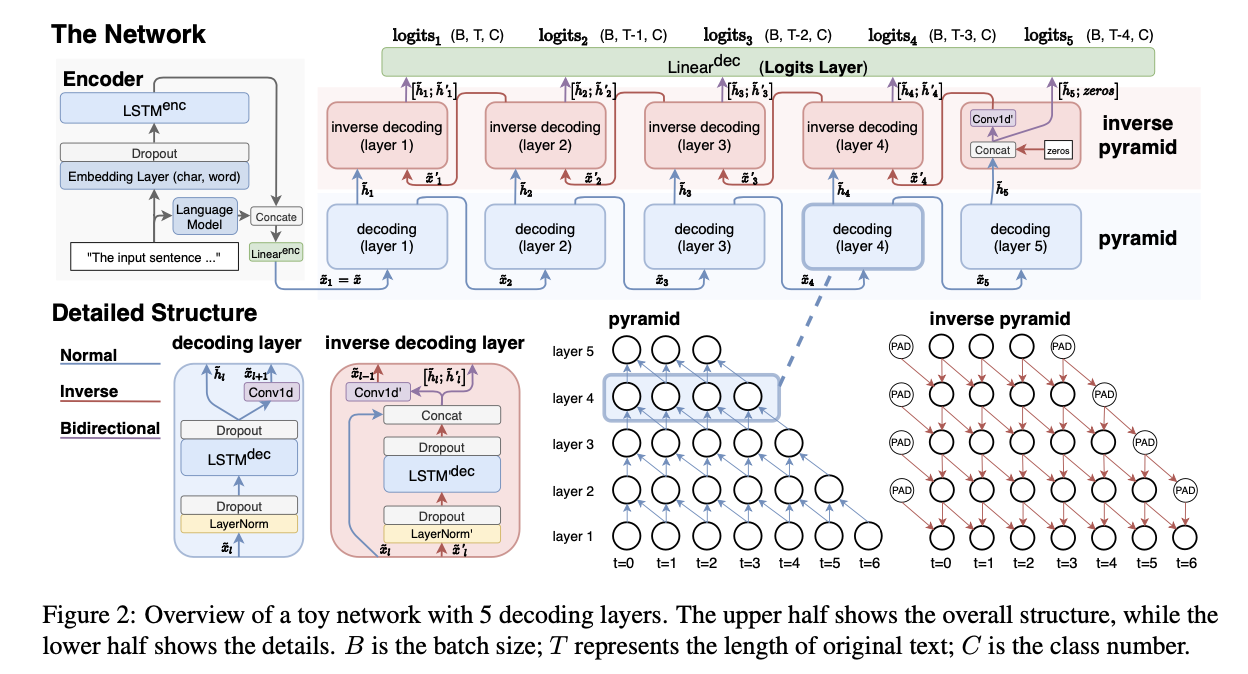
论文阅读笔记一

Pyramid: A Layered Model for Nested Named Entity Recognition

1. 贡献
2. 本文针对嵌套实体识别提出了一个金字塔分层模型，通过识别实体的长度来识别实体指称，不会出现层错位(layer disorientation)和错误传播，该模型也可以解决更通用的重叠NER任务。
3. 设计了一个逆向金字塔，允许相邻层之间进行双向交互。
4. 构建了一个包含重叠但非嵌套的实体，实验结果表明模型具备处理重叠实体的能力。
5. 方法

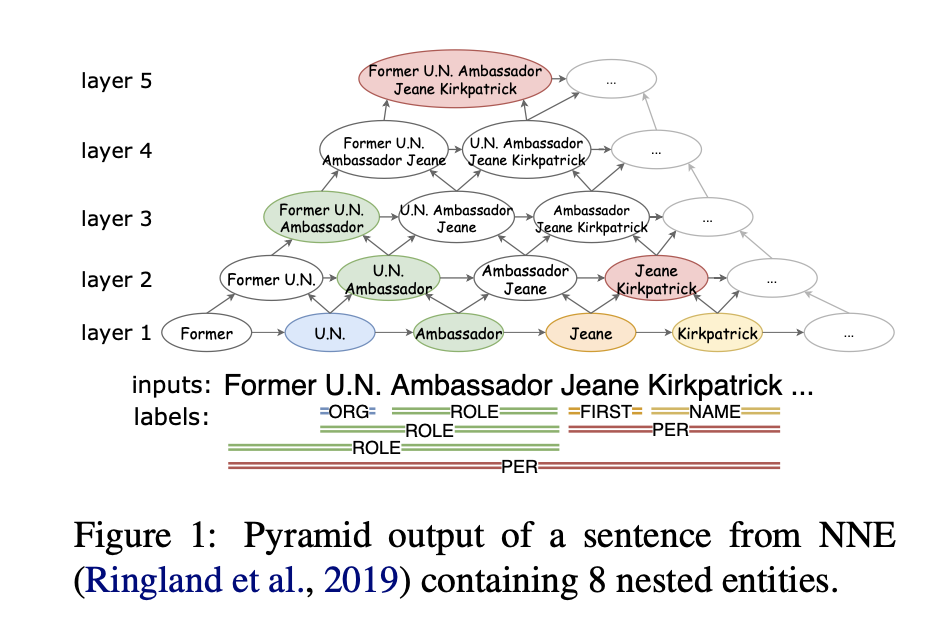
模型包括编码器、金字塔、逆向金字塔和一个logits层。



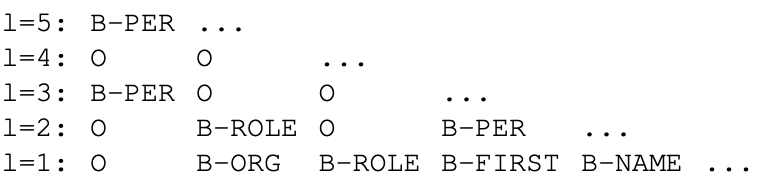
上半部分是模型整体结构，下半部分是模型实现的细节。

蓝色的是5个自下而上的解码层，粉色的是5个自上而下的逆向解码层。

1. 模型输入和输出

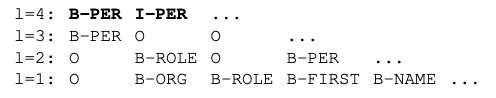


模型输入为长度为T的文本，经过编码器之后，词嵌入被送到平面实体解码层中，得到IOB2格式的L标签序列，标签序列的长度为T-1,T-2,…,T-L+1,L是解码层的个数。如下图所示，输入为Former U.N. Ambassador Jeane Kirkpatrick，这句话包括8个实体，金字塔输出将包含分层的标签序列：



这样的分层序列不能包含超过5个词的任何实体。即L层的解码器无法预测包含超过L个词的实体。所以作者提出在最底层的平面实体预测层上预测所有长度大于L的实体。就是下面L-1个层为完整的实体预测B-class标签，而顶层则可以预测B-class和I-class标签以及完整的实体，因此规定当嵌套两个实体时，如果其中一个实体长度大于L，则另一个实体长度不能大于L-1。

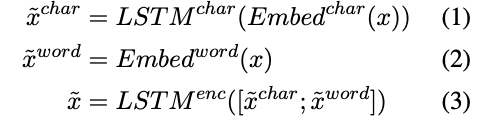
在上面的实例中，假设L=4，即共4个解码层，最长的实体Former U.N. Ambassador Jeane Kirkpatrick在第四个解码层本识别为B-PER I-PER…，具体的识别结果如下：



使用这样的修正方案可以处理比L长的食堂，但是大多数(99%)实体的长度都不超过15，而且嵌套实体的长度也很少比15还长，所以将解码层L设为16以最大程度地减小解码器个数的影响，参数L可以根据精确度和推理速度需要进行调整。

1. 编码器

作者通过将字符嵌入和单词嵌入拼接起来来表示每个单词，首先，字符嵌入通过LSTM产生，该嵌入包括单词的自省特征和形态特征，这一步有利于处理未登录OOV词汇；第二，单词嵌入通过预训练词向量进行初始化，对于未登录词，作者采用随机初始化的方式进行初始化，训练时再进行微调。这两部分词嵌入将被送到BiLSTM中进行编码提取上下文信息。

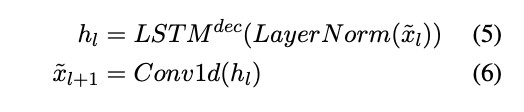


本文采用了BERT作为预训练上下文模型词向量，BERT被拼接到LSTM的输出层，通过线性层降维输出：  


1. 金字塔

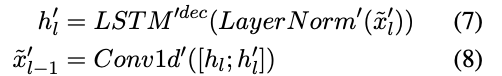
金字塔模块采用自上而下的方式识别实体。L个解码层，每层对应一个平面的实体识别器，每个解码层包括一个LSTM和一个两个内核的CNN。在第l层中，LSTM识别长度为l的实体，而CNN则整合两个相邻的隐藏层状态并将包含层信息的文本区域嵌入送到较高(l+1层)解码层中，通过l个解码层之后，每个隐藏状态实际上代表了l个原始单词的区域编码。因此，第l个解码层枚举了长度为l的文本跨度，所有这些l层的信息即可产生可能的实体跨度。

金字塔结构本质上提供了有用的归纳偏差：层越高、输入序列越短，迫使模型捕获用于预测长实体的高级信息和用于预测短实体的低级信息。此外，随着跨度表示减少到1，每个层上的预测任务都简单明了——预测在该层中表示长度为1的实体。

由于第一个解码层的输入来自编码器，而其他解码层的输入来自较低相邻层的输出，因此输入偏置和缩放因子在各层之间会有所不同，为了解决这个问题，作者将区域嵌入送到LSTM解码之前，用层归一化该嵌入进行处理：令  


1. 逆向金字塔

自下而上的金字塔每个解码层都考虑了地层的层信息，但是一层无法从较高的相邻层获得反馈，此外，对于长实体，他们的嵌入需要经过许多较低的层次，往往会丢失重要信息。故作者加入了一个逆向金字塔，以自上而下的方式识别实体，解决上述问题。在这个金字塔中，首先用CNN减少序列长度，然后送到较高的解码层。为了重构第l-1个解码层的文本区域嵌入，作者将第l个正向解码层和逆向解码层的隐藏状态串联起来，并将其送到反向CNN中，从反向解码层中得到的输出为：

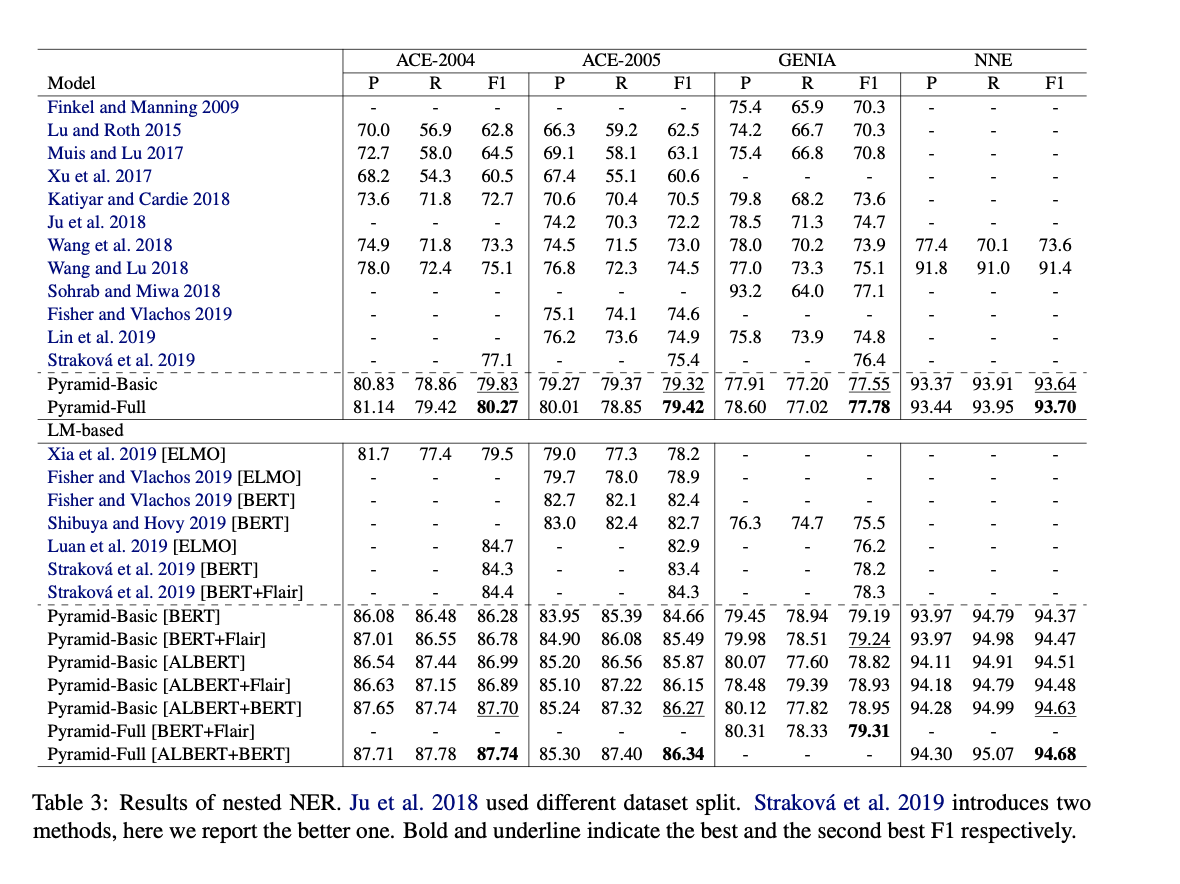


对于顶部逆向解码层，用0代替h L’。

最后结合正向解码层和逆向解码层的隐藏状态，使用前馈蹭来预测其类别：



1. 实验结果

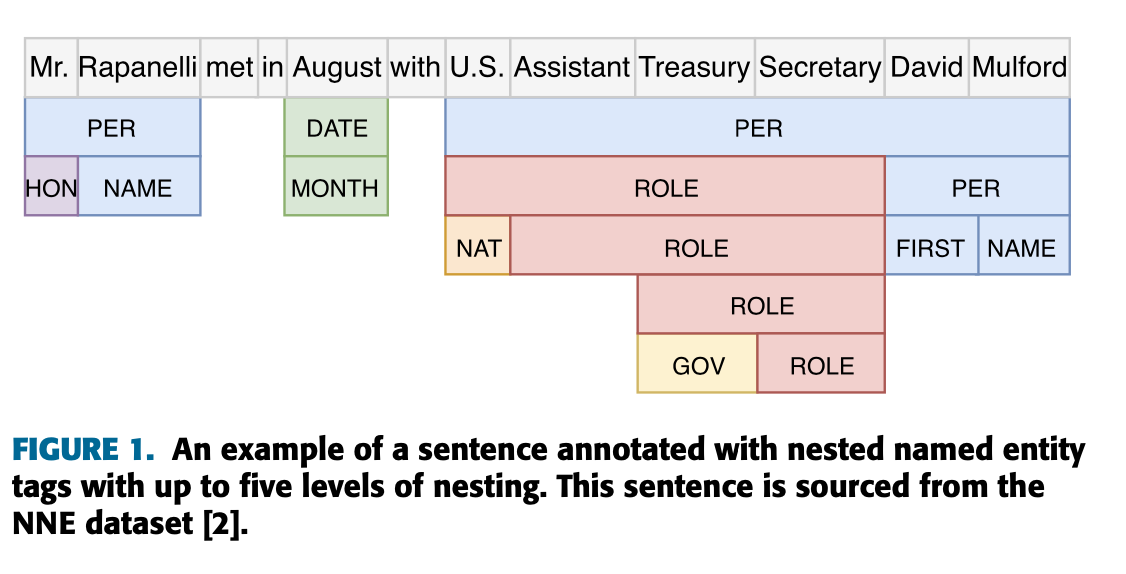


一般完整模型的时间复杂度为O(TL)，其中T为文本长度，L为解码器个数。解码器个数越少，推理速度越快，但性能较低。

论文阅读笔记二

A Bidirectional Iterative Algorithm for Nested Named Entity Recognition

1. 贡献
2. 作者提出了一种以迭代方式识别嵌套实体的神经网络模型，该模型对实体长度、类别、深度或预测输出的结构没有任何限制。
3. 引入了一种用于嵌套实体识别的双向算法，算法结合了由内而外和由外而内的两个独立训练的迭代模型，两个模型的预测结果将被合并为一个实体集。
4. 方法



1. 神经迭代模型

对于长度为n的单词序列x：



令m为输入序列x中连续k个单词的实体类型，实体标签l和置信度得分c(m)为：

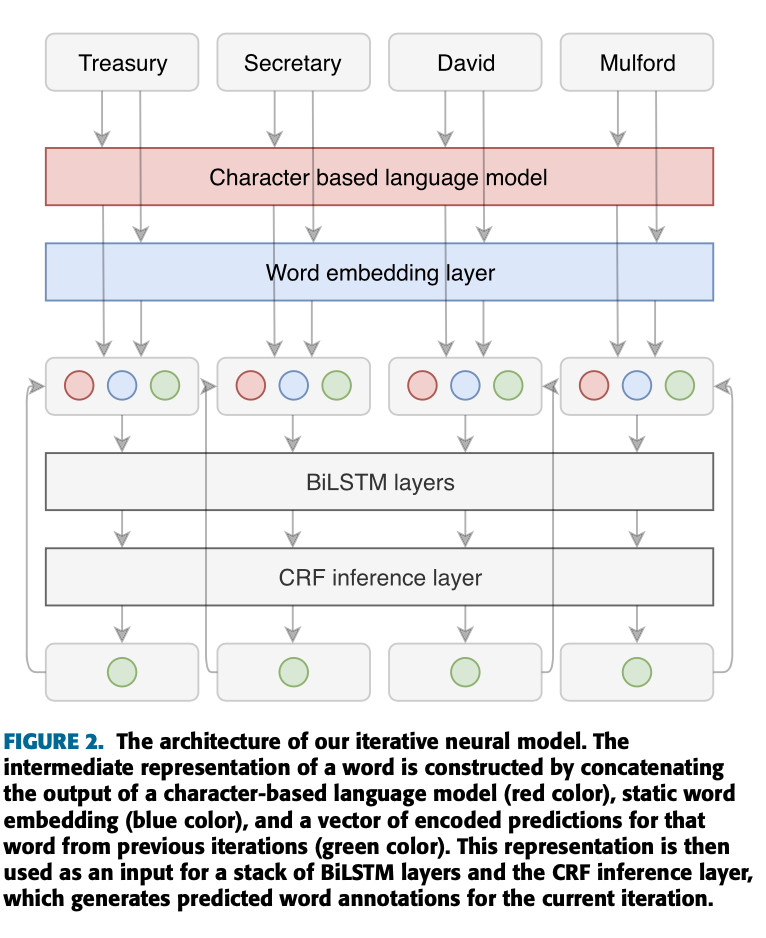


实体识别模型是为了找到并正确分类句子中所有实体，作者提出的模型通过多个步骤推理实现。

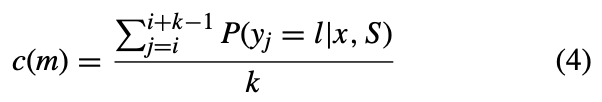
令S表示句子x中所有预测的唯一实体的集合，对给定输入x和一组之前预测得到的实体集合S，函数f可得到实体的标签y：



在第一次迭代之前，S集合为空，每次迭代过程中都加入从模型输出的所有实体，重复这个迭代过程直到找不到新的实体。最终迭代之后，S代表模型最终的预测结果。



图二是模型的简化结构，对于单词序列，作者构建了单词的向量表示，每个单词都由上下文词向量、单词词嵌入和多热向量拼接组成。多热编码是一个二进制矢量，其维数等于标签数加1，每个位置对应一个特定标签，多余的位用于标记初始迭代。这三部分作为BiLSTM的输入，用于计算上下文相关的隐藏状态。最后一个BiLSTM的输出被送到CRF推理层，该层负责预测标签y的序列，该序列使当前迭代的条件概率P(y|x,S)最大化。公式2中类型为l的实体c(m)的置信度得分定义为属于类别l的单个单词的概率的算术平均值：



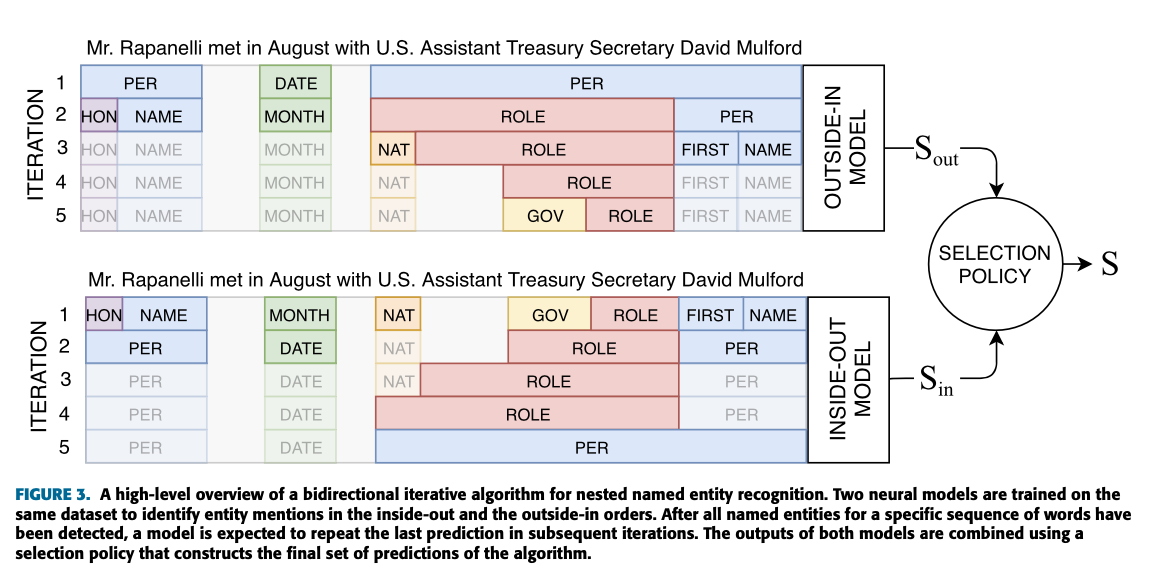
1. 训练过程

将每个嵌套的NER示例划分为几个由内而外和由外而内的层，并将每一层都视为单独的训练样本。作者统计两个标签之间的父子关系频率，并根据频率对标签进行排序。一个训练实例包括单词序列和代表先前迭代预测的向量序列。训练阶段，先前的预测序列是根据训练集中的真实标签综合创建的，仅在模型推测阶段使用模型的真实预测。通过这个过程，样本在训练过程中彼此独立，并且无需更改过程即可适应模型的迭代性质。CRF输出层与维特比解码算法一起使用，故使用句子级负对数似然函数作为损失函数。

1. 双向迭代算法

本节提出一种利用由内而外和由外而内两种模型进行训练输出结果来生成最终实体集的方法。将相同输出结构的两个视图组合在一起，可以确定单个模型无法检测到的嵌套模式。

算法核心概念如图3所示：

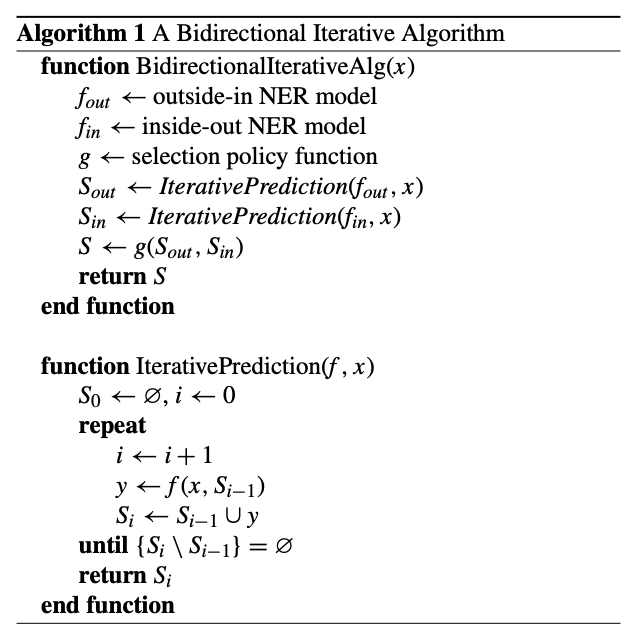


两个方向的模型都使用了前面描述的方法独立地产生他们的输出，然后由函数g处理由内而外和由外而内的两个模型的预测结果。函数g是一种选择策略，g接受两组实体指称，并产生一个过滤后的集合S，其中包括Sin和Sout中的零个或多个元素。



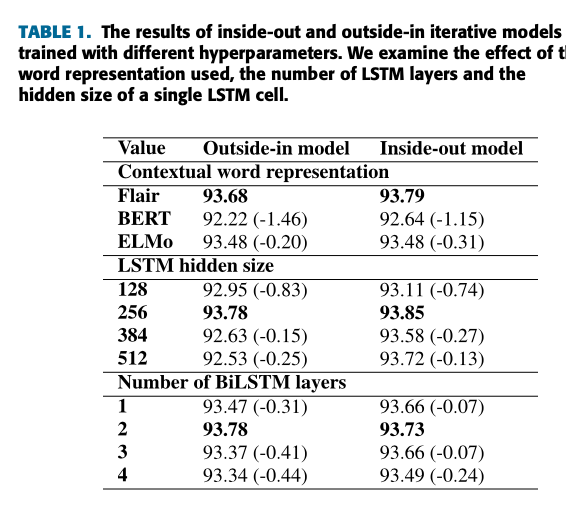
作者评估了6种选择策略对模型的影响：

1. 仅由内而外模型：基线策略，仅选择由外而内模型中的实体，等效于仅使用由内而外模型。
2. 仅由内而外模型：另一种基线策略，但仅考虑了由内而外模型的预测。
3. 模型相交：当且仅当Sin和Sout中都存在实体指称时，该指称才出现在实体集合中。
4. 模型联合：策略3取交集，本策略取并集
5. 基于概率的选择策略：仅当其置信度得分c(m)大于等于指定阈值超参数时才会接受该实体指称m，对集合Sin和Sout使用单独的阈值，使用线性搜索找到这些超参数，以使得验证集上的F1得分最大化。
6. 线性分类器选择策略：使用附加的逻辑回归模型决定接受还是拒绝m，在从训练数据集生成的预测集Sin和Sout中提取的特征列表上训练模型。特征包括Sin和Sout中m的置信度得分，Sin和Sout中m的标签，Sin和Sout中父实体的标签，Sin和Sout中父实体的置信度。令s(x)表示这个逻辑回归模型的S形函数，仅接受s(x)大于或等于指定阈值超参数的实体，设置阈值使验证集中的F1分数最大



1. 实验结果

作者使用不同的超参数训练了由内而外的迭代模型，检查了所用单词表示的效果，LSTM层的数量以及单个LSTM单元的隐藏大小。



作者也对比了不同的选择策略下的算法性能：

