

### 题目

一种嵌套命名实体识别的边界感知神经模型

### 动机

现有的两种提取嵌套实体的方法：

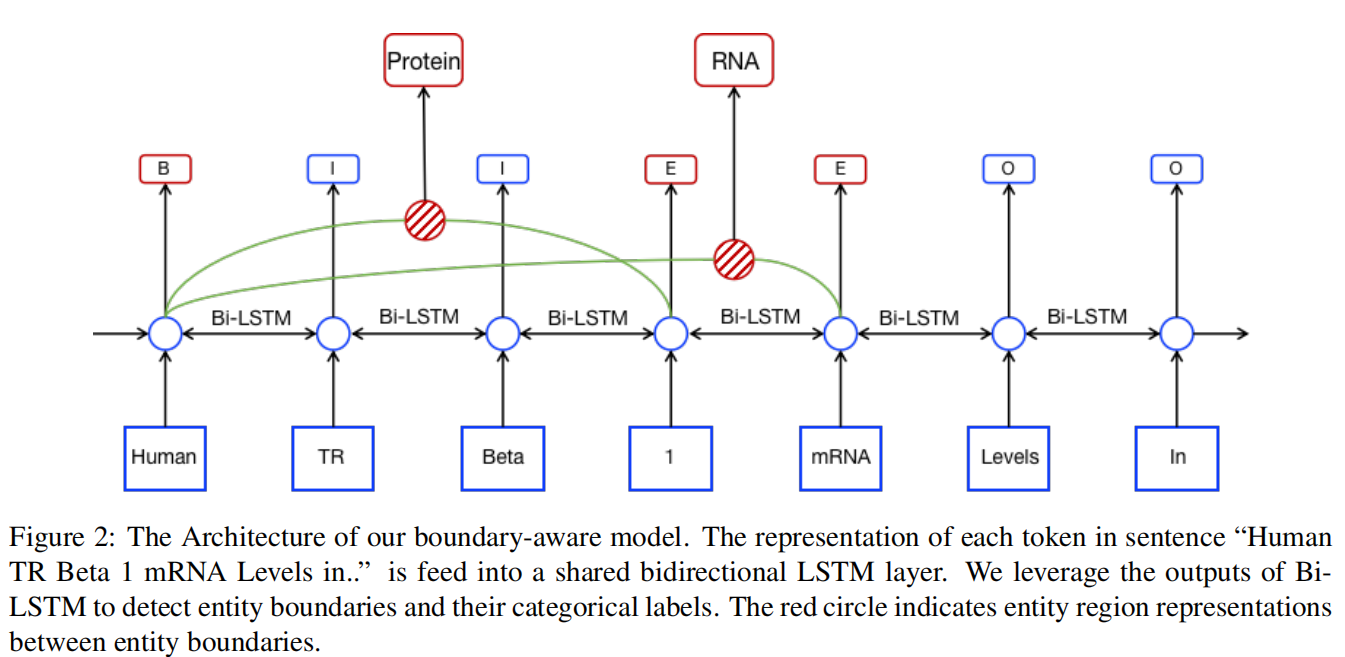
1. 分层序列标记模型将首先提取内部实体（由其他实体包含），并将它们馈送到下一层以提取外部实体。因此，该模型存在误差传播的问题。当上一层提取错误的实体时，将会影响下一层的性能。此外，当首先提取外部实体时，将不会检测到内部实体。
2. 详尽的区域分类模型枚举了句子中所有可能的区域或跨度来预测单个层中的实体。他们的方法的一个问题是忽略了显式的边界信息，导致了一些非实体的提取。我们考虑一个例子。在GENIA数据集中的一个标记序列中，“新的TH蛋白”是一个实体，而“一个新的TH蛋白”不是一个实体。然而，由于它们共享许多令牌，它们合并后的区域表示是相似的。“新颖”和“蛋白质”是实体的边界。如果没有边界信息，这两个候选区域都被提取为实体。

这两种模型可以进行互补。我们利用序列标记模型来考虑边界信息来定位实体。我们还利用区域分类模型来预测实体，而不考虑内部和外部实体的依赖关系。

### 贡献

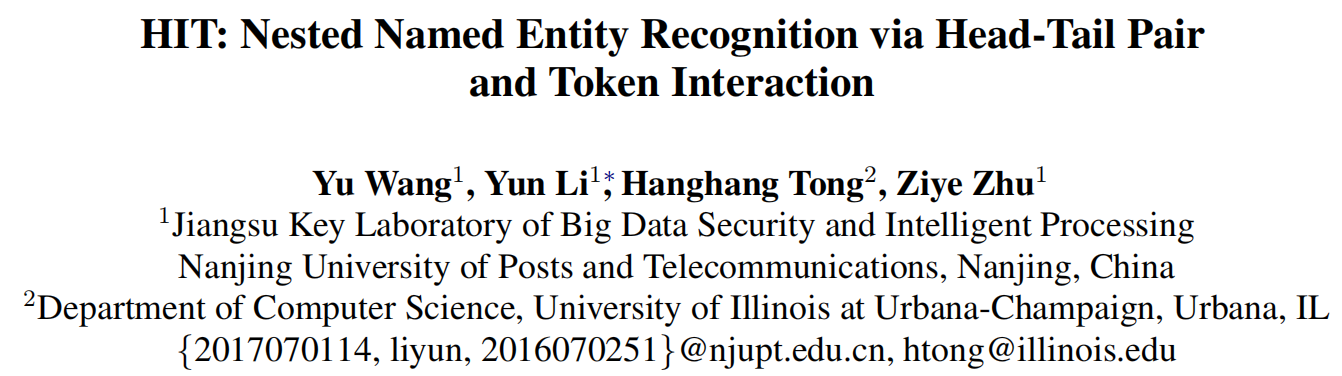
1. 我们提出了一个边界感知的神经模型，它利用实体边界来预测分类标签。我们的模型可以通过使用序列标记模型检测边界来精确地定位实体。基于检测到的边界，我们的模型利用边界相关区域来预测实体分类标签，这可以降低计算成本，缓解误差传播问题。
2. 我们引入了多任务学习来捕获实体边界及其分类标签的依赖关系，这有助于提高识别实体的性能。
3. 我们在公共嵌套的NER数据集上进行实验。实验结果表明，我们的模型优于以前最先进的方法，而且我们的模型在推理速度上要快得多。

### 模型



### 思路

在本文中，我们提出了一个边界感知神经模型，使融合序列标记模型和区域分类模型。我们应用单层序列标记模型来识别实体边界，因为嵌套实体中的标记可以共享相同的边界标签。例如，如图1所示，“人类”可以用标记B进行标记，尽管它是两个不同实体的开始。基于检测到的实体边界，我们通过对边界相关区域进行分类来预测实体分类标签。如图1所示，我们将每个带有标签B的令牌与带有标签e的令牌进行匹配。它们之间的区域被认为是候选实体。候选实体的表示将被用于分类标签的分类。



### 题目

HIT：通过头尾对和令牌交互进行的嵌套命名实体识别

### 进展

Seq2seq：(Sohrab和Miwa，2018；Ju等人，2018；Zheng等人，2019)经常利用传统的序列标记方法来学习嵌套结构。例如，Ju等人（2018）利用分层的LSTM网络从内部实体到外部实体捕获嵌套命名实体。然而，由于序列标记方法在表示嵌套结构方面的限制，这些方法仍然可能存在错误传播的问题。

基于超图：(Lu和Roth，2015；Wang和Lu，2018)介绍了用于学习嵌套命名实体的超图结构。其优点在于超边可以自然地表达嵌套结构。他们的方法的一个问题(Lu和Roth，2015)是超图的伪结构。Wang和Lu（2018）进一步提出了神经节段超图来解决这一问题。然而，如果输入句子太长或存在许多实体类别，它们的超图结构就会变得太复杂，这反过来使得这些模型的优化变得非常困难。

### 动机

基于传统的序列标注的方法和有向超图的方法，往往未能在嵌套结构的表达式能力和模型的复杂性之间取得良好的平衡。

我们提出的HIT模型利用了与（嵌套)命名实体相关的两个关键属性，包括(1)显式边界令牌和(2）边界内令牌之间的紧密内部连接。具体来说，我们设计了(1)基于多头自注意机制和双自射分类器的头尾检测器来检测边界标记，以及(2)基于传统序列标记方法的令牌交互标记器来表征边界内的内部令牌联系。在三个公共NER数据集上的实验表明，所提出的HIT达到了最先进的性能。

### 贡献

1. 我们证明了头尾对能够有效、精确地表达嵌套结构实体的边界信息。
2. 我们利用令牌交互标记器来描述边界内令牌之间的内部连接，其中我们揭示了令牌交互对识别实体有很大的影响。
3. 我们用头尾对和令牌交互序列完成实体分类，同时引入多任务损失来同时训练我们的模型。

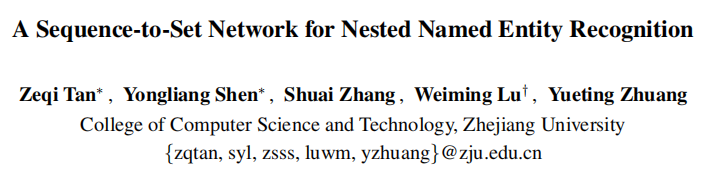
### 模型

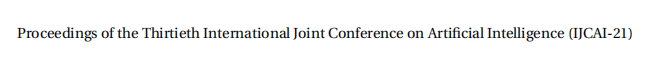
### 

### 思路

将表示序列x输入头尾检测器进行预测是否每个成对的标记都是一个实体的头尾。同时，使用令牌交互标记器根据上下文捕获相邻令牌之间的内部连接，这些联系可以指示当前令牌之前或之后的令牌是否属于一个实体。最后，利用区域分类器将头尾检测器和标记交互标记器进行集成，完成实体识别。

# Nested NER





### 题目

一种用于嵌套的命名实体识别的序列到集合的网络

### 进展

Seq-to-seq：将嵌套的NER视为一个标签被依次解码的序列生成任务。但是，在命名实体识别任务中，输出标签本质上是一个无序的集合。Strakov´a et al. [2019]依次生成对标签顺序敏感的输出标签。通过这种方式，即使模型预测了所有正确的标签，也可能由于顺序不一致而导致不合理的训练损失。

Span-based：对通过各种方法从文本序列中提取的候选跨度进行分类。 Sohrab and Miwa [2018]枚举限制长度内的所有可能的跨度，然后预测它们的类别。Zheng et al. [2019]首先分别识别左右边界，然后将它们进行匹配以形成候选跨度。ARN [Lin et al.,2019]根据先前确定的锚定词来识别感兴趣的跨度。, Yu et al. [2020]通过双仿射模型对跨度评分，并达到了最先进的水平。跨度选择方法面临着错误传播的问题，因为边界或锚点往往会被错误地识别，而跨度枚举方法需要搜索所有可能的区域。此外，这些方法中的候选跨度不会直接相互交互，因此错误地忽略了命名实体之间的依赖关系。

### 动机

基于跨度的方法（span-based）将嵌套NER作为一个跨度分类任务。优点：可以自然的处理嵌套NER；缺点：搜索空间大、缺乏实体间的交互。

为解决基于跨度的方法的问题，提出了一种新的序列到集合的神经网络。使用一个能够捕获实体之间的依赖关系的非自回归解码器在一次传递中预测最终的实体集。使用基于二分图匹配的损失函数计算整体训练损失。

### 数据集

ACE 2004, ACE 2005 and KBP 2017

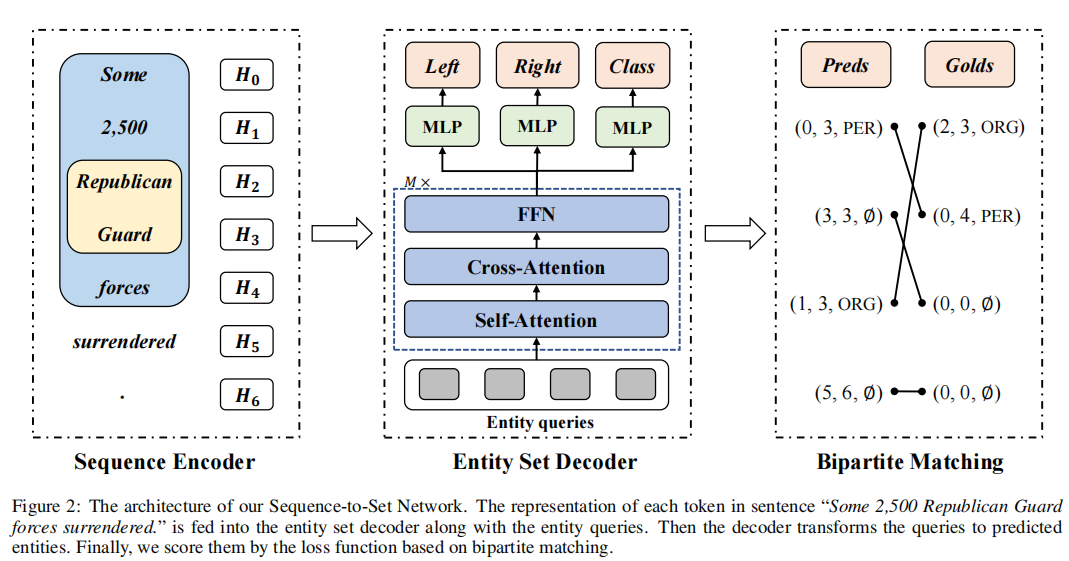
### 贡献

1：我们提出了一种新的序列到集的网络来预测实体集。据我们所知，我们是第一个将命名实体识别作为序列到集的任务。此外，我们一次预测最终的实体集，而序列到序列模型一个接一个地预测实体。由于实体本质上是无序的，我们的对标签顺序不敏感的模型获得了更好的性能。

2：我们提供了一组固定的实体查询来替换显式的候选跨度，从而不需要搜索所有可能的跨度。此外，我们能够通过使用执行实体查询之间直接交互的自注意机制来捕获实体之间的依赖关系。

3：实验结果表明，我们的模型在三个广泛使用的数据集上取得了最先进的水平，在2004年得分高0.56%，在ACE2005上领先1.65%，在KBP2017上领先2.99%。

### 模型



### 思路