# 利用示例间相关性的远程监督实体关系抽取方法

## 引言

传统的有监督关系抽取系统需要大量手工标注的数据，标注数据需要耗费大量的时间和精力。为减少模型对人工标注数据的依赖，2009年，Mintz等[18]使用远程监督的方法自动标记数据，但该方法容易出现标记错误的问题，影响关系分类器的性能。

为缓解噪声样本对模型的影响，多示例学习框架被引入到远程监督实体关系抽取任务中。在多示例学习框架下，句子水平的选择注意力模型被用于动态减小包内噪声示例的影响。但是，第一，多示例学习方法将具有相同实体对的句子放入同一个包中，包中的示例具有一定的联系，这是包内示例的重要监督信息。传统的句子水平的注意力模型假设包内的示例都是相互独立的，因此忽略了包内示例之间的相关性，这不可避免地导致监督信息丢失。第二，使用句子水平的注意力模型依靠目标关系选择包内的有效示例，使得计算得到的包内示例的权重分布不准确，使一些非噪声示例的权重非常小，导致正确标记示例的监督信息丢失。

针对上述问题，提出一种利用示例间相关性的远程监督实体关系抽取方法。该方法在多示例学习框架下进行建模，使用一层卷积和自注意力机制对示例进行编码，利用包内示例间的相关性为包内正确标记示例分配较高的权重，从而关注到所有正确标记的示例。

本章的主要内容及结构安排如下：4.2节对远程监督实体关系抽取方法进行分类讨论并对问题进行总结分析；4.3节说明利用示例间相关性的远程监督实体关系抽取方法的算法原理和实现细节；4.4节构建实验验证该方法的效果；4.5节归纳总结本章的主要内容。

## 相关工作

### 主要技术和方法

远程监督方法标记的数据中存在大量噪声样本。大部分远程监督实体关系抽取方法均在多示例学习框架下缓解噪声的影响，这些方法主要分为：基于“至少一个”假设的方法；句子水平的选择注意力方法；过滤噪声示例的方法；改进注意力机制的方法。

#### 基于“至少一个”假设的方法

基于“至少一个”假设的方法假定在多示例学习框架下一个包内至少有一个提及这两个实体的示例将表达它们的关系，并且从每个包内选择最高概率的示例用于训练和预测。Zeng等[20]为解决远程监督方法带来的噪声样本的问题，将远程监督实体关系抽取任务看作一个多示例学习问题，首先使用PCNN对示例进行编码，然后从每个包中选取置信度最高的一个示例用于训练模型以减少噪声样本对模型的影响。He等[55]同样在多示例学习框架下，选取包中置信度最高的样本训练模型，并使用LSTM作为编码器。但该方法在训练和预测时仅选择包内最可能的示例，将丢失大量包含在被忽略示例中的信息，从而影响关系分类器的性能。

#### 句子水平的选择注意力方法

句子水平的选择注意力方法在包内建立句子级别的注意力，能够动态减少噪声示例的权重。Lin等[21]提出一种句子水平的选择注意力模型，该模型使用注意力机制为包内的示例动态分配权重。该模型首先对包中的句子进行编码得到句子的向量表示；其次通过计算句子向量表示和包的关系向量的相似度赋予该句子权重；然后对包内的各个句子进行加权求和得到包向量表示；最后在包上进行关系分类。Ji等[33]也提出一种句子水平的注意力模型，并且引入了背景知识。该模型首先使用“实体-实体=关系”的方法表示实体之间存在的关系；然后使用CNN学习实体描述页面特征；最后通过计算实体间关系和包内句子的相似度为句子分配权重。

Wu等[56]在多示例多标签框架下也使用注意力机制来选择包中的句子，并引入噪声以加强模型的健壮性。Xing等[57]同样在多示例学习框架下使用句子水平的注意力机制，并且提出一种新颖的独立头尾卷积神经网络提取更好的句子语义特征，采用粗略策略过滤不存在关系的示例以减轻嘈杂的数据问题。

但该方法将包内的示例视为相互独立的，忽略了包内示例间的相关性；该方法计算得到的权重分布不准确，使一些标记正确的示例的权重非常小，造成监督信息的丢失，从而影响关系分类器的精确率；另外，该方法忽略包内的所有示例都是噪声的情况，限制了关系分类器性能的提升。

#### 过滤噪声示例的方法

过滤噪声示例的方法利用强化学习、生成式对抗网络的方法过滤噪声示例，选择包中高质量的示例用于模型的训练。Feng等[36]提出一种包含示例选择器和关系分类器的模型。首先示例选择器通过强化学习方法选择高质量的示例并将其输入到关系分类器中，然后关系分类器预测示例的关系类型并计算相应的概率值作为示例选择器的奖励。实验结果表明，该模型可以有效地处理噪声，在句子级别上能获得较好的关系分类性能。Zeng等[58]提出一种具有强化学习的模型，该模型将实体对之间的关系用作监督信息，借助强化学习的方法指导关系抽取器的训练。Qin等[59]也尝试使用一种深度强化学习框架自动识别噪声样本，该方法中奖励函数的奖励根据关系分类器的性能变化来计算。Qin等[60]使用一种生成式对抗网络在没有任何监督信息的情况下自动识别每个关系类型实例集中的噪声样本。

但是该方法在句子水平过滤噪声数据，不能处理包内的所有示例都是噪声的情况，从而影响关系分类器性能的提升。

#### 改进注意力机制的方法

改进注意力机制的方法通过改进注意力机制减小噪声示例的影响。Du等[34]发现传统的一维注意力模型无法学习不同的上下文来预测实体对之间的关系，因此提出一种多层结构的自注意力机制用于关系抽取。Han等[61]发现大多数方法都是独立的处理每个关系类型，忽略了关系层次结构中存在的丰富语义相关性。因此提出一种结合关系层次信息的注意力方法，在多个层提供从细到细的粒度，以更好地识别有效示例，这对于抽取长尾关系特别有效。

此外，Yuan等[35]提出一种跨关系跨包的选择注意力模型，该方法通过考虑关系之间的相关性改善句子水平的注意力机制，并在包级别上使用注意力机制，第一个注意力机制能够减少嘈杂或不匹配句子的影响，第二个注意力机制选择高质量的包。Ye等[29]在包内和包间同时使用注意力机制，以分别处理句子水平和包水平的噪声。

### 问题总结与分析

远程监督实体关系抽取方法在监督信息利用、减轻噪声影响等方面还有提升空间：

1. 基于“至少一个”假设的方法在训练和预测时仅选择包内概率值最高的示例，造成被忽略示例中的信息丢失，从而影响关系分类器的性能。
2. 句子水平的选择注意力方法将包内示例视为相互独立，忽略了包内示例间的相关性且部分正确标记示例的权重非常小，造成监督信息的丢失，从而影响关系分类器的精确率；另外该方法忽略了包内的所有示例都是噪声的情况，噪声包会影响关系分类器性能的提升。
3. 过滤噪声示例的方法在句子水平过滤噪声数据，不能处理包内的所有示例都是噪声的情况，从而影响关系分类器性能的提升。

综上所述，针对选择注意力方法将包内示例视为相互独立使得信息利用不充分导致关系分类器精确率低的问题，提出一种利用示例间相关性的远程监督实体关系抽取方法，该方法利用相似度衡量示例间的相关性，示例与包内其他示例的相似度越高，其对应的权重越大，实现为正确标记示例分配较高的权重，从而能够关注到包内所有正确标记的示例。

## 算法原理

### 算法框架

提出一种利用示例间相关性的远程监督实体关系抽取模型，该模型在多示例学习框架下进行建模，主要由4个模块组成，分别是嵌入表示、编码器、示例自选择注意力模块和关系分类器，该模型的原理框架如图4.1所示。



图4.1 利用示例间相关性的远程监督实体关系抽取方法原理框架

其中，嵌入表示模块首先通过预训练的方法得到输入句子中每个单词的词向量，然后为指定输入句子中实体对的位置需要在词向量中添加位置向量，最后将输入句子中所有单词的向量表示叠加为嵌入矩阵；编码器首先读取嵌入矩阵和两个实体，然后通过卷积和自注意力机制学习示例的向量表示；示例自选择注意力模块首先将具有相同实体对的示例的向量表示放入同一个包中，然后依据包内示例间的相关性为包内示例计算权重，最后通过加权求和的方法计算得到包向量表示；关系分类器首先读取包向量表示，然后在包向量表示上预测关系类型。

### 嵌入表示

嵌入表示模块用于将输入中的每个单词转换为低维的实值向量表示。此外，为指定实体对的位置，还需要对输入中的所有单词使用位置向量。

对于包中的第个示例，其中，表示包中示例的数量，表示示例长度，表示示例中第个单词的向量表示，该向量由词向量和位置向量两部分组成。首先通过查找由Word2vec工具训练得到的嵌入表示矩阵将单词映射为维度是的向量表示；然后使用Zeng等[5]提出的位置嵌入方法得到每个单词距离头实体和尾实体的相对距离的向量表示，维度是；最后将每个单词对应的词向量和位置向量拼接起来形成一个新的向量表示，维度。

### 编码器

编码器用于学习示例的语义向量表示。将示例输入编码器中学习得到该示例的向量表示，其中，表示编码器的维度，如式(4.1)所示。



其中，表示编码器中涉及到的函数 。

本模型不使用CNN或者PCNN作为编码器，而是使用一种新的编码器学习示例更好的向量表示。该编码器由4层组成，分别是位置编码层、卷积层、自注意力层和前馈层。编码器结构如图4.1的中间部分所示。在编码器的开始位置添加一个位置编码，这是因为自注意力机制无法学习输入序列中单词的位置信息。位置编码使用Vaswani等[37]提出的方法编码每个单词的位置信息。卷积层使用一层记忆效率和泛化能力更好的深度可分离卷积神经网络。自注意力层使用多“头”注意力机制。前馈层由一个卷积操作和一个最大池化操作组成。

### 示例自选择注意力模块

在远程监督实体关系抽取方法中，选择注意力模型使用远程监督方法标记的关系标签作为“查询”选择正确标记示例，未考虑包内示例间的相关性。由于同一个包内的示例共享同一个实体对，因此示例之间存在或多或少的联系，如果包内的两个示例能表达同一种关系类型，则这两个示例的语义向量表示越相似。为此，本方法使用示例向量间的相似性衡量包内示例间的相关性，为与包内其他示例相似的示例分配更高的权重。该方法在选择示例时不依赖目标关系，仅依赖包内的示例选择正确标记的示例。

给定一个包含个示例的包，这些示例均包含相同的实体对。将该包内的所有示例进行编码之后，可以得到每个示例的向量表示。因此，包，其中，。为充分利用包内正确标记示例的信息，利用示例间的相关性为包内的示例动态的分配权重，通过加权求和的方法得到包的向量表示，计算过程如式(4.2)所示。



其中，，表示包中第个示例对应的权重，的计算过程如式(4.3)所示。



其中，表示包内第个示例与包内其他示例的相似度的总和，使用简单的点乘计算包内示例之间的相似度，计算过程如(4.4)所示。



其中，表示示例的向量表示进行归一化之后的结果，归一化的过程如式(4.5)所示。



### 关系分类器

得到包的向量表示之后，直接在包水平上进行关系预测。

首先，为避免过拟合，在包向量表示上使用dropout[62]策略；然后，将包的向量表示与一个关系矩阵进行相似度计算得到该包被预测为每种关系类型的置信度，其中，表示关系的数量，计算过程如式(4.6)所示。



其中，表示偏值向量。

最后，使用Softmax函数计算将包预测为第种关系类型的条件概率，计算过程如式(4.7)所示。



其中，表示模型中所有可训练的参数。

在本模型中，损失函数定义为交叉熵损失函数，计算过程如式(4.8)所示。



其中，表示训练集中包的数量，表示模型中所有可训练的参数，使用随机梯度下降的方式更新模型的参数。

## 实验分析

### 实验目的和数据源

为验证利用示例间相关性的远程监督实体关系抽取方法的有效性，在广泛使用的New York Times（NYT）数据集上进行实验。

该版本的NYT语料的训练集包含较多的噪声样本，本节将其命名为NYT-MORE。该数据集由Riedel等[19]发布，并在远程监督实体关系抽取的研究中被广泛使用。该数据集通过远程监督的方法将Freebase[63]与New York Times语料库自动对齐而生成。该数据集将2005-2006年的文章中的句子用作训练集，2007年的文章中的句子被用作测试集。该数据集含有53种关系类型，其中包括一种NA（Not A relation）关系，表示两个实体之间不存在关系类型。该数据集的详细统计信息如表4.1所示。

表4.1 NYT-MORE数据集的统计结果

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 测试集 |
| 句子数量 | 570,088 | 172,448 |
| 实体对数量 | 291,699 | 96,678 |
| 三元组数量 | 19,429 | 1,950 |

### 实验环境和条件

实验使用的硬件资源如表4.2所示，包括实验计算机和服务器，分别用于本地编写代码和远程运行代码。实验使用的软件资源如表4.3所示。

表4.2 实验所用硬件资源

|  |  |  |
| --- | --- | --- |
| 硬件名称 | 数量 | 配置 |
| 实验计算机 | 1 | Inter i7-6700，CPU 3.40GHz，RAM 8.00G，Windows 7 64 system |
| 服务器 | 1 | GPU GTX 1080 Ti，显存 10.92G，Ubuntu 16.04.4 LTS |

表4.3 实验所用软件资源

|  |  |  |
| --- | --- | --- |
| 软件名称 | 来源 | 说明 |
| Python3.6.8 | https://www.python.org/downloads/release/python-368/ | 开源的跨平台计算机程序设计语言 |
| Tensorflow1.8 | https://www.tensorflow.org/ | 开源的深度学习框架 |
| Pytorch1.0.1 | https://pytorch.org/ | 开源的Python机器学习库 |

### 评价方法

在NYT-MORE数据集的held-out测试集上评估本模型，将模型预测的关系和Freebase中的关系进行对比。

评价指标使用精确率/召回率曲线（*Precision/Recall Curve*，*PR*）、曲线下面积（*Area Under Curve*，*AUC*）。

*PR*曲线是以*Precision*和*Recall*为变量的曲线，其中*Precision*为纵坐标轴，*Recall*为横坐标轴。*PR*曲线上的点表示：在某一阈值下，模型将大于该阈值的结果判定为正，小于该阈值的结果判定为负，此时结果所对应的*Precision*和*Recall*。通过将阈值从高到低移动生成*PR*曲线，*PR*曲线的生成流程如图4.2所示。



图4.2 *PR*曲线的生成流程图

*AUC*值定义为*PR*曲线下与坐标轴围成的面积，是一种用来度量分类模型好坏的一个标准，*AUC*值越大，模型的效果越好。

### 对比分析实验

#### 实验过程和参数

共选取8种不同的模型和本文提出的SelfCON+SATT模型进行对比。

Mintz[18]（2009）、MultiR[64]（2011）和MIMLRE[65]（2012）是3种基于特征的方法；PCNN+ATT[21]（2016）和PCNN+ATT+soft-label[22]（2017）是2种使用选择注意力机制的方法；PCNN+ATT+RL[59]（2018）、PCNN+ATT+GAN[60]（2018）和PCNN+ATTRA+BAGATT[29]（2019）是远程监督实体关系抽取领域3种新方法，对于这3种方法，直接运行论文作者公开的源代码获得*PR*曲线数据。

其中，PCNN表示使用分段卷积神经网络作为编码器，ATT表示使用Lin等[21]提出的选择注意力方法，SelfCON表示使用卷积和自注意力机制作为编码器，SATT表示示例自选择注意力模块。

实验过程包括2个部分。

第一部分，为进行可视化比较，绘制SelfCON+SATT和这8种模型的前2000个点组成的*PR*曲线；

第二部分，为进行定量比较，依据第一部分绘制的*PR*曲线计算SelfCON+SATT模型、PCNN+ATT+RL[59]、PCNN+ATT+GAN[60]和PCNN+ATTRA+BAGATT[29]这4种模型的*AUC*值。

实验中SelfCON+SATT模型的参数设置如表4.4所示。

表4.4 SelfCON+SATT模型的参数设置

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 参数 | 词向量维度 | 位置向量维度 | 批的大小 | 学习率 | 卷积核窗口大小 | 头数 | dropout |
| 值 | 50 | 5 | 32 | 0.01 | 7 | 8 | 0.9 |

其中，使用Lin等[21]公布的词嵌入矩阵，表中头数指编码器中多“头”注意力机制的“头”的数量。

#### 实验结果和结论

##### 实验结果

SelfCON+SATT模型和这8种模型的*PR*曲线如图4.3所示。根据图4.3的*PR*曲线计算SelfCON+SATT、PCNN+ATT+RL[59]、PCNN+ATT+GAN[60]和PCNN+ATTRA+BAGATT[29]这4种模型的*AUC*值，实验结果如表4.5所示。

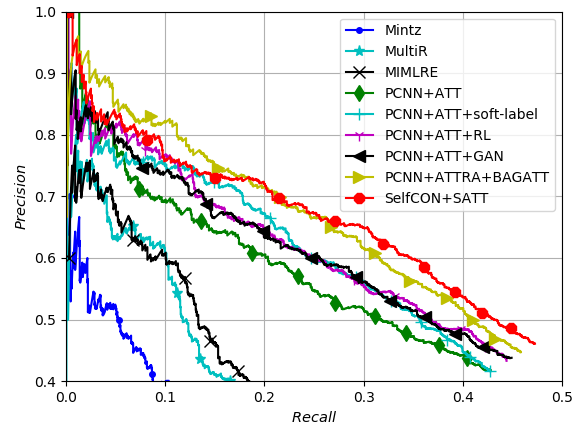


图4.3 SelfCON+SATT和8种对比模型的*PR*曲线

表4.5 SelfCON+SATT和3种对比模型的*AUC*值

|  |  |
| --- | --- |
| 模型 | *AUC*值 |
| PCNN+ATT+RL[59] | 0.281 |
| PCNN+ATT+GAN[60] | 0.281 |
| PCNN+ATTRA+BAGATT[29] | 0.312 |
| SelfCON+SATT | **0.320** |

##### 分析讨论

由图4.3的实验结果可知：(1) SelfCON+SATT模型与其他对比模型相比整体上具有更好的*PR*性能。(2) SelfCON+SATT模型的整体性能更加平衡。在*Recall*较小时，PCNN+ATTRA+BAGATT模型[29]表现更好，但是在*Recall*大于0.18之后，该模型的*Precision*相比于SelfCON+SATT模型下降更快。表明SelfCON+SATT模型具有更平衡的整体性能。

从表4.5的实验结果可知，SelfCON+SATT模型取得最高的*AUC*值0.320，与PCNN+ATT+RL和PCNN+ATT+GAN相比*AUC*值提升3.9%，与PCNN+ATTRA+BAGATT相比提升0.8%。

综上所述，提出的利用示例间相关性的远程监督实体关系抽取方法是有效的，能够利用到更多正确标记示例的监督信息，从而提升了关系分类器的性能。

### 自选择注意力有效性实验

#### 实验过程和参数

共实现7种模型验证示例自选择注意力模块的有效性，包括CNN+ATT、CNN+SATT、PCNN+ATT、PCNN+SATT、PCNN+ATTRA+BAGATT[29]、SelfCON+ATT、SelfCON+SATT。CNN表示使用卷积神经网络作为编码器，PCNN表示使用分段卷积神经网络作为编码器，ATT表示使用Lin等[21]提出的选择注意力方法，SelfCON表示使用卷积和自注意力机制作为编码器，SATT表示示例自选择注意力模块。实验过程包括3个部分：第一部分，实现CNN+ATT和CNN+SATT模型，绘制这2种模型的*PR*曲线，计算这2种模型的完整*PR*曲线的*AUC*值；

第二部分，实现PCNN+ATT和PCNN+SATT模型，绘制这2种模型的*PR*曲线，计算这2种模型的完整*PR*曲线的*AUC*值；

第三部分，实现SelfCON+ATT和SelfCON+SATT模型，直接运行PCNN+ATTRA+BAGATT[29]模型的源代码获得*PR*曲线的数据，绘制这3种模型的*PR*曲线，计算这3种模型的完整*PR*曲线的*AUC*值。

#### 实验结果和结论

##### 实验结果

CNN+ATT和CNN+SATT模型的*PR*曲线如图4.4所示。

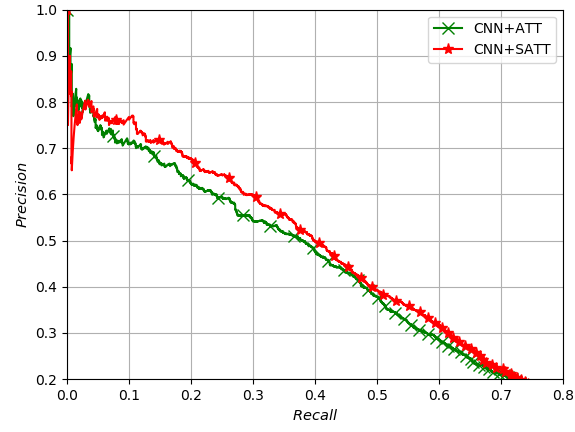


图4.4 CNN+ATT和CNN+SATT模型的*PR*曲线

PCNN+ATT和PCNN+SATT模型的*PR*曲线如图4.5所示。

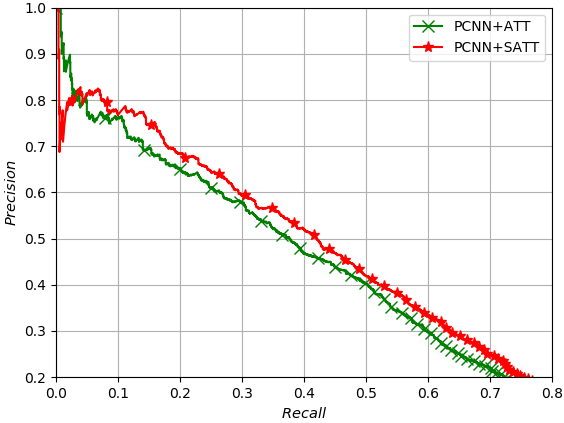


图4.5 PCNN+ATT和PCNN+SATT模型的*PR*曲线

SelfCON+ATT、SelfCON+SATT和PCNN+ATTRA+BAGATT[29]模型的*PR*曲线如图4.6所示。

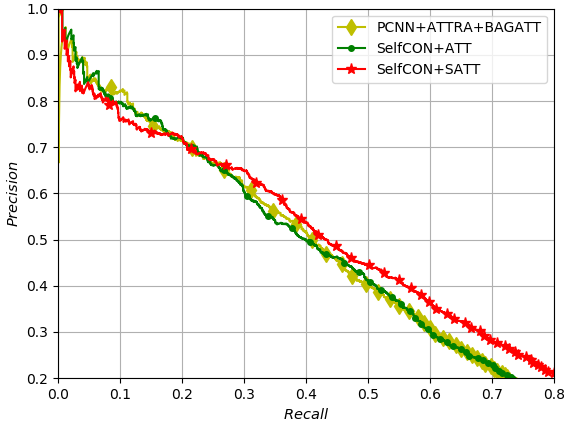


图4.6 SelfCON+SATT和对比模型的*PR*曲线

这7种模型的完整*PR*曲线的*AUC*值如表4.6所示。

表4.6 不同模型的完整*PR*曲线的*AUC*值

|  |  |
| --- | --- |
| 模型 | *AUC*值 |
| CNN+ATT | 0.384 |
| CNN+SATT | 0.405 |
| PCNN+ATT | 0.399 |
| PCNN+SATT | 0.424 |
| PCNN+ATTRA+BAGATT[29] | 0.422 |
| SelfCON+ATT | 0.426 |
| SelfCON+SATT | **0.453** |

##### 分析讨论

通过图4.4、图4.5、图4.6和表4.6的实验结果可知：

1. 利用示例间的相关性能够更加充分的利用正确标记示例的监督信息。CNN+SATT模型与CNN+ATT模型相比，*AUC*值提升2.1%，PCNN+SATT模型与PCNN+ATT模型相比，*AUC*值提升2.5%，SelfCON+SATT模型与SelfCON+ATT模型相比，*AUC*值提升2.7%。实验结果证明SATT方法要优于ATT方法。SATT方法利用包内示例间的相关性计算注意力权重，可以为所有正确标记示例学习到较大的权重，能够关注包内所有正确标记示例。
2. SelfCON+SATT模型取得最好的*AUC*值0.453，与对比模型PCNN+ATTRA+BAGATT相比，*AUC*值提升3.1%。
3. 使用卷积和自注意力机制组成的编码器比PCNN和CNN更有效。

综上所述，提出的利用示例间相关性的远程监督实体关系抽取方法利用包内示例间的相关性能够更加充分的利用所有标记正确示例的监督信息，从而提高了关系分类器的性能。

### 实例分析实验

#### 实验过程和参数

为进一步验证示例自选择注意力模块的有效性，随机选择一个包，使用自选择注意力模块（SATT）和传统的选择注意力方法（ATT）计算该包内所有示例的权重，实验过程如下：

1. 从训练集中随机选择1个示例包；
2. 使用SelfCON+SATT模型计算该包内示例的权重分布，并记录结果；
3. 使用SelfCON+ATT模型计算该包内示例的权重分布，并记录结果。

#### 实验结果和结论

##### 实验结果

使用SATT和ATT方法计算得到的包内示例的权重分布如表4.7所示，在该包实例中，包对应的两个实体为“chingy”和“st.\_louis”，这两个实体对应的关系类型为“*/people/person/place\_lived* ”，第3个示例是一个噪声示例，因为该示例不能表达实体“chingy”和实体“st.\_louis”之间的关系“*/people/person/place\_lived* ”。

表4.7 使用SATT和ATT方法计算得到的一个包内示例的权重分布

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 示例 | 噪声 | ATT | SATT |
| 1 | it sounds a little like the **st.\_louis** rapper **chingy**'s mega-single ''right thurr''. | 否 | 0.980 | 0.356 |
| 2 | chingy after a string of pop hits, the **st.\_louis** rapper **chingy** split with his former label boss, ludacris. | 否 | 0.003 | 0.353 |
| 3 | it's a good time for **st.\_louis** thanks to the doors that nelly and **chingy** and j-kwon were able to open. | 是 | 0.017 | 0.291 |

##### 分析讨论

由表4.7中的实验结果可知，通过ATT方法计算得到的噪声示例的权重大小为0.017，正确标记示例的权重大小分别为0.980和0.003。可以发现ATT方法在减少噪声示例权重的同时只能关注某一个正确标记示例而忽略了其他标记正确的示例，导致正确标记示例的监督信息丢失。SATT方法计算得到的正确标记示例的权重分别为0.356和0.353。因此SATT方法可以更加充分利用所有正确标记示例的监督信息，有助于提升关系分类器的性能。

## 小结

针对选择注意力模型将包内示例视为相互独立使得信息利用不充分导致关系分类器精确率低的问题，提出一种利用示例间相关性的远程监督实体关系抽取方法。该方法将远程监督实体关系抽取任务视为一种多示例学习任务，首先使用一层卷积加自注意力机制组成的编码器对包内的每个示例进行编码得到示例的向量表示；然后在包内使用示例自选择注意力模块为包内的示例动态的分配权重，并将包内的示例向量加权求和得到包向量表示；最后将包向量表示输入Softmax函数进行关系分类。

在NYT-MORE数据集上进行实验，结果表明，该方法不依赖目标关系，仅利用包内示例的相关性可以为正确标记示例分配较高的权重，能够关注到所有正确标记的示例，提高了关系分类器的性能。