Partial Annotation based CRF

朱运

April 13, 2018

1 符号定义

$\mathcal{D} = \{S^j, Y^j\}_{j=1}^N$：表示一个数据集，包含 N 个句子和对应的 N 个人工标注的分词序列。

$S^j = w_1^j ... w_i^j ... w_{n_j}^j$：表示第 j 个句子，由 n_j 个汉字组成。

$Y^j = y_1^j ... y_i^j ... y_{n_j}^j$：表示第 j 个句子对应的标签序列。

$\mathcal{T}$：表示标签集合，即隐状态的所有可能取值，$y_i^j \in \mathcal{T}$。

$\mathcal{V}$：表示字表 (vocabulary)，即数据 $\mathcal{D}$ 所有汉字的集合，$w_i^j \in \mathcal{V}$。

2 概念定义

以汉语分词任务举例，对于一个例句“我是中国人。”，我们采用四标签集（BMES），B、M、E、S 分别代表一个词的开始，中间，结尾和单字成词。

全标注是指句子中的每个字都给出了正确的分词标签，即 $|Y^j| = |S^j|$。假设全标注对应的 tag 序列是 Y =（S,S,B,M,E,S）, 如图 1 所示，那么其每个汉字对应的 tag 集合就是（{S},{S},{B},{M},{E},{S}）

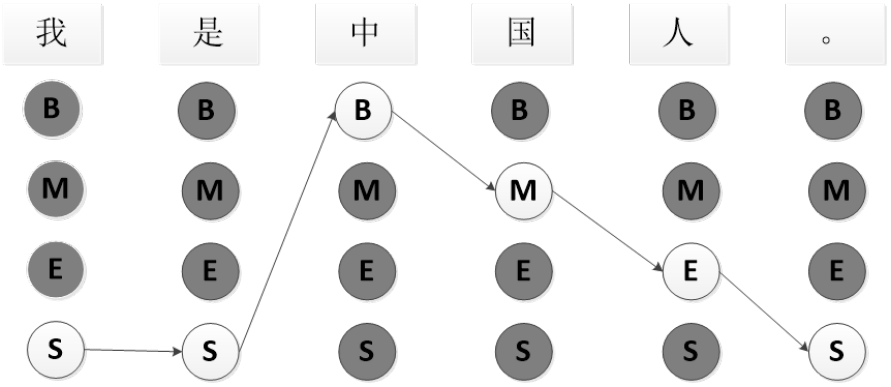

图 1: 全标注

部分标注是指句子只有某些字的分词标签给出了，而其余的字的标签没有给出，我们可以将全标注作为局部标注的一种特殊情形。假设部分标注已知的切分信息是“中国人”是一个词，其余位置未知，那么“中国人”对应的 tag 序列是（B,M,E），如图 2 所示，那么其每个汉字对应的 tag 集合就是（{B、M、E、S},{B、M、E、S},{B},{M},{E},{B、M、E、S}）

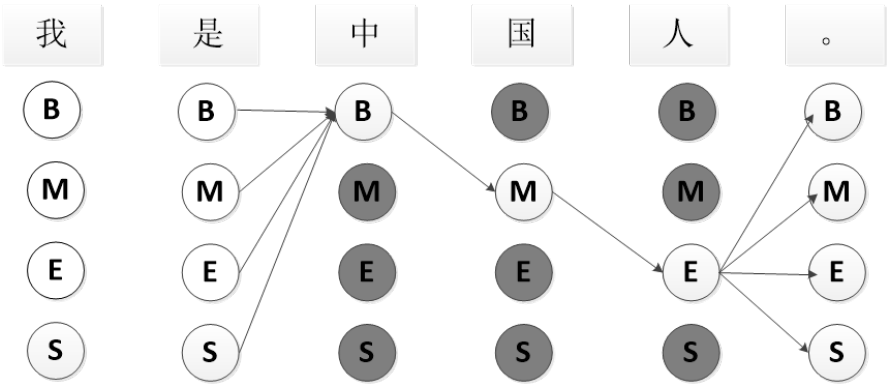

图 2: 部分标注

这里，我们定义了模糊标注：即句子中，每个字的标签是不唯一的，即我们允许一个字可以有多个标签。假设模糊标注已知的切分信息是“中国人”或者“中国”和“人”，其余位置未知，如图 3 所示，那么其每个汉字对应的 tag 集合就是（{B、M、E、S},{B、M、E、S},{B},{M、E},{E、S},{B、M、E、S}）

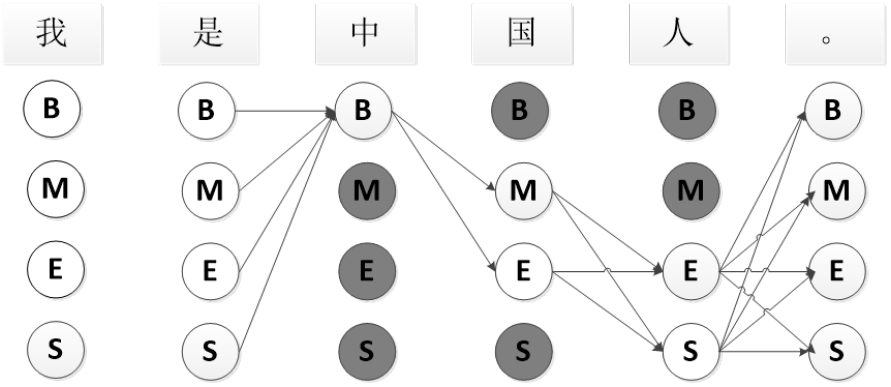

图 3: 模糊标注

3 公式推导

我们采用 CRF 模型来处理分词的序列标注问题。给定一个输入的字符序列，模型的作用是给这个序列每个位置的字赋予对应的一个标签。

全标注中，假设给定的一个句子 S = $w_1...w_n$，其对应的正确 tag 序列为 Y = $y_1...y_n$。那么 CRF 定义句子 S 标注为序列 Y 的概率为：

$$p(Y|S) = \frac{e^{\mathit{Score}(S,Y)}}{Z(S)} \tag{1}$$

其中

$$Z(S) = \sum_{Y' \in \mathcal{T}^n} e^{\mathit{Score}(S,Y')} \tag{2}$$

部分标注中，如图 2 所示，只有句子里面给定的词有固定的 tag，其余位置标签不确定，那么符合图 2 情况的序列集合

Y_L = {(B,B,B,M,E,B),(M,B,B,M,E,B),(E,B,B,M,E,B),(S,B,B,M,E,B),
(B,M,B,M,E,B),(M,M,B,M,E,B),(E,M,B,M,E,B),(S,M,B,M,E,B),
(B,E,B,M,E,B),(M,E,B,M,E,B),(E,E,B,M,E,B),(S,E,B,M,E,B),
(B,S,B,M,E,B),(M,S,B,M,E,B),(E,S,B,M,E,B),(S,S,B,M,E,B)...}

集合里面一共 4*4*1*1*1*4 = 64 种情况。其中 Y_L 表示所有可能的序列集合，那么 Y_L 的边缘概率可以表示为

$$p(Y_L|S) = \sum_{y \in Y_L} \frac{e^{\mathit{Score}(S,y)}}{Z(S)} \tag{3}$$

我们定义 Z_L 为：

$$Z_{Y_L} = \sum_{y \in Y_L} e^{Score(S,y)} \tag{4}$$

那么部分标注的边缘概率就可以归一化为：

$$p(Y_L|S) = \frac{Z_{Y_L}}{Z(S)} \tag{5}$$

根据全标注的 CRF 似然函数

$$LL(\mathcal{D};\mathbf{w}) = \sum_{j=1}^{N} \left[\mathit{Score}(S^j, Y^j) - \log Z(S^j)\right] \tag{6}$$

我们可以得到对应的部分标注的似然函数：

$$LL(\mathcal{D};\mathbf{w}) = \sum_{j=1}^{N} \left[\log Z_{Y_L}(S^j) - \log Z(S^j)\right] \tag{7}$$

根据 CRF 似然函数求解：

$$\frac{\partial \log Z(S^j)}{\partial \mathbf{w}} = \sum_{Y' \in \mathcal{T}^n} p(Y'|S) \cdot \mathbf{f}(S^j, Y') \tag{8}$$

可以得到：

$$\frac{\partial \log Z_{Y_L}(S^j)}{\partial \mathbf{w}} = \sum_{Y' \in Y_L} p(Y'|S) \cdot \mathbf{f}(S^j, Y') \tag{9}$$

Z_{Y_L} 和 Z 的形式以及计算梯度方式很相似，通过对 Z_{Y_L} 的前向 alpha 和后向 beta 的计算加上一些约束，我们可以将 Z_{Y_L} 和 Z 的计算统一起来。

$$\alpha_{Z_{Y_L}}(k,t) = \begin{cases} \sum_{(t',t)\in Y_L} e^{\mathbf{Score}(S,k,t',t)} \cdot \alpha(k-1,t') & (t',t) \in Y_L \\ 0 & (t',t) \notin Y_L \end{cases} \tag{10}$$

$$\beta_{Z_{Y_L}}(k,t) = \begin{cases} \sum_{(t,t')\in Y_L} e^{\mathbf{Score}(S,k+1,t,t')} \cdot \beta(k+1,t') & (t,t') \in Y_L \\ 0 & (t,t') \notin Y_L \end{cases} \tag{11}$$

同理，模糊标注的计算方式和局部标注的计算一样。

4 说明

全标注，部分标注以及模糊标注在训练和测试的时候不需要考虑 tag 之间的约束关系（即标签 B 之后只能接标签 E 或者 M，而不能接标签 S），只有在算 prf 值需要解码的时候需要考虑 tag 之间的约束关系。