探索用于事件提取和生成的预训练语言模型

杨森†, 冯大为†, 乔林波, 阚志刚, 李东生‡
国防科技大学，长沙，中国
{yansen_yang,linbo.qiao,kanzhigang13 @nudt.edu.cn
davyfeng.c@gmail.com, lds1201@163.com

摘要

传统的ACE事件提取方法通常依赖于人工标注的数据，而这些数据的创建往往很费力，而且规模有限。因此，除了事件提取本身的难度之外，训练数据的不足也阻碍了学习过程。为了促进事件提取，我们首先提出了一个事件提取模型，以克服角色重叠的问题，在角色方面对论据进行分离预测。此外，为了解决训练数据不足的问题，我们提出了一种方法，通过编辑原型自动生成标记数据，并通过质量排名筛选出生成的样本。在ACE2005数据集上的实验表明，我们的外牵引模型可以超过大多数现有的外牵引方法。此外，结合我们的生成方法还表现出进一步的重大改进。它在事件提取任务上获得了新的最先进的结果，包括将触发器分类的F1得分推到81.1%，将参数分类的F1得分推到58.9%。

1 简介

对于许多NLP应用来说，事件提取是一项关键而具有挑战性的任务。它的目标是检测事件的触发和参数。图1展示了一个包含由"*会议*"触发的*Meet*类型的事件的语料，有两个参数。"*布什总统*"和"*几个阿拉伯领导人*"，这两个人都扮演着"*实体*"的角色。

在事件中，有两个有趣的问题需要更多的努力。一方面，事件中的角色在频率上有很大的不同（图2），而且它们在某些词上会有重叠。

活动类型。 *会议* [实体]
句子： ***布什总统*** 将要***会见***[触发器] 与***几位阿拉伯领导人***[实体]

图1：句子中突出了一个*Meet*类型的事件，包括一个触发器和两个参数。

甚至共享同一个论点（**角色重叠问题**）。例如，在句子"*爆炸杀死了炸弹客和三个购物者*"中，"*杀死*"触发了一个*攻击*事件，而参数"*炸弹客*"同时扮演着"*攻击者*"和"*受害者*"的角色。在ACE2005数据集中有大约10%的事件（Doddington等人，2004）存在角色重叠的问题。然而，尽管有证据表明存在角色重叠问题，但很少有人注意到它。相反，在许多方法的评估设置中，它经常被简化。例如，在以前的大多数工作中，如果一个参数在一个事件中同时扮演多个角色，只要预测击中了其中的任何一个，模型就能正确分类，这显然远远不能适用于现实世界。因此，我们设计了一个有效的机制来解决这个问题，并在实验中采用更严格的评价标准。

另一方面，到目前为止，大多数基于深度学习的事件提取方法都遵循监督学习的范式，这需要大量的标记数据进行训练。然而，准确标注大量的数据是一项非常艰巨的任务。为了缓解现有方法在预定义事件数据方面的不足，事件生成方法经常被用来产生额外的事件进行训练（Yang等人，2018；Zeng等人，2018；Chen等人，2017）。而远距离监督（Mintz等人，2009）是为此常用的技术，用于标记- ing外部语料。但质量和数量

†这两位作者贡献相同。
‡通讯作者。

计算语言学协会第57届年会论文集》，第5284-5294页 意大利佛罗伦萨，2019年7月28日至8月2日。Qc 2019年计算语言学协会

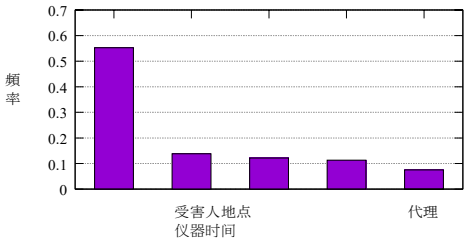

图2：在ACE2005数据集中，出现在*Injure*类型的事件中的角色频率。

在远距离监督下生成的事件，高度依赖于源数据。事实上，外部语料库也可以被预训练的语言模型利用来生成句子。因此，我们转向预训练的语言模型，试图利用它们从大规模语料库中学到的知识来生成事件。

具体来说，本文提出了一个基于预训练语言模型的框架，其中包括一个事件提取模型作为我们的基线和一个标记的事件生成方法。我们所提出的事件提取模型由一个触发器提取器和一个参数提取器组成，参数提取器参考前者的结果进行推理。此外，我们通过对基于角色重要性的损失函数的重新加权来提高参数提取器的性能。

预训练的语言模型也被应用于生成标记的数据。受Guu等人（2018）的工作启发，我们将现有样本作为事件生成的原型，其中包含两个关键步骤：参数替换和附属标记重写。通过对生成的样本的质量进行评分，我们可以挑选出那些高质量的样本。将它们与现有的数据结合起来可以进一步提高我们的事件提取器的性能。

2 相关工作

事件提取

就分析粒度而言，有文档级事件提取（Yang et al. ，2018）和句子级事件提取（Zeng et al.）我们在本文中主要讨论后者的统计方法。这些方法可以进一步分为两个详细的类别：基于特征的方法（Liao and Grishman，2010；Liu et al.，2010；Miwa et al.，2009；Liu et al.，2016；Hong et al.，2011；Li et al.，2013b）和基于神经的方法，利用神经网络的优势自动学习特征（Chen et al. , 2015; Nguyen and Grishman, 2015; Feng et al. , 2016)。

事件生成 外部资源，如Freebase、FrameNet和WordNet，通常被用来生成事件并丰富训练数据。以前的一些事件生成方法（Chen等人，2017；Zeng等人，2018）。

基于远距离监督的一个强有力的假设[1]来标记无监督语料库中的事件。但事实上，共同出现的实体可能没有预期的重新关系。此外，Huang等人（2016）在提取事件时加入了抽象的意义表示和分布语义。而Liu等人（2016，2017）设法从FrameNet中的框架中挖掘额外的事件。

预训练的语言模型

预训练的语言模型能够在考虑到语境的情况下动态地捕捉单词的含义。McCann等人（2017）在目标任务中利用预训练的语言模型来监督翻译语料库。ELMO（Embeddings from Language Models）（Peters等人，2018）通过用堆叠的双向LSTM（Long Short Term Memory）和残余结构对字符进行编码，获得对语境敏感的嵌入（He等人，2016）。Howard和Ruder（2018）在文本分类上获得了相似的再结果。GPT（生成预训练）（Radford等人，2018）在12个任务中的9个任务中提高了技术水平。BERT（Bidirectional Encoder Representations from Transformers）（Devlin等人，2018）打破了11项NLP任务的记录，受到了广泛的关注。

3 提取模式

本节描述了我们提取纯文本中发生的事件的方法。我们认为事件提取是一个两阶段的任务，其中包括触发器提取和参数提取，并提出了一个基于预训练语言模型的事件提取器（**PLMEE**）。图3说明了PLMEE的结构。它由一个触发器提取器和一个参数提取器组成，两者都依赖于BERT的特征表示。

3.1 触发器提取器

触发器提取器的目标是预测一个标记是否触发了一个事件。因此，我们将触发器的提取表述为一个标记级的分类任务，其标签为

[1]如果两个实体在知识库中具有某种关系，那么所有提到这两个实体的句子都会表达这种关系。

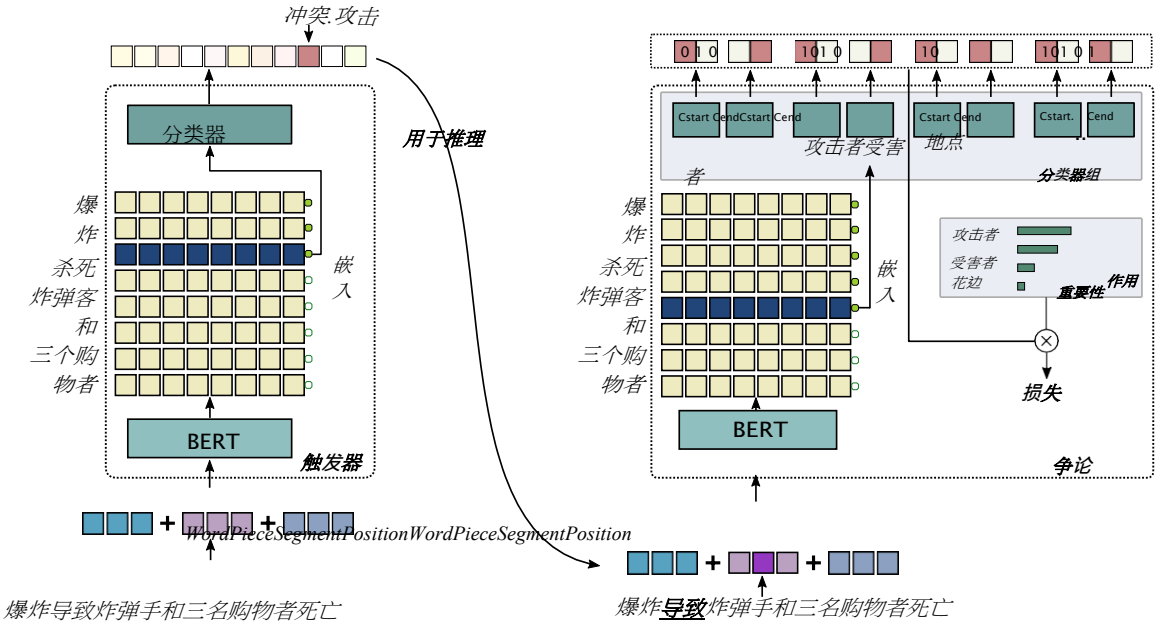

图3：PLMEE架构的说明，包括一个触发器提取器和一个参数提取器。图中还显示了由 "杀 "字触发的事件实例的处理过程。

是事件类型，只需在BERT上添加一个多分类器来构建触发器提取器。

触发器的输入遵循BERT，即三类嵌入的总和，包括WordPiece嵌入（Wu等人，2016），位置嵌入和段嵌入。由于输入只包含一个句子，所以它的所有片段id都被设置为零。此外，to- ken [CLS] 和 [SEP][2] 被放置在句子的开始和结束。

在许多情况下，触发器是一个短语。因此，我们将具有相同预测标签的连续标记视为整个触发器。一般来说，我们采用交叉熵作为微调的损失函数。

3.2 论据提取器

给定触发器后，参数提取器的目的是要将

牵引相关的论点和它们发挥的
与触发器的提取相比，参数的提取更为复杂，因为有三个问题：参数对触发器的依赖性，大多数参数是长名词短语，以及角色重叠问题。我们正是采取了一系列行动来处理这些障碍。

与触发器提取器一样，参数提取器也需要三种嵌入物。然而，它需要知道哪些标记构成了触发器。因此，我们给参数提取器提供的是触发器标记的片段id是一个。

为了克服参数提取中的后两个问题，我们在BERT上增加了多组二进制分类器。每一组分类器对一个角色进行断裂，以确定扮演该角色的所有论据的跨度（每个跨度包括一个开始和一个结束）。这种方法类似于SQuAD上的问题回答任务（Rajpurkar等人，2016），其中只有一个答案，而扮演同一角色的多个参数可以同时出现在一个事件中。由于预测是用角色分开的，一个论据可以扮演多个角色，而一个标记可以属于不同的论据。因此，角色重叠的问题也可以得到解决。

3.3 论证跨度的确定

在PLMEE中，一个标记t被预测为扮演角色r的参数的开始，其概率为1。

$$P^r(t) = Softmax(W^r \cdot B(t))。$$

所有作用。

而作为结束与概率。

$$P_e^r(t) = Softmax(W_e^r \cdot B(t))。$$

其中我们用下标 "s "表示 "开始"，下标 "e "表示 "结束"。W^r 是旨在检测扮演r角色的参数的开始的二元分类器的权重，而W^r 是另一个旨在检测结束的二元分类器的权重。B 是BERT嵌入。

的0和1，根据 r r

对于每个角色r，我们可以得到两个列表B^r和 B^r

签署

$_{Ps}$和$_{Pe}$.它们表明分别是句子中的一个标记是否是

[2][CLS]、[SEP]和[MASK]是BERT的特殊标记。

扮演角色r*的参数的*开始或结束。[3].Al- gorithm 1被用来依次检测每个标记，以确定扮演r角色的所有参数的跨度。

算法1 论据跨度的确定

在。P_s^r 和 P_e^r，B_s^r 和 $B^r{}_e$句子长度 l.
Out:扮演角色r的论据的跨度列表L **启动**：$A_s \leftarrow$-1, $A_e \leftarrow$-1
1: **对于**$i \leftarrow 0$到l**做**
2: **如果**在状态1**和**i^{th} token**是一个开始，那么**
3: $a_s \leftarrow i$，并改变为状态2
4: **结束 如果**
5: **如果**在状态2，**那么**
6: **如果**i^{th} token**是一个新的开始，那么**
7: $a_s \leftarrow i$ **如果**$P^r_s[i] > P^r_s[a_s]$
8: **结束 如果**
9: **如果**i^{th} token**是一个结束，那么**
10: $a_e \leftarrow i$并改变为状态3
11: **结束，如果**
12: **结束 如果**
13: **如果**在状态3，**那么**
14: **如果**i^{th} token**是一个新的结束，那么**
15: $a_e \leftarrow i$ **如果**$P^r_e[i] > P^r_e[a_e]$
16: **结束 如果**
17: **如果**i^{th} token**是一个新的开始，那么**
18: 将$[a_s, a_e]$添加到L*中*。
19: $a_e \leftarrow$ -1, $a_s \leftarrow i$并改变为状态2
如果 **结束**
21: **结束**
如果
22: **结束**

算法1包含一个有限状态机，它在对B^r 和B^r的重新响应中从一个状态变化到另一个状态。有三个状态：IEF定义为按对数比例计算的逆裂纹。理。1) 没有检测到开始或结束。2）只检测到一个开始；3）检测到一个开始以及一个结束。特别是，状态根据以下规则变化。当当前标记是一个开始时，状态1变为状态2；当当前标记是一个结束时，状态2变为状态3；当当前标记是一个新开始时，状态3变为状态2。值得注意的是，如果已经有了一个开始，又出现了另一个开始，我们将选择概率较高的那个，对于结束也是如此。

输出概率和黄金标签$\boldsymbol{y}$之间的熵。

$$L_s = \frac{1}{|R| \times |S|} \sum_{r \in R} CE(P^r, \boldsymbol{y}^r)。$$

其中CE是交叉熵，是角色集，是输入句子是，其中的标记数。同样地，我们把e定义为损失所有检测器的二元分类器的功能。

$$L_e = \frac{1}{|R| \times |S|} \sum_{r \in R} CE(P^r_e, \boldsymbol{y}^r)。$$

我们最后平均L_s 和L_e，作为argument extractor的损失。

如图2所示，在角色之间存在着很大的自由度差距。这意味着角色在事件中具有不同程度的"重要性"。这里的"重要性"是指一个角色表示特定类型事件的能力。例如，"*受害者*"这个角色比"*时间*"这个角色更有可能表示一个*死亡*事件。受此启发，我们根据角色的重要性重新对L_s和L_e，并建议用以下定义来衡量重要性。

角色频率（RF）
我们将RF定义为角色r出现在类型v的事件中的频率。

$$\text{射频}(r, v) = \frac{v}{\sum_{k \in Rv} N^k_v},$$

其中，N^r_v是在类型v的事件中出现的角色r的数量。

反向事件频率（IEF）作为一个角色的普遍重要性的衡量标准，我们去掉了

3.4 损失重计

含有角色r的事件类型的概念。

$$\text{IEF}(r) = \log \frac{|V|}{|\{v \in V : r \in v\}|},$$

其中V是事件类型的集合。

最后我们把RF-IEF作为RF和IEF的乘积。RF-IEF(r, v) = RF(r, v) IEF(r)。随着通过RF-IEF，我们可以衡量一个角色的重要性r在类型v的事件中。

$$I(r, v) = \frac{\text{expRF-IEF}(r,v)}{\sum_{r' \in R} \text{expRF-IEF}(r',v)}。$$

我们选择了三种事件类型，并在表1中列出了每种类型中最重要的两种角色。这表明，尽管可能有多种角色

我们最初定义L_s 作为所有的损失函数二元分类器，负责检测参数的开始。它是交叉的平均数。

[3]如果$B^r[i]=1$，则i^{th} token是一个开始，如果$B^r_{签署}[i]=1$，则是一个结束。

事件类型	前2个角色	总数
运输(15)	神器，起源	0.76
攻击(14)	攻击者，目标	0.85
模具(12)	受害者，代理人	0.90

表1：每个事件类型的前两个角色和他们的重要性总和。事件类型后面括号里的数字是在其中出现过的角色的数量。

在某人类型的事件中，只有少数人是不可或缺的。

给予输入的事件类型v，我们重新权衡L_s和L_e，基于每个角色在v中的重要性。

$$L_s = \frac{I(r,v)}{|S|} \mathrm{CE}(P^r, y_s^r)_s$$

$$L_e = \frac{I(r,v)}{|S|} \mathrm{CE}(P^r, y_{ee}^r)。$$

论据提取器L的损失仍然是平均水平。
L_s 和 L_e 的年龄。

4 训练数据的生成

除了PLMEE之外，我们还提出了一种基于预先训练的语言模型的事件生成方法，如图4所示。通过编辑原型，该方法可以生成一个数量可观的标记样本作为额外的训练语料。它由三个阶段组成：预处理、事件生成和评分。

为了方便生成方法，我们将附属标记定义为句子中除触发器和论据之外的标记，不仅包括单词和数字，还包括标点符号。以图1中的句子为例，"*是*"和"*去*"是辅助标记。很明显，辅助标记可以调整表达的流畅性和多样性。因此，我们试图在保持触发器和参数不变的情况下，改写它们以扩大生成结果的多样性。

4.1 预处理

有了黄金标签，我们首先收集ACE2005数据集中的论据以及它们所扮演的角色。然而，那些与其他参数重叠的参数被排除在外。因为这样的参数是十个长的复合短语，包含了太多的意外信息，将它们纳入参数替换可能会带来更多不必要的错误。

我们还采用BERT作为目标模型，在下一阶段重写附属标记，并且在ACE2005数据集上用屏蔽语言模型任务对其进行微调（Devlin等人，2018），使其预测偏向数据集的分布。与BERT的预训练程序相同，每次我们都会对一批感性的句子进行采样，并屏蔽15%的标记。
它的目标仍然是在没有监督的情况下预测出正确的标记。

4.2 事件生成

为了生成事件，我们在一个原类型上进行两个步骤。首先，我们将原类型中的参数替换为那些扮演过相同角色的类似参数。接下来，我们用微调的BERT重写附属标记。通过这两个步骤，我们可以发现

保留一个带有注释的新句子。

争论替换 第一步是重新

在事件中放置论据。参数都是
被取代，而新的人应该扮演过同样的角色。虽然角色是继承的
替换后，所以我们仍然可以对生成的样本使用原点标签。

为了不大幅度改变意义，我们采用相似性作为选择新论据的标准。它基于以下两点考虑：一是扮演相同角色的两个论据在语义上可能有很大的差异；二是一个论据所扮演的角色在很大程度上取决于它的语境。
因此，我们应该选择那些在语义上相似且与上下文一致的论据。

我们使用嵌入之间的余弦相似度来衡量两个论据的相似度。由于ELMO有能力处理OOV的问题，我们采用它来嵌入论据。

$$\mathrm{EE}(a) = \frac{1}{|a|} \sum_{t \in a} (t),$$

其中a是参数，是ELMO嵌入。我们选择前10%最相似的论据作为候选人，并对其相似性使用softmax操作来分配概率。

一个参数被替换为80%的概率，同时保持20%的概率不变，以偏向实际事件的表述（Devlin等人，2018）。请注意，触发器保持不变，以避免de-pendency关系的不理想偏差。

辅助标记重写

争论替换的结果已经可以被认为是生成的数据，但是恒定的上下文可能会增加过拟合的风险。因此，为了平滑

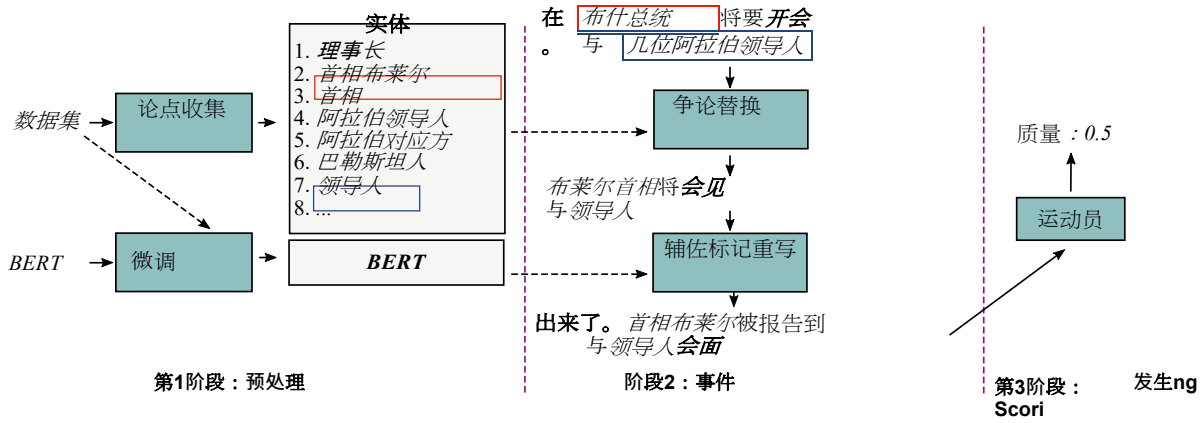

图4：生成方法的流程图。

生成的数据并扩大其多样性，我们设法用微调的BERT重写附属标记。

改写是用与当前语境更匹配的新词来替换原型中的一些附属词组。我们把它看作是一个Cloze任务（Taylor, 1953），其中一些附属标记被随机屏蔽，在第一阶段微调的BERT被用来预测基于上下文的合适标记的词汇ID。我们用一个参数m来表示需要改写的附属标记的比例。

辅助标记的改写是一个循序渐进的过程。每次我们掩盖15%的辅助标记时

(有标记[MASK])。然后，该句子被送入进入BERT，以产生新的附属标记。还没有被改写的辅助标记会

暂时保留在句中。

为了进一步说明上述两个步骤，我们在图4中给出一个实例。在这个例子中，我们将m设置为1.0，这意味着所有的附属标记将被改写。最终的输出结果是"*首相布莱尔被报告参加了与领导人的会议*"，它与原型中的原始事件有相同的标签。很明显，尽管m是1.0，但一些附属标记还是被保留了。

4.3 评分

理论上，用我们的生成方法可以生成无限多的事件。然而，并非所有的事件都对提取器有价值，有些甚至会降低其性能。因此，我们增加了一个额外的阶段，对每个生成的样本的质量进行量化，以挑选出那些有价值的后者反映了数据之间的差异。

困惑度（PPL）与对数版本的掩蔽困惑度（Devlin等人，2018）不同，我们采取的是那些平均概率的广告。

衔接标记，这些标记已被改写为生成句子S的per-plexity。

$$PPL(S)=A\frac{1}{(S)}\frac{}{P(t)}。$$

$$t\in A(S)$$

其中，A是被改写的附属标记的集合S。

距离（DIS）

我们用余弦相似度来衡量S和数据集D之间的距离。

$$邓丽君(s,d)=1-\frac{1b(s)-b(s)}{|d|\quad |b(s)|\times|b(s)|}$$

样本。我们评估质量的关键在于它与两个因素密切相关，即困惑度和与原始数据集的距离。前者反映了基因的合理性。

S∈D

与ELMO的嵌入论点不同，我们利用BERT来嵌入句子，并将第一个标记[CLS]的嵌入作为句子的嵌入。

PPL和DIS都被限制在[0,1]内。我们认为，生成的高质量样本应该同时具有低的PPL和DIS。因此，我们将质量函数定义为。

$$Q(S) = 1 - \lambda \text{PPL}\ S + (1 - \lambda)\ \text{DIS}\ S, D$$

(((

，其中λ [0, 1] ∈ 是平衡参数。这个函数用于在实验中选择高质量的生成样本。

5 实验

在这一部分，我们首先在ACE2005数据集上评估了我们的事件提取器PLMEE。然后，我们给出了一个生成样本的案例研究，并通过将它们添加到训练集中进行了真实的评估。最后，我们说明了该生成方法的局限性。

阶段 / 模型	触发器鉴定(%)			触发器钙化(%)			争论鉴定(%)			争论钙化(%)		
	P	R	F	P	R	F	P	R	F	P	R	F
交叉活动		不适用		68.7	68.9	68.8	50.9	49.7	50.3	45.1	44.1	44.6
跨越实体		不适用		72.9	64.3	68.3	53.4	52.9	53.1	51.6	45.5	48.3
最大熵	76.9	65.0	70.4	73.7	62.3	67.5	69.8	47.9	56.8	64.7	44.4	52.7
DMCNN	80.4	67.7	73.5	75.6	63.6	69.1	68.8	51.9	59.1	62.2	46.9	53.5
JRNN	68.5	75.7	71.9	66.0	73.0	69.3	61.4	64.2	62.8	54.2	56.7	55.4
DMCNN-DS	79.7	69.6	74.3	75.7	66.0	70.5	71.4	56.9	63.3	62.8	50.1	55.7
ANN-FN		不适用		79.5	60.7	68.8		不适用			不适用	
ANN-AugATT		不适用		78.0	66.3	71.7		不适用			不适用	
PLMEE(-)	84.8	83.7	**84.2**	81.0	80.4	**80.7**	71.5	59.2	64.7	61.7	53.9	57.5
PLMEE							71.4	60.1	**65.3**	62.3	54.2	**58.0**

表2：所有方法的性能。黑体字表示最好的结果。

与之前的研究一样（Li et al. , 2013b; Chen et al. , 2015; Hong et al. , 2011），我们将40个新闻报道文档作为测试集，30个其他文档作为验证集，其余529个文档作为训练集。然而，与之前的工作不同，我们采用以下标准来评估每个预知事件的正确性。

1. **只有当触**发器的跨度和类型与黄金标签相匹配时，触发器的预测才是正确的。
2. 只有当一个参数的跨度和它所扮演的所有角色与黄金标签相匹配时，该参数的预测才是正确的。

值得注意的是，一个参数的所有预测角色都需要与黄金标签相匹配，而不是只有其中一个。我们采用*预判（P）*、*召回（R）*和*F测量（F1）*作为评价指标。

5.1 事件提取的结果

我们把以前的几个经典作品进行比较，并把它们分为三类。

基于特征的方法 在*Cross event*（Liao and Gr- ishman, 2010）中，文档级的信息被用来协助事件的提取。而*Cross entity*（Hong等人，2011）在提取中使用了跨实体推理。*Max Extropy* (Li et al., 2013a)在结构化预测的基础上提取触发器和论据。

基于神经的方法*DMCNN*（Chen et al. , 2015）首先采用动态多池CNN来自动提取句子级特征。*JRNN*（Nguyen等人，2016）提出了一个联合的

基于双向RNN的事件提取框架。

基于外部资源的方法 *DMCNN-DS*（Chen等人，2017）使用FreeBase在无监督语料库中通过distance监督来标记潜在事件。*ANN-FN*（Liu等人，2016）通过从FrameNet中检测到的额外事件来提高提取率，而*ANN-AugATT*（Liu等人，2017）通过监督的注意力机制来利用参数信息，进一步提高性能。

为了验证损失再加权的有效性，我们进行了两组实验来进行比较。即，损失函数在所有分类器的输出上简单求平均值的一组（表示为**PLMEE(-)**）和根据角色重要性对损失进行重新加权的一组（表示为**PLMEE**）。

表2比较了上述模型与PLMEE在测试集上的结果。如表所示，在触发器提取任务和参数提取任务中，PLMEE(-)在所有的比较方法中取得了最好的结果。触发器提取的改进是相当明显的，在F1得分上有接近10%的大幅增长。而在论据提取方面的改进并不明显，只达到2%左右。这可能是由于我们采用了更严格的评估指标，以及论据提取任务的难度。此外，与基于特征的方法相比，基于神经系统的方法可以取得更好的性能。在比较基于外部资源的方法和基于神经系统的方法时，也出现了同样的观察结果。它表明，外部资源

原型	m	产生的事件
*布什总统*将**会见***与几位阿拉伯领导人会面*	0.2	*俄罗斯总统普京*将参加与*阿拉伯领导人的***会议**
	0.4	据报道，*总统*将与*一位阿拉伯国家的同行***会面**
	0.6	*布什先生*被传唤参加与*一些什叶派穆斯林团体*的**会议**
	0.8	*总统*正在参加与*巴勒斯坦人的***会议**
	1.0	据报道，*布莱尔首相*在与*领导人的***会晤***中*

表3：用不同比例的重写附属标记生成的样本。斜体表示参数，粗体表示触发器。

在事件提取方面，PLMEE(-
)模型对提高事件提取能力很有帮助。此外，PLMEE模型在论据提取任务上能取得更好的再结果--
在识别的F1分数上提高了0.6%，在分类上提高了0.5%--比PLMEE(-
)模型要好，这意味着重新加权损失能有效地提高性能。

5.2 案例研究

表3展示了参数m在0.2到1.0之间的原型及其生成。我们可以观察到，替换后的参数与原型中的语境相对吻合，这表明它们与语义中的原始参数相近。

另一方面，重写附属标记可以使生成的数据平滑，并扩大其二维性。然而，由于没有明确的指导，这个步骤也可能引入不可预测的噪音，使生成的数据不象预期的那样流畅。

5.3 自动评估的生成

到目前为止，主要有三个方面的生成方法可能会对提取模型的性能产生重大影响，其中包括生成样本的数量（用n来代表，表示生成大小是数据集大小的倍数），重写附属标记的比例m，以及生成样本的质量。前两个因素在生成过程中是可控的。特别是，我们可以重用一个原型，并通过基于相似性的重新放置得到各种参数的组合，这将为重写附属标记带来不同的语境。此外，改写的附属标记的比例也可以调整，从而产生进一步的变化。尽管生成的质量不能被任意控制，但可以通过评分函数Q来量化，这样就可以把那些质量较高的样本挑出来，加入到训练集中。随着Q中λ的变化可以用不同的选择策略来筛选出生成的样本。

我们首先通过网格搜索对开发集的前两个参数进行调整。特别是，我们将m设定为0.2到1.0，间隔为0.2，并将n设定为0.5、1.0和2.0，同时在生成过程中保持其他参数不变。我们用这些参数进行了实验。通过分析结果，我们发现PLMEE在触发器提取和参数提取方面的最佳性能可以通过m=0.4和n=1.0实现。这表明，无论是生成的样本太少还是太多，都是提取的较好选择。太少的影响有限，而太多的影响会带来更多的噪音，扰乱数据集的分布。为了获得更好的提取性能，我们在下面的实验中使用这种参数设置。

我们还调查了样本选择方法的有效性，在三组不同的选择策略之间进行了比较。我们使用我们的生成方法获得了四倍于ACE2005数据集的大小，m=0.4，并挑选出其中的四分之一（n=1.0），λ分别为0、0.5和1.0。当λ为0或1.0时，完全由困惑度或距离来决定质量。我们发现，质量函数中λ=0.5的选择方法能够挑选出对促进提取性能更有利的样本。

模型	触发(%)	论据(%)
PLMEE	80.7	58.0
PLMEE(+)	**81.1**	**58.9**

表4：测试集上触发器分类和参数分类的F1得分。

最后，我们将上述生成的数据与ACE2005数据集结合起来，研究我们的生成方法在测试中的有效性。

集。在表4中，我们用PLMEE(+)表示用额外生成的样本训练的PLMEE模型。结果表明，使用我们的事件生成方法，PLMEE模型可以达到最先进的事件提取效果。

5.4 限制条件

通过比较生成的样本和人工标注的样本中的注释，我们发现我们的生成方法的一个问题是，角色可能会有偏差，因为语义可能会有很大的变化，而只有几个附属标记被重写。以图5为例。参数 "*匹兹堡* "和 "*波士顿* "所扮演的角色应该是 "*目的地* "和 "*原点*"，而不是像原型中那样的反义词。这是因为 "*从* "被替换成了 "*为*"，而 "*开车去* "被替换成了 "*从*"。

原型。 从尼亚加拉大瀑布出发，开车到多伦多，行驶85英里		
触发器	离开	
事件类型	*运动.运输*	
论点	尼亚加拉大瀑布	多伦多
角色	*原产地*	*目的地*
伪触发 前往匹兹堡，在200英里内从波士顿返回 leave ✓		
事件类型	*运动.运输*	
论点	匹兹堡 ✗	波士顿 ✗
角色	*原产地*	*目的地*

图5：其中一个生成的样本有错误的注释。

6 结论和讨论

在本文中，我们提出了一个框架，通过使用前牵引模型和生成方法的组合来促进事件提取，这两种方法都是基于预先训练的语言模型。为了解决角色重叠的问题，我们的提取方法试图在角色方面将论据预测分开。然后，它利用角色的重要性来重新权衡损失函数。为了进行事件生成，我们提出了一种新的方法，将现有事件作为原型。这种事件生成方法可以通过参数替换和附属标记重写产生可控的低龄化样本。它还得益于能够量化生成样本质量的评分机制。实验结果表明，生成的数据质量是有竞争力的，将它们与现有的语料库结合起来，可以使我们提出的事件提取器优于一些先进的方法。

另一方面，我们的工作仍有局限性。同一类型的事件往往具有相似性。而共同出现的角色往往有着紧密的关系。这些特征在我们的模型中被忽略了，但它们值得在改进提取模型方面进行更多研究。此外，尽管我们的生成方法可以控制生成样本的数量并进行高质量的过滤，但它仍然可以解决与远处超视距的角色相似的偏差问题。因此，在未来的工作中，我们将在预训练的语言模型中加入事件之间的关系和论据之间的关系，并采取有效措施来克服生成中的角色偏差问题。

鸣谢

该工作得到了国家重点研发计划（2018YFB0204300）和国家自然科学基金（61872376和61806216）的资助。

参考文献

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao.2017.用于大规模事件提取的自动标记数据生成。In *Proceed- ings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Pa- pers)*, volume 1, pages 409-419.

陈玉波，徐立恒，刘康，曾 道 坚，赵军。2015.通过动态多池卷积神经网络提取事件。在*计算语言学协会第53届年会和第7届自然语言处理国际联合会议论文集（第一卷：长篇论文）中*，第一卷，第167-176页。

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.2018.Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: 1810.04805*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel.2004.自动内容提取（ace）程序-任务、数据和评估。在*LREC*，第二卷，第1页。

冯晓成，黄立夫，唐德玉，纪恒，秦兵，刘婷。2016.一个独立于语言的神经网络用于事件检测。见*《第54届美国计算语言学年会论文集》（第二卷：短文）*，第二卷，第66-71页。

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang.2018.通过编辑原型生成句子。*Transaction s of the Association of Computational Linguistics*, 6:437-450.

何开明，张翔宇，任少卿，和孙健。2016.用于图像识别的深度残差学习。在*IEEE计算机视觉和模式识别会议论文集中*，第770-778页。

洪宇，张剑锋，马斌，姚建民，周国栋，朱巧明。2011.
使用跨实体推理来改善事件外延。在*计算语言学协会第49届年会的论文集中。人类语言技术-第一卷*，第1127-1136页。计算语言学协会。

Jeremy Howard和Sebastian Ruder。2018.用于文本分类的通用语言模型微调。见*《美国计算语言学协会第56届年会论文集》（第一卷：长篇论文）*，第328-339页。

黄立夫、泰勒-卡西迪、冯晓成、纪恒、克莱尔-R-沃斯、韩佳伟和阿维鲁普-
西尔。2016.自由事件提取和事件模式归纳。见*《第54届美国计算语言学年会论文集》（第一卷：长篇论文）*，第一卷，第258-268页。

李培峰，朱巧明，和周国栋.2013a.用话语层面的信息对中文事件提取中的论据识别和角色确定进行联合建模。In *Twenty-Third International Joint Conference on Artificial Intelli- gence*.

李琦，纪恒，和黄亮。2013b.通过具有全局特征的结构化预测进行联合事件提取。在*计算语言学协会第51届年会论文集（第一卷：长篇论文）*，第一卷，第73-82页。

Shasha Liao和Ralph Grishman.2010.使用文档级别的跨事件推理来改善事件提取。见*《计算语言学协会第48届年会论文集》*，第789-797页。计算语言学协会。

刘兵，钱龙华，王红玲，周国栋。2010.基于依赖性驱动的特征学习，从生物医学文本中提取蛋白质-
蛋白质的相互关系。在*第23届国际计算语言学会议论文集中。海报*，第757-
765页。计算语言学协会。

刘书林，陈玉波，何世柱，刘康，和赵军。2016.利用framenet改善自动事件检测。见《*第54届国际语言学协会年会论文集》（第一卷：长篇论文）*，第一卷，第2134-2143页。

刘书林，陈玉波，刘康，和赵军。2017.利用论据信息，通过监督注意机制改进事件检测。见*《第55届美国计算语言学年会论文集》（第一卷：长篇论文）*，第一卷，第1789-1798页。

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher.2017.在翻译中学习。Contextualized word vectors.In *Advances in Neural In-Form Processing Systems*, pages 6294-6305.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky.2009.在没有标记数据的情况下对关系外推的远程监督。In *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 2Volume 2*, pages 1003-1011.计算语言学协会。

Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi
Tsujii.2009.用于从多个语料库中提取蛋白质-蛋白质相互作用的丰富特征向量。In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Process- ing:Volume 1-Volume 1*, pages 121-130.Associa- tion for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman.2016.通过递归神经网络的联合事件提取。在*2016年计算语言学协会北美分会会议上。Human Lan-guage Technologies*，第300-309页。

Thien Huu Nguyen和Ralph Grishman。2015.用卷积神经网络进行事件检测和领域适应。见*《第53届计算语言学协会年会暨第7届自然语言处理国际联合会议论文集》（第2卷：短文）*，第2卷，第365-371页。

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer.2018.深度语境化的单词表达。In *Proceedings of the 2018 Confer- ence of the North American Chapter of the Associ- ation for Computational Linguistics:Human Lan-guage Technologies，第一卷（长篇论文）*，第2227-2237页。

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever.2018.通过生成性预训练提高语言理解能力。*URL https://s3- us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language under-standing paper. pdf*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy
Liang.2016.Squad:100,000多个用于机器理解文本的问题。在*2016年自然语言处理的经验方法会议上*，第2383-2392页。

Wilson L Taylor.1953."cloze程序"。一个测量可读性的新工具。*Journalism Bulletin*, 30(4):415-433.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016。谷歌的神经医学翻译系统。*arXiv preprint arXiv:1609.08144*.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao.2018.Dcfee:一个基于自动标记的训练数据的文档级中文金融事件提取系统。*ACL 2018会议论文集，系统演示*，第50-55页。

Ying Zeng, Yansong Feng, Rong Ma, Zheng Wang, Rui Yan, Chongde Shi, and Dongyan Zhao.2018.通过自动生成训练数据扩大事件提取学习规模。在*第三十二届AAAI人工智能会议上*。