定的句子使用 Stanford

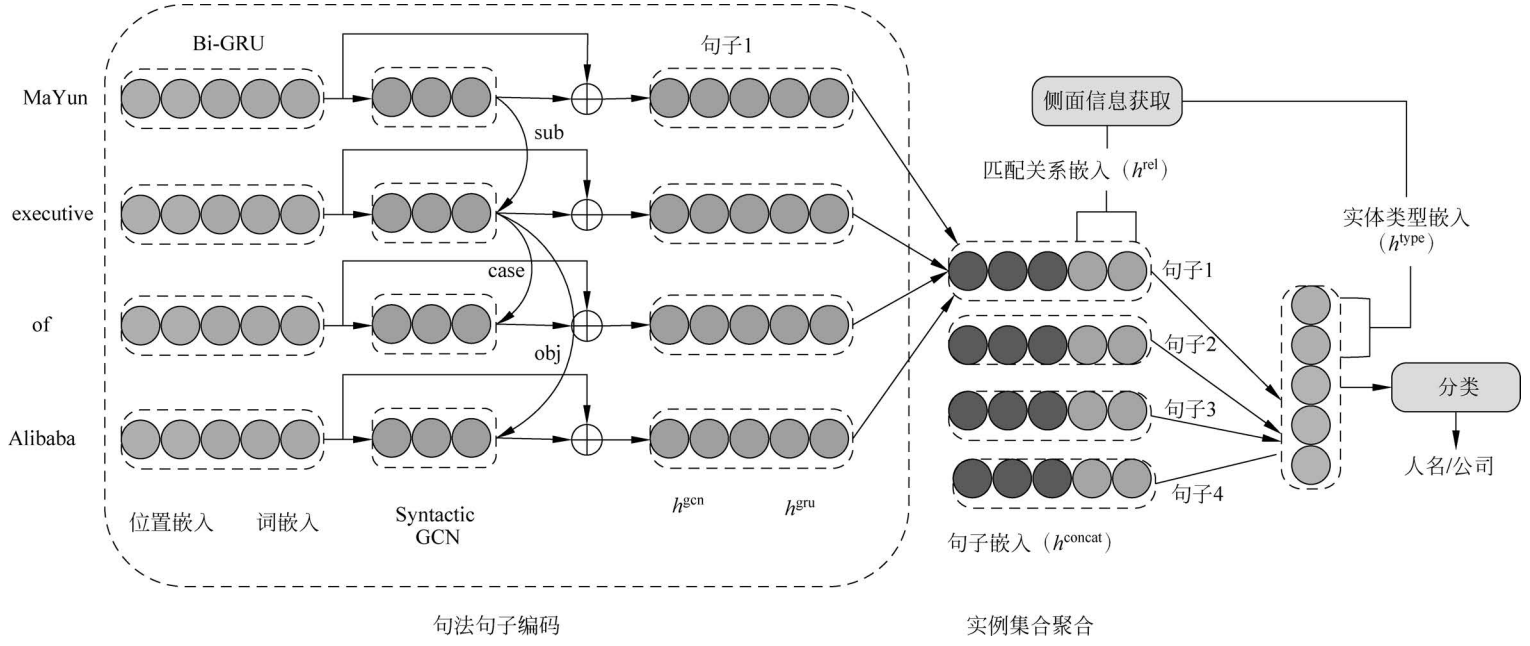

图 5.6　模型的整体结构(见彩插)

CoreNLP 生成它的依赖树。然后在依赖关系图上运行 GCN，使用公式(5.25)更新嵌入，以 h^{gru} 作为输入：

$$h_v^{k+1}=f\Big(\sum_{u\in N(v)} g_{uv}^k\times(W_{l_{uv}}^k h_u^k+b_{l_{uv}}^k)\Big) \tag{5.25}$$

对于有向图，从节点 u 到带有标签 l_{uv} 的节点 v 的边表示为 (u,v,l_{uv})；$N(v)$ 表示基于更新后的边集的相邻 v 的邻居集合；f 是任意非线性激活函数；k 表示 GCN 层。

对于每个令牌 w_i，GCN 嵌入 $h_{i_{k+1}}^{\text{gcn}}\in\mathbf{R}^{d_{k+1}}$ 后的层定义如下所示：

$$h_{i_{k+1}}^{\text{gcn}}=f\Big(\sum_{u\in N(i)} g_{iu}^k\times(W_{l_{iu}}^k h_{u_k}^{\text{gcn}}+b_{L_{iu}}^k)\Big) \tag{5.26}$$

其中，g_{iu}^k 表示定义的边缘门控；l_{iu} 表示边缘标签。在整个实验中，使用 ReLU 作为激活函数 f。将 GCN 的句法图编码添加到 Bi-GRU 输出中，得到最终令牌表示 $h_i^{\text{concat}}=[h_i^{\text{gru}};h_{i^{k+1}}^{\text{gcn}}]$。

2. 侧面信息的获取

在侧面信息的获取中，为了增强模型在关系抽取任务中的性能，这里利用了知识图谱的额外监督信息，并结合了开放信息抽取方法获取相关的侧面信息。在基于远程监督的关系抽取中，由于实体来自知识库，因此可以利用有关实体的知识来改进关系抽取。

知识图谱是一种结构化的知识库，其中包含了大量的实体、关系和属性信息。可以从知识图谱中获取额外的监督信息，例如实体的描述、关系的定义等，这些信息可以用于指导关系抽取模型的训练和推理过程。通过利用知识图谱中的监督信息可以提高模型对于关系的准确性和泛化能力。

另外，侧面信息的获取还使用了开放信息抽取方法，该方法可以自动地从大规模未标注的文本数据中抽取出潜在的关系事实。与传统的信息抽取方法不同，开放信息抽取不依赖于预定义的本体或模式，而是通过无监督学习的方式发现可能的关系三元组。这些抽取出的关系事实可以作为侧面信息，为关系抽取模型提供额外的背景知识和上下文信息。

通过结合知识图谱的监督信息和开放信息抽取方法提取的侧面信息，可以改善关系抽取模型的性能。这些额外的信息可以丰富模型对于关系的理解，提供更多的语义上下文，并帮助模型更准确地推断和分类关系。这种综合利用侧面信息的方法为关系抽取任务带来了更全面和准确的信息支持。

具体步骤如下。

用 P 表示提取目标实体之间的关系短语，使用释义数据库进一步扩展关系别名集 R。为了将 P 与扩展关系别名集 R 匹配，使用 GloVe 嵌入在 d 维空间中进行投影。

使用词嵌入来投射短语有助于进一步扩展这些集合，因为语义上相似的词在嵌入空间中更接近。然后，对于每个短语 $p \in \mathrm{P}$，计算其与 R 中所有关系别名的余弦距离，并将最接近的关系别名对应的关系作为该句子的匹配关系。在余弦距离上使用一个阈值来去除噪声别名。然后为每个关系定义一个 k_r 维嵌入，将其称为匹配关系嵌入 h^{rel}。对于给定的句子，h^{rel} 与其表征物 s 连接，由句法编码器得到，对于带有 $|\mathrm{P}| > 1$ 的句子，可能会得到多个匹配关系。在这种情况下，取它们嵌入的平均值。

3. 实例集聚合

实例集聚合指将句法句子编码器生成的句子表示与前一步得到的匹配关系嵌入进行连接的过程。这样可以将句子的语义信息与匹配关系的特征进行融合，进一步提升关系抽取模型的性能。在进行实例集聚合之后，应用注意力机制将注意力放在句子级别上，以学习整个句子集合的表示。通过对不同句子的重要性进行加权，注意力机制能够更加关注对关系抽取任务最有贡献的句子。接下来，将注意力加权后的句子表示与实体类型嵌入进行连接。实体类型嵌入是对实体的类型信息进行编码的向量表示。通过将实体类型嵌入与句子表示相连接，可以引入实体的语义特征，帮助模型更好地理解实体与句子之间的关系。最后，将连接后的特征输入 Softmax 分类器中，进行关系预测。Softmax 分类器根据输入的特征向量，通过计算各个关系类别的概率分布，确定最可能的关系类别。

具体步骤如下：为了使用所有有效的句子，使用句子的注意力来获得整个袋子的表示。将每个句子的嵌入与匹配关系嵌入 h^{rel} 连接起来。第 i 个句子的注意分值如下所示：

$$\alpha_i = \frac{\exp(\hat{s}_l \cdot \boldsymbol{q})}{\sum_{j=1}^{n} \exp(\hat{s}_l \cdot \boldsymbol{q})}, \quad \hat{s}_l = [s_i ; h_i^{\mathrm{rel}}] \tag{5.27}$$

其中，$\boldsymbol{q}$ 表示一个随机查询向量。包表示法 β 是句子的加权和，然后进行主体 $h_{\mathrm{sub}}^{\mathrm{type}}$ 和客体 $h_{\mathrm{obj}}^{\mathrm{type}}$ 实体类型嵌入连接。得到 $\hat{\beta}$，如下所示：

$$\hat{\beta} = [\beta ; h_{\mathrm{sub}}^{\mathrm{type}} ; h_{\mathrm{obj}}^{\mathrm{type}}], \quad \beta = \sum_{i=1}^{n} \alpha_i \hat{s}_l \tag{5.28}$$

最后 $\hat{\beta}$ 被送到 Softmax 分类器，得到关系的概率分布公式为

$$p(y) = \mathrm{Softmax}(W \cdot \hat{\beta} + b) \tag{5.29}$$

5.7.4 实验结果

在本次实验中，在“中国少数民族古籍总目提要”数据集上对模型进行了评估。本节实现的目标是从“中国少数民族古籍总目提要”数据集中提取人物之间的关系，例如

师徒关系、夫妻关系等。本次实验的抽取示例如图 5.7 所示。

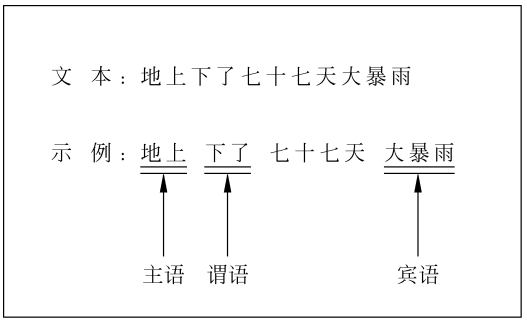

图 5.7　本次实验的抽取结构(见彩插)

本实验收集了大量的中国少数民族古籍总目提要数据，每篇文章都包含多个句子。经过多次模型优化。其实验结果如表 5.6 所示。PCNN 是一个由 Zeng 等提出的基于 CNN 的关系抽取模型，使用分段最大池来表示句子。PCNN＋ATT 是由 Lin 等提出的一种基于 CNN 的分段最大汇聚模型，以获取句子表示，然后在句子上进行注意力机制。BGWA 是由 Jat 等提出的一种基于 Bi-GRU 的词和句级注意的关系抽取模型。这里用不同数量的句子来评估这个方法，结果显示，该方法的精度比用于对比的方法的精度要高。

表 5.6　不同数量句子的评估精度

模　　型	100 个句子的评估精度	200 个句子的评估精度	500 个句子的评估精度
PCNN	0.71	0.69	0.65
PCNN＋ATT	0.77	0.72	0.68
BGWA	0.81	0.75	0.71
RESIDE	0.85	0.79	0.76