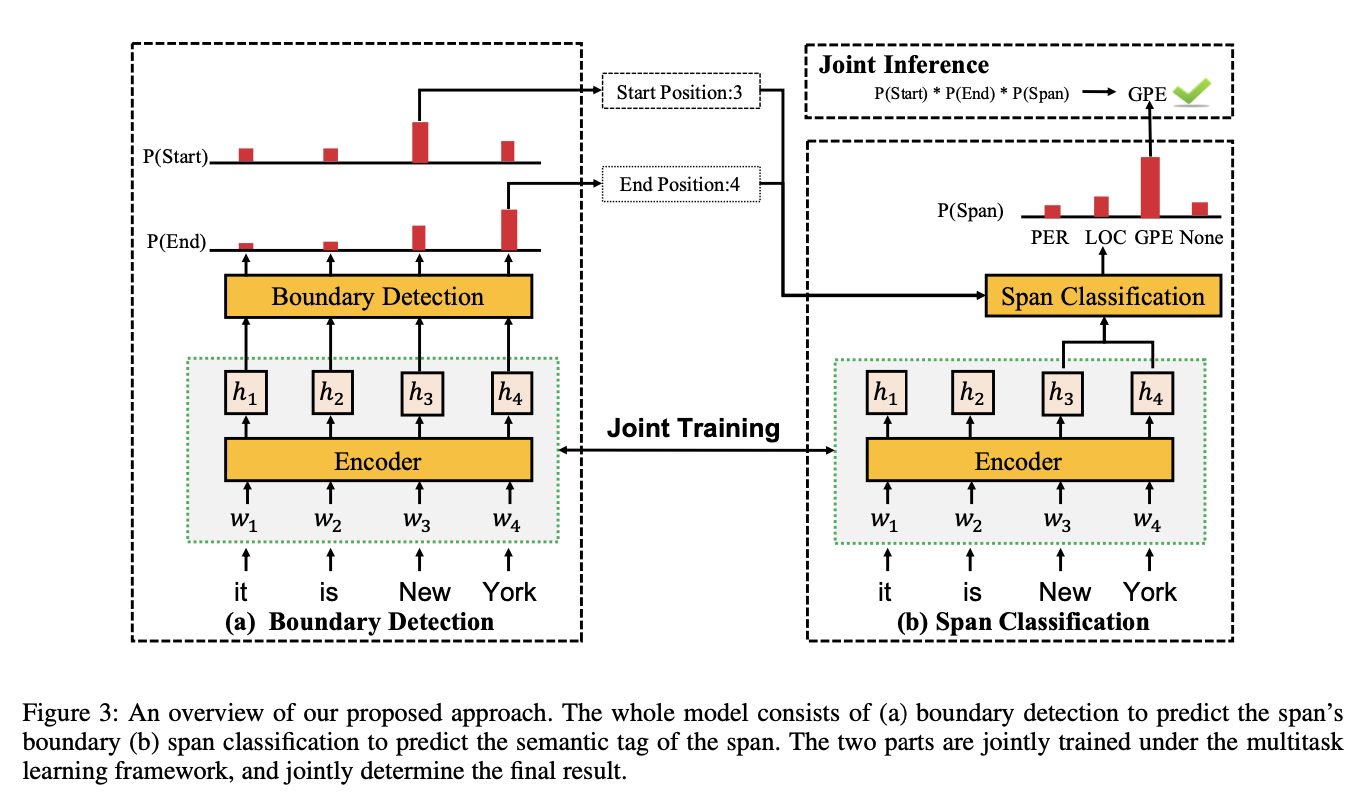
**论文阅读笔记一**

Boundary Enhanced Neural Span Classification for Nested Named Entity Recognition

1. Motivation
2. NER往往被当作一个序列标注问题，序列标注模型难以处理嵌套实体。
3. Span-base方法可以发现不同子序列中的嵌套实体，很适合处理嵌套实体问题，但是先前的span-base方法对所有的候选实体都进行分类，该方法计算复杂度高、推理速度慢，同时该方法主要学习跨度表示，缺乏有效的边界学习监督方式。
4. Contribution
5. 为了解决上述俩问题提出了一个边界增强神经跨度分类模型BENSC，除了对跨度进行分类外，也添加了一个边界探测任务来监督边界跨度表示学习。
6. 分类和边界检测任务在多任务训练框架下联合训练。边界探测模型负责预测每个词是否为实体的边界，跨度分类模型负责整合跨度的内部信息并预测其实体类别
7. 边界探测模型能够产生高质量的候选跨度，大大降低了时间复杂度。
8. 推理阶段，该框架通过边界检测模型获取边界置信度得分Ps和Pe，通过跨度分类模型获得跨度标签C得分Psp。这三个分数共同决定一个跨度是否为一个实体C。
9. Method

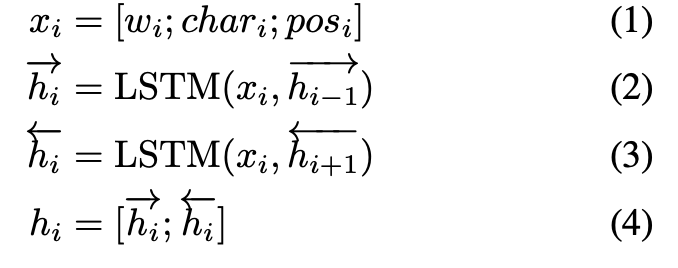


边界检测模块主要用于检测一个词是否为实体的起始词或终止词；跨度分类模型主要对跨度进行分类。这两个模块在多任务学习框架下联合训练。

1. 编码器

作者分别用LSTM和BERT对序列进行编码，对于给定句子S中的单词序列，句子长度为N，编码类型包括字符级、单词级和上下文词嵌入。

首先是LSTM编码：



对于BERT编码，首先对序列用Wordpiece算法进行分词，然后加入分类标识[CLS]、句子表示[SEP]得到输入序列，用一系列L堆transformer块进行编码：

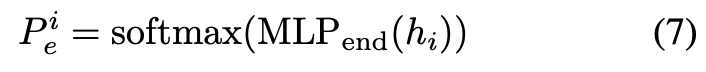


1. 边界检测

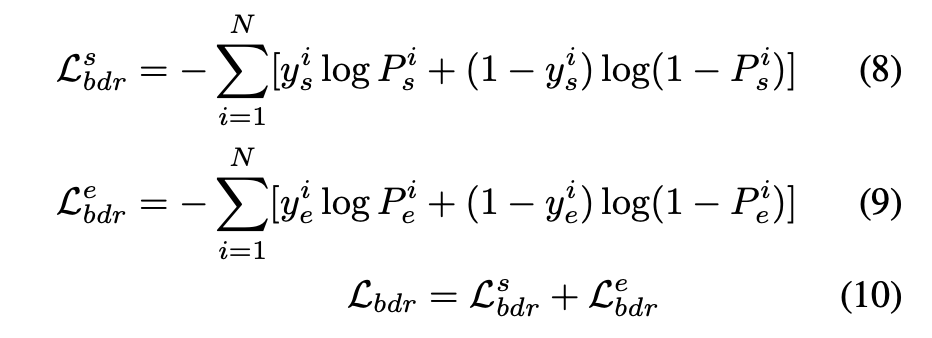
将单词的上下文表示hi送到多层感知机MLP里面，用一个softmax层获取单词wi是实体的起始词的概率：



用同样的方法获取一个词wi为实体的终止词的概率：



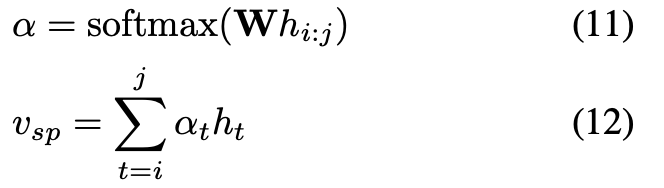
训练目标函数为起始词和终止词的交叉熵损失函数之和：



其中和分别表示单词i为实体的起始词和终止词的标签。

1. 跨度分类

对LSTM编码，作者使用注意力机制对相应边界(i,j)的标记进行汇总：



其中W是待学习的参数。

对于BERT编码，作者通过自关注机制获取跨度表示，对于跨度，用transformer块对跨度进行编码：

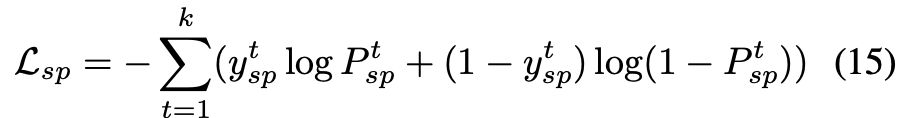


其中表示从词i到j这一跨度。

然后将跨度表示vsp送到多层感知机分类器中用softmax获取其类别标签：



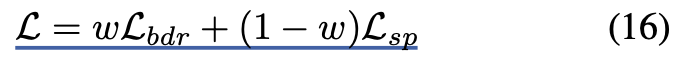
最后最小化交叉熵损失函数：



其中k是标签个数，表示跨度是否为标签t中的标签。

1. 联合训练和推理

训练目标是最小化边界和跨度的交叉熵损失函数之和：



其中w是平衡两个子任务的超参数。

在推理阶段，给定实例，首先用边界检测模型获取边界概率和，如果\*大于指定阈值，则对所有合理的(i＜j)跨度进行分类，然后将跨度送到跨度分类器中进行预测没如果比阈值大，则将该跨度归类为C。

1. 模型细节

LSTM编码：300维静态预训练GloVe词嵌入，未登录词用0补充，字符嵌入和磁性嵌入的大小为50，隐藏层向量维度为150，模型用Adam进行优化，学习率为0.002。

BERT编码：用base版BERT，隐藏层向量长度为768，12个注意力头，用Adam训练，学习率为3e-5。

1. 优点
2. 相较于序列标注的方法，span-base的方法能够有效识别嵌套实体。
3. 相较于hypergraph-base的方法，该方法能够节约人力、有效分类。
4. Idea

Multi-CRF+BENSC

BENSC，识别边界词；

Multi-CRF，依次对单个类别实体进行识别。

**论文阅读笔记二**

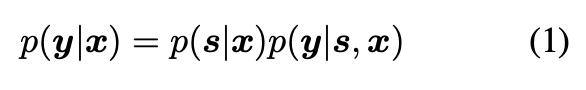
Combining Spans into Entities: A Neural Two-Stage Approach for Recognizing Discontiguous Entities

1. Motivation
2. 识别包含一些不连续的单词序列或是与其他实体重合的目标实体。
3. 通过两段式的框架来降低底层编码对下游任务的影响。
4. Contribution
5. 将不连续实体提取问题分解成两个子任务，提出了一种不存在歧义的两阶段不连续实体识别方法。
6. 针对两个子任务设计了两个神经网络组件，使用共享文本编码器来联合训练两个任务。
7. 相比于之前的方法，该模型下的任务不受外部特征的影响。
8. Method
9. 用x={x1,…x|x|}表示句子，x={}表示一组不连续的实体，其中类型k的每个实体都包含一个跨度列表，下标是跨度开始和结束的位置。比如一个句子“中国人民银行”，则被表示为{中国，中国人民，人民银行，中国人民银行}。

该模型采用两段式框架，第一步提取不连续实体中的跨度，然后将这些跨度合并到不连续实体中去。一般来说这种不连续实体都是被标注好的，该方法首先联合提取和标注实体中的跨度，然后仅将同类型的跨度进行合并，比如人民银行和中国人民银行都是ORG，则这俩合并，而中国为GPE，不合并。

本文将中间类型的跨度称为段（segment）。

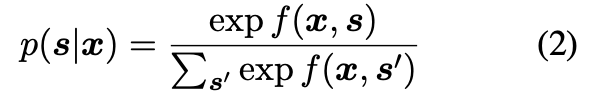
最后模型用CRF进行解码：



其中表示被标记为k类型实体的一组段。

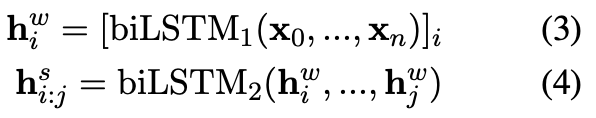
1. 段提取

给定句子x={x1,…x|x|}中的可能的实体片段可能互相重叠，为了使模型能够提取此类重叠的片段组合，作者使用了基于超图表示的神经段超图模型来编码无歧义的可能的段组合，具体来说，段超图用一个线性的方式来建模每个段组合的条件概率：



其中f(x,s)是句子x和输出段组合s中任意一对的得分。比如f(中国人民银行, 中国)。

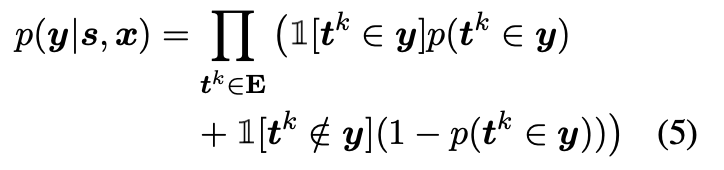
在分段超图中，每个段组合都对应一个超路径，超路径的得分是每条超边得分的总和，基于LSTM的单词级和跨度级表示如下：



其中xK是第k个单词的词嵌入，表示第i个单词的隐藏层状态，表示从i到j的段的隐藏层状态。

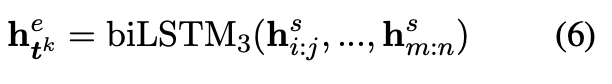
1. 段合并

给定一组段，要段合并为合并为实体。首先基于相同的实体中的段应具有相同类型且彼此不重叠的假设，枚举所有有效段组合，将这些组合表示为E，然后模型独立决定段是否有效，有效的实体段为候选实体，用tk表示候选实体段，∈s，给定段组合s，产生实体y的概率为：



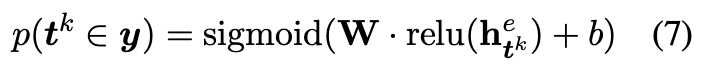
其中1是一个指示函数，使用二元分类器对p进行建模。

为了获取同一组合中各个段之间的相互作用，在段的顶部采用采用了一个LSTM：



其中表示段组合的向量形式，然后将这个向量作为二元分类器的输入以确定是否为实体。这一步重用了公式4的段编码，这意味着模型在段提取和合并阶段共享了词嵌入和段编码。

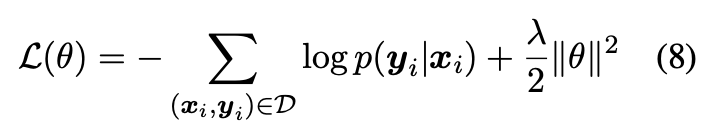
二元分类器如下：



在上面的联合训练的模型中，从单词编码到跨度的不同级别对文本用3个单独的LSTM进行编码，然后对不连续的实体进行编码。直观上，单词级LSTM获取地基信息以进行段提取，而实体级LSTM获取较高级别的信息以进行段合并。

1. 学习和解码

对于包含句子-实体对(x,y)的数据集D，实验目标是最小化负对数似然函数：



其中θ是模型参数，λ是l2正则化的系数。P(y|x)通过计算而来。

解码阶段，首先通过神经超图模型预测最佳段，然后将进行合并并通过输出实体类型。

1. 优点
2. 相比于人工构造规则识别实体的方法有了很大的进步。
3. 联合抽取和合并段，使得段结果没有歧义
4. 需要改进的地方Idea

段结果的正确结果传给实体识别阶段了，而错误结果被遗弃。

建议可以构建损失函数联合训练，避免一错再错。

论文阅读笔记三

Neural Segmental Hypergraphs for Overlapping Mention Recognition

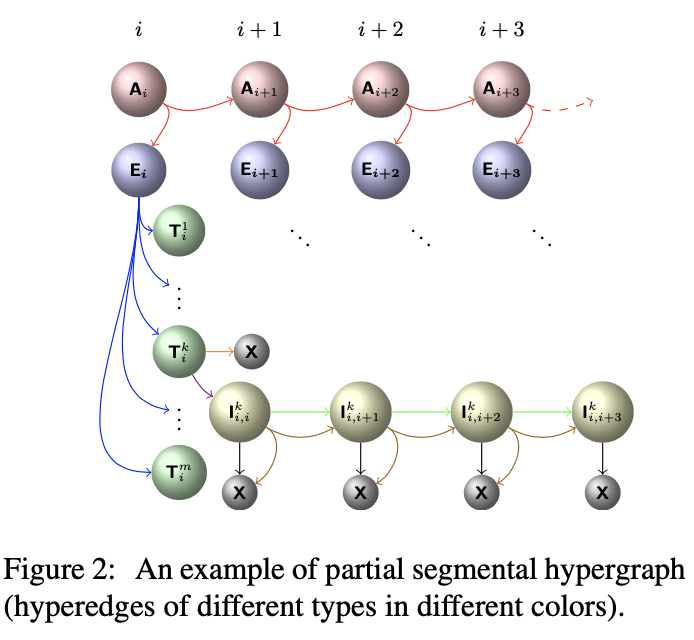
1. Motivation
2. 通过分段超图来表示数据集。
3. 建模数据集中的重叠实体。
4. Contribution
5. 作者提出了一种基于分段超图的表示方法，完成了重叠实体的建模。该模型可以获取其他模型无法获取的特征并保持较低的推理复杂度O(cmn)，其中m为实体类型数量，n是句子中单词的数量，c是每个实体的最大单词数量。
6. 作者从理论上证明了该模型不受到外部特征的限制，即加入BERT等预训练词嵌入也不会对模型结果造成太大的影响。
7. Method
8. 分段超图简介

作者用分段超图表示句子中所有的可能为实体的单词组合，是一种有向超图，其中每个超边都由单个指定的父节点(边e的头部)和一个有序的子节点(边e的尾部)组成。本文的分段超图由以下5种类型的节点组成：

1. 是对所有以第i个词或之后的词开头的实体进行编码
2. 对所有以第i个词开头的实体进行编码
3. 表示从第i个单词开始类型为k的实体
4. 表示所有类型为k、包含了从i开始到第j个单词的实体
5. X表示实体的结尾

连接节点的边旨在引导模型根据子节点重新表达父节点的语义。

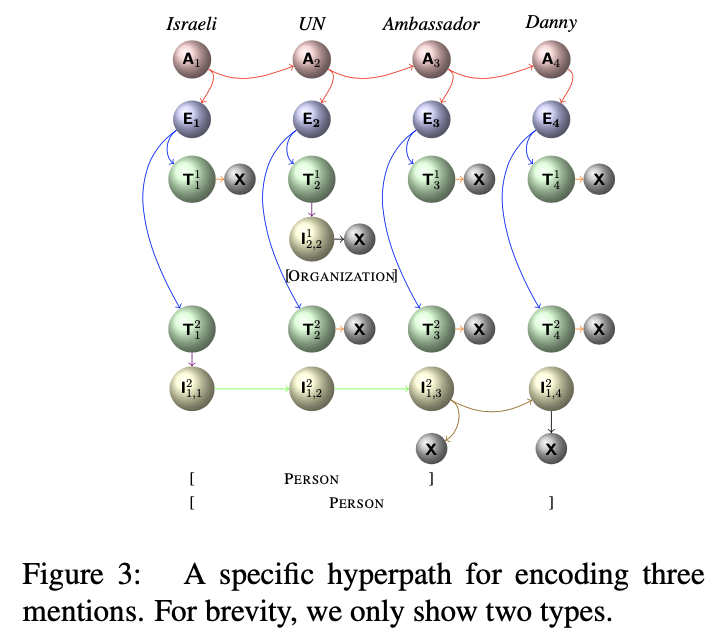
图2是由4个单词组成的跨度为[I,i+3]的所有实体组合：



图中的超边有4种类型：

* 1. 从到它的子节点的超边表示由完全从第i个单词开始的边或严格出现在第i个单词之后的组成。
  2. 从到它的子节点的超边表示我们应该考虑所有可能以第i个单词开始的实体的实体类型。
  3. 从开始的两条超边和表示至少存在一个从第i个单词开始的实体，或者不存在任何实体。
  4. 从开始的三条超边、、分别表示以下三种情况：第j和第j+1个单词至少有一个由第i个单词开始的实体；存在一个由第i个单词到第j个单词组成的实体；前两种情况都有可能。

完整的超图紧凑地编码了一个句子中可能出现的所有可能的实体，这些实体可能重叠或是不重叠。当我按照超边所指的方向编码完整的分段超图时，一次在每个节点上仅选择一个节点传出超边，如此就能得到超路径。



如图3，PERSON类型Israeli UN Ambassador来自以下的节点顺序(沿着超路径): A1, E1,,,,，而Israeli UN Ambassador Danny的PERSON类型来自以下节点序列：,,,,,X。这样的节点序列共同形成一个单一的超路径，该路径编码这种相互重叠的特定的实体组合。

1. 理论分析

定理3.1：(无结构歧义)对于任意句子及其分段超图，令S为给定句子的所有可能实体组合的集合，P为包含的所有超路径的集合，P与S中的元素之间是一一对应的关系。

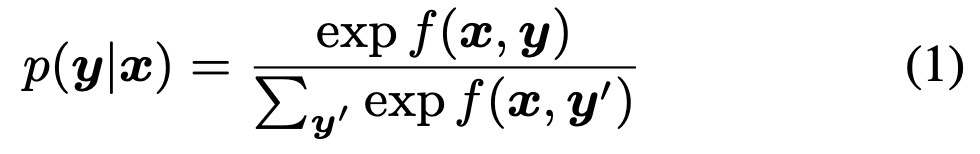
每条超路径的独特之处在于它是涉及X节点的超边的集合，这些超边可以唯一确定实体集合，相反，实体集合可以由这一类超边集合唯一地表示，从而产生唯一的超路径。

比如短语ABCD，假设只含有BC和ABCD两个实体，则B和A为正确的左边界，C和D为正确的右边界，如果基于匹配的思想进行预测，则有可能产生BD和AC这样错误的结果，这就是有歧义的地方。

相比之下，基于超图的模型定义了新的I节点和X节点共同获取完整的边界信息，因此具有无结构歧义的属性。

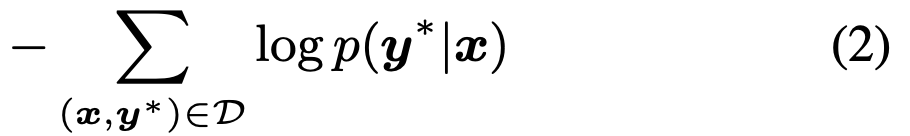
1. 基于超图的实体识别学习

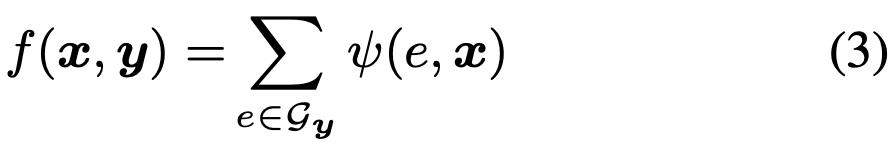
作者采用对数线性的方法对每个超路径的条件概率进行建模：



其中f(x,y)表示任意一对输入句子x和输出的实体组合y的得分，该分数对应唯一的超路径。

训练目标是最小化训练集中所有实例的负对数似然函数：



本文在超路径内的每个超边上定义特征，得分函数可被分解为以下形式：  


其中e∈是出现在超路径内的超边，而φ(e,x)是当输入句子为x时e在句子中的分数。这个函数在softmax边缘训练中介绍。

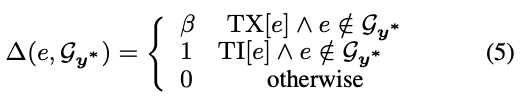
除了单词级特征，分段超图还允许定义跨度级特征，可以在跨度[i,j]的节点上提取局部特征。

* 1. Softmax边缘训练

φ(e,x)函数定义如下：



其中φ(e, x)是特征函数，∆(e, Gy∗ )是边缘损失函数：



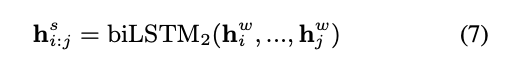
其中，y\*是黄金实体组合，TX[e]和TI[e]是指示符函数，如果e分别在T和X之间且在T和I之间，则返回true，否则返回false，令β≥1使得损失函数将给假阴性分配的惩罚比给假阳性分配的惩罚更多。

* 1. 特征表示

作者用LSTM来学习单词级和跨度级的特征表示，形成神经段超图模型。首先将句子中的第i个单词映射到其预训练单词，得到其预训练词嵌入ei，然后将其POS标签映射到标签嵌入pi中(如果存在的话)。第i个单词的表示形式是两种词嵌入的拼接结果：

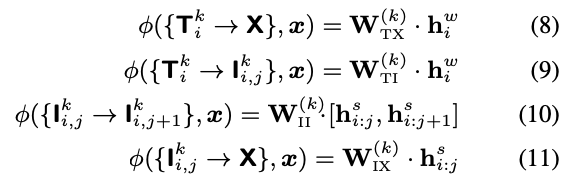


将该词嵌入结果作为LSTM的输入，生成跨度级的特征，（由Segmental recurrent neural networks得）作者使用O(cn)的复杂度计算所有跨度的词嵌入：



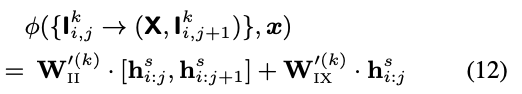
其中n是句子x的单词数，c是实体最大长度。

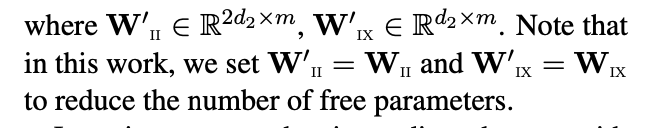
超图中有4种类型的超边，由此我们可以定义得分函数。每个有效实体的超路径都包括第一类和第二类超边，无需在此类超边上定义分数，因为它们的分数将成为可在对数线性模型的整体损失函数中消除的常数因子。只需在后两种类型的超边上定义得分函数，对于仅包含两个节点的超边，用线性层计算他们的分数：



其中矩阵WTX和WTI的维度为d2×m，WII维度为2d2×m，WIX维度为d2×m，上标k表示矩阵的第k列，d1是hw的维数，d2是hs的维数，m是实体类型的数量。

对于涉及两个以上节点的超边，得分计算如下：





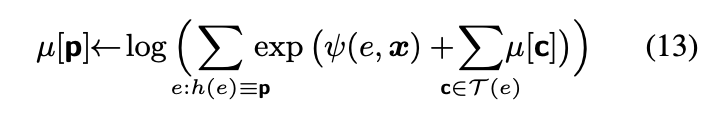
该过程使用随机梯度下降下并用Adam更新规则，用l2正则化减少过拟合。

* 1. 字符级表示

为了与主流模型进行对比，作者整合了字符级的组件来获取字符和形态特征，使用LSTM嵌入作为输入。每个单词的字符级表示chi是将LSTM的前向和后向隐藏层向量进行拼接得到的，每个单词的表示形式为

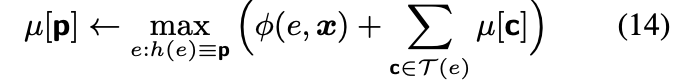
1. 推理和预测

推理阶段使用inside-outside style message-passing algorithm进行推理，从叶节点X遍历到根节点A1，将消息从其子节点传递到父节点p节点：



其中h(e)是超边e的头，T(e)是形成边e的尾巴的节点集合，它们是给定e的h(e)的子节点。

解码阶段：作者在完整的超图上执行标准的MAP推理，寻找最佳的超路径，生成的过程类似于最大乘积消息传递算法，仅考虑特征函数φ来构造消息：



每个节点对应于一个总和/最大值计的计算，一个节点最多与3个超边相连，推理的时间复杂度可通过图中的节点数来表示，即O(cmn)，c为实体最大长度。

1. 优势
2. 避免了实体边界检测的歧义。
3. 相比于BERT等神经网络模型，具有体积小、复杂度低的优势
4. Idea
5. 将超图识别出的候选实体用LSTM进行编码并结合CRF进行实体识别
6. 边界检测和实体识别联合训练，避免误差传播。