**论文阅读笔记一**

Joint Mention Extraction and Classification with Mention Hypergraphs

1. Motivation
2. 获取长度不受限的重叠实体
3. 线性时间复杂度
4. Contribution
5. 作者设计了一个能够有效处理任意长度的重叠指称的模型
6. 作者设计的模型的学习和推理算法具有线性复杂度
7. 该模型能够以线性复杂度额外获取指称的头部信息
8. 实现的Method
9. 关于超图

对于由n个单词组成的句子，一个指称m可以用一个三元组表示为<bm, em, τ>，其中b和e是指称m的第一个词和最后一个词，τ表示实体类型，

可能的候选实体的个数为tn(n+1)/2个，t表示实体类型数。如果直接建模这些候选实体，计算复杂度非常高。

一种解决方案就是对候选实体指称引入紧凑表示，以指数方式编码实体指称，从而降低计算量。本文介绍的方法是超图表示法。

超图的构成：节点+超边

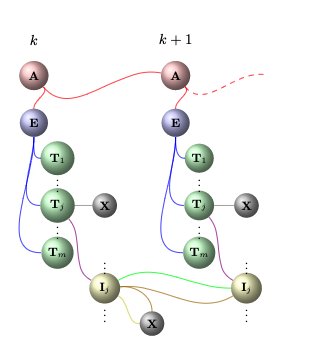
A节点：按顺序排列具有不同左边界的指称。具体来说，位置k(第k个词)或Ak处的每个A节点用于紧凑地表示句子中所有开始于第k个词的指称。

E节点：用于**严格**表示左边界开始于第k个单词的指称。

T节点：用于紧凑表示所有**严格**开始于第k个词的类型为j的指称。

I节点：用于紧凑表示包含第k个词的类型为j的指称。

X节点：终止节点。



连接父子节点的红色超边，

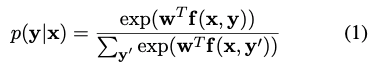
对每个I节点，顶部绿色边表示当前单词出现在指称中间，底部黄色边表示当前单词在指称中是最后出现的词，中间的棕色边编码了两种同时发生的事实，即单词属于同一类型的多个重叠实体。

句子中的指称的组合都可以用完整的指称超图的一个子超图表示，但是反过来不行，因为两个不同的重叠实体组合可能共享同一指称超图。

一个指称可被一条超路径唯一表示，所有超路径构成了原始超图的唯一超子图。

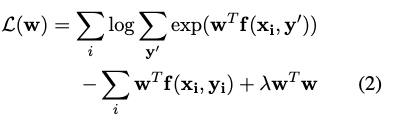
1. 对数线性模型 log-linear modeling

参考CRF模型，作者用了一个对数线性方法来进行联合指称抽取和分类，具体来说，给定输入句子X，输出y(表示指称的特定组合的指称子超图)的概率被表示为如下形式：



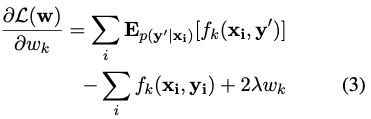
其中f(x,y)是定义在输入输出对(x,y)上的特征向量，w是模型的权重参数。

训练的目标是最小化数据集的正则化负对数似然函数：



其中表示第i个训练实例，最后一项表示L2正则项，λ是一个正标量，定值为0.01.

目标函数的梯度为：



其中wk表示第k个特征函数fk的权重。

1. 学习算法

在每个时间步k，要计算指称m的I节点、T节点、E节点、A节点的得分，总的时间复杂度为O(mn)(假设每个节点的特征函数的分数是一个恒定的时间)。若直接用CRF，则由于一阶假设，复杂度为O(m2n)。

1. 特征表示
2. 出现在当前单词周围的上下文窗口大小为3的单词特征，如果有POS特征也加入其中
3. 包含当前单词的单词语法信息n-gram，n=2，3，4
4. 当前单词的BoW特征，上下文窗口大小为5
5. 单词模板特征：特征定义在(x,y)上，内容包括定义在输入上的指标函数和节点上的指标函数的乘积。节点T和I的类型和连接节点T和I节点的超边。
6. 联合建模

指称提取和分类任务的另一个假设就是每个指称要有一个指称头，头就是一个指称的子串。

用三种不同的节点代替I节点，即表示出现在类型j中且在其头部之前出现的单词，节点表示出现在类型为j的指称的头中间的词，表示出现在类型为j的指称的头后面的单词。超边也需要根据节点组合情况进行构建。在这种情况下，计算的复杂度依然为O(mn)。

1. 优点

线性时间复杂度