Partial Annotation based CRF

朱运

April 16, 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，$y_i^j \in \mathcal{T}$。

2 概念定义

我们以汉语分词任务为例，讲解基于 CRF 如何实现局部标注，从而得到用于模型更新的 loss。汉语分词任务中，我们采用四标签集 {B, M, E,S} 来表示每个字的分词结果，其中，B、M、E 分别代表一个词的开始、中间、结尾，S 表示单字成词的字。那么，$\mathcal{T} = \{B, M, E, S\}$。

全标注是指句子中的每个字都给出了人工标注的分词标签，即 $|Y^j|$ = $|S^j|$。以句子“我是中国人。”为例，我们知道句子的分词答案是“我 是 中国人 。”，那么全标注对应的 tag 序列是 Y =（{S},{S},{B},{M},{E},{S}）, 如图 1 所示。

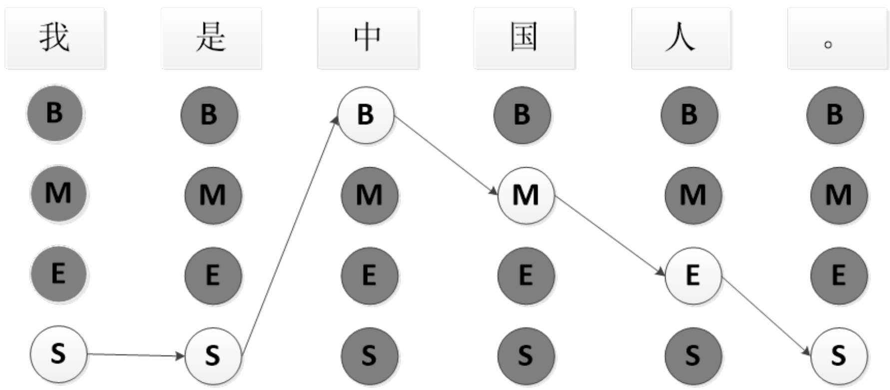

图 1: “我是中国人”的全标注示例

部分标注是指句子中只有部分字的分词标签给出，而其余字的标签没有给出。我们可以将全标注作为局部标注的一种特殊情形。假设在句子“我是中国人”中，“中国人”是一个词，其余字的分词信息未知，那么“中国人”对应的 tag 序列是（B,M,E）, 句子中的其他字，我们认为它的标签可能是 $\mathcal{T}$ 中的每个标签，那么句子的标签序列可以表示为（{B、M、E、S},{B、M、E、S},{B},{M},{E},{B、M、E、S}），如图 2 所示。

模糊标注是指：句子中，每个字或者某些字的标签不是唯一的，即我们允许一个字可以有多个标签。句子“我是中国人。”，按照不同粒度，“中国人”可以看做一个整体，

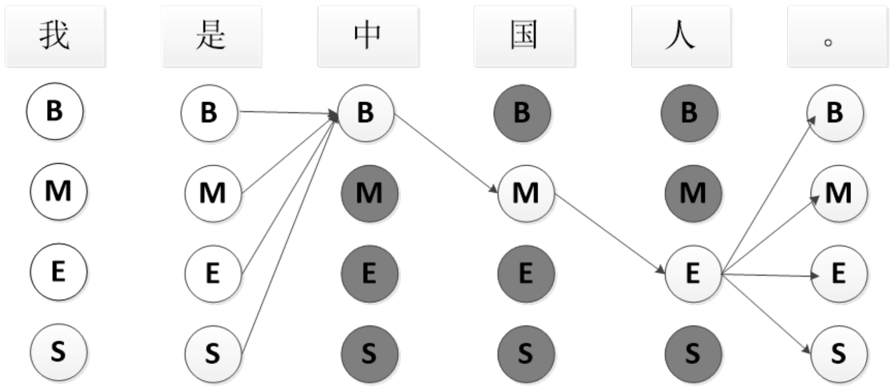

图 2: “我是中国人”的局部标注示例，其中，“中国人”是已知的分词信息

也可以切分为：“中国”和“人”。对于句子中的其他字，因为缺少其具体的分词信息，我们认为它的标签可能是 $\mathcal{T}$ 中的每个标签，那么句子的标签序列为（{B、M、E、S}, {B、M、E、S}, {B}, {M、E}, {E、S}, {B、M、E、S}），如图 3 所示。

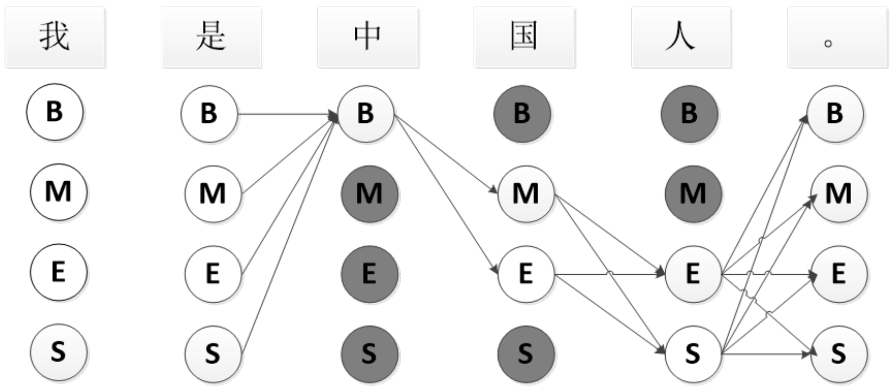

图 3: “我是中国人”的模糊标注示例

3 公式推导

我们采用 CRF 模型来处理分词的序列标注问题。给定一个输入的字序列，模型的作用是计算序列中每个字赋予每个标签的概率。

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

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

其中，

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

部分模糊标注，如图 3 所示，句子中，只有部分词语给定了某个或者某些 tag，其余位置标签完全不确定，那么符合图 3 情况的序列集合：Y^p = （{B、M、E、S}, {B、M、 E、S}, {B}, {M、E}, {E、S}, {B、M、E、S}），集合 Y^p 里面一共 4*4*1*2*2*4 = 256 种情况。其中 Y^p 表示所有可能的序列集合，那么 Y^p 的边缘概率可以表示为

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

我们定义 Z^P 为：

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

那么局部模糊标注 Y^p 的边缘概率就可以归一化为：

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

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局部模糊标注的似然函数：

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

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：

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

Z_{Y^p} 和 Z 的形式以及计算梯度方式很相似，区别在于，在局部模糊标注情况下，我们在计算前向 alpha 和后向 beta 的过程中需要加一些约束，将不符合已有标注的情形禁止掉即可。

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

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

其中，Y_k^p 表示，在局部模糊标注中，第 k 个字对应的所有分词标签。

4 说明

全标注，局部标注及模糊标注在训练时，不需要考虑分词标签之间的约束关系（如：B 之后只能接 E 或 M，而不能接 S），只有在测试解码时才考虑标签之间的约束关系。