9.序列标注

1. 应用场景

序列标注任务

- 对于序列中的每个时间步做分类
- 得到每个时间步的标签

- 对于输入：X1X2X3X4....Xn
- 预测输出：Y1Y2Y3Y4.....Yn

- 应用场景：
- 分词，词性标注，句法分析，命名实体识别等

命名实体识别（NER）：可以提取文本中的实体，抽取有效信息更好地完成下游任务，广义的信息抽取

输出长度为句子长度的标签

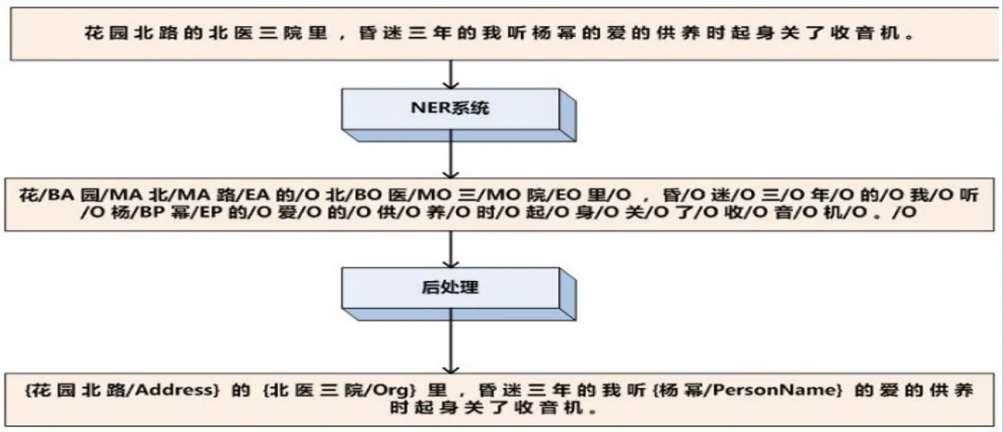
-盗版必究-

- BA:地址左边界
- MA:地址内部
- EA:地址右边界

- BO:机构左边界
- MO:机构内部
- EO:机构右边界

- BP:人名左边界
- MP:人名内部
- EP:人名右边界

- O:无关字



对每个字进行标签分类，下面相当于10分类

2. 条件随机场

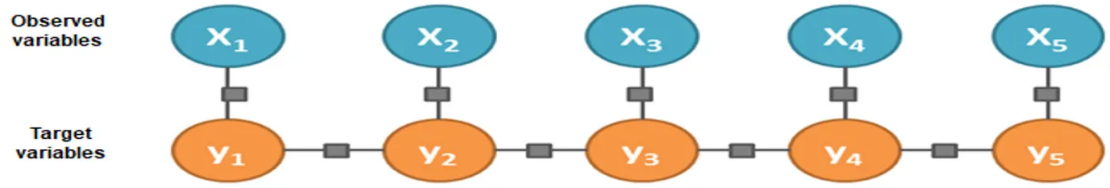
CRF-条件随机场

--八斗人工智能
-盗版必究-

Find a label sequence y that maximizes:

$$p(\boldsymbol{y}|\boldsymbol{x};\theta) = \frac{1}{Z(x)}\exp\{\sum_{i=1}^{M}\theta \cdot f(y_{i-1}, y_i, x_i)\}$$

- Input observation sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_M)$
- Output label sequence $\boldsymbol{y} = (y_1, y_2, \ldots, y_M)$
- $f(y_{i-1}, y_i, x_i)$: feature function vector
- θ: weights
- $Z(x)$: term for normalization



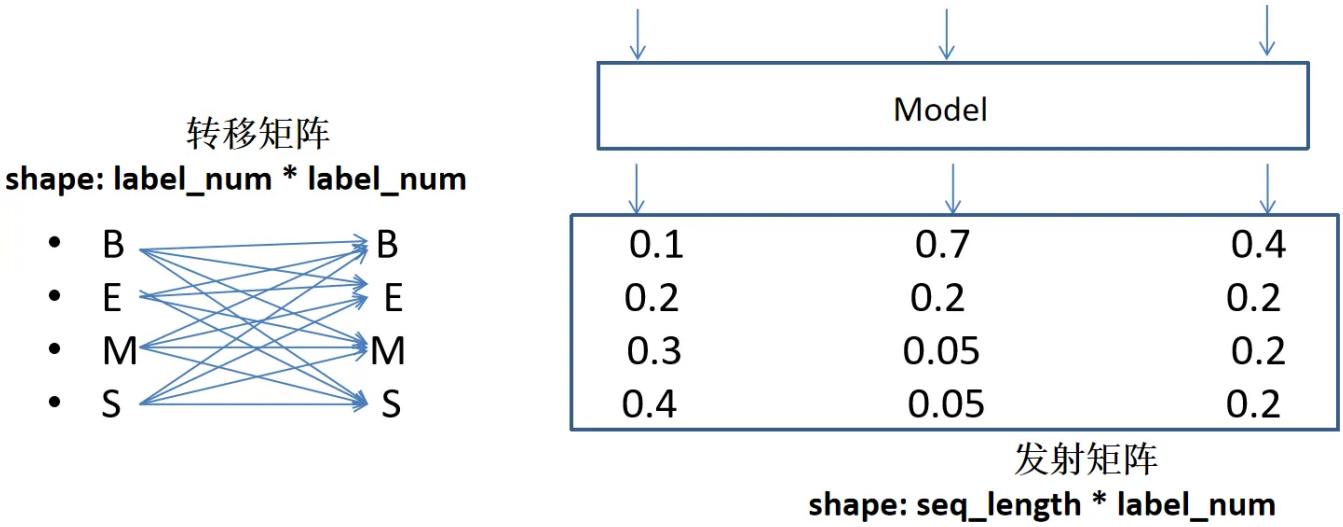
要解决的问题：地址的左边界后面不可能直接接一个人名的右边界，因此需要合理的条件约束，限制类别之间的转移关系

CRF-转移矩阵

以↓开始
转移到→

	START	B-Person	I-Person	B-Organization	I-Organization	O	END
START	0	0.8	0.007	0.7	0.0008	0.9	0.08
B-Person	0	0.6	0.9	0.2	0.0006	0.6	0.009
I-Person	-1	0.5	0.53	0.55	0.0003	0.85	0.008
B-Organization	0.9	0.5	0.0003	0.25	0.8	0.77	0.006
I-Organization	-0.9	0.45	0.007	0.7	0.65	0.76	0.2
O	0	0.65	0.0007	0.7	0.0008	0.9	0.08
END	0	0	0	0	0	0	0

CRF转移矩阵也需要机器学习来学习



发射矩阵即为每个字对应类别概率矩阵，在训练时，需要同时考虑转移矩阵和发射矩阵（原token的标签概率），即同时训练这两个矩阵。

输入序列X，输出序列为y的分数：$s(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i}$ A为转移矩阵 P为发射矩阵

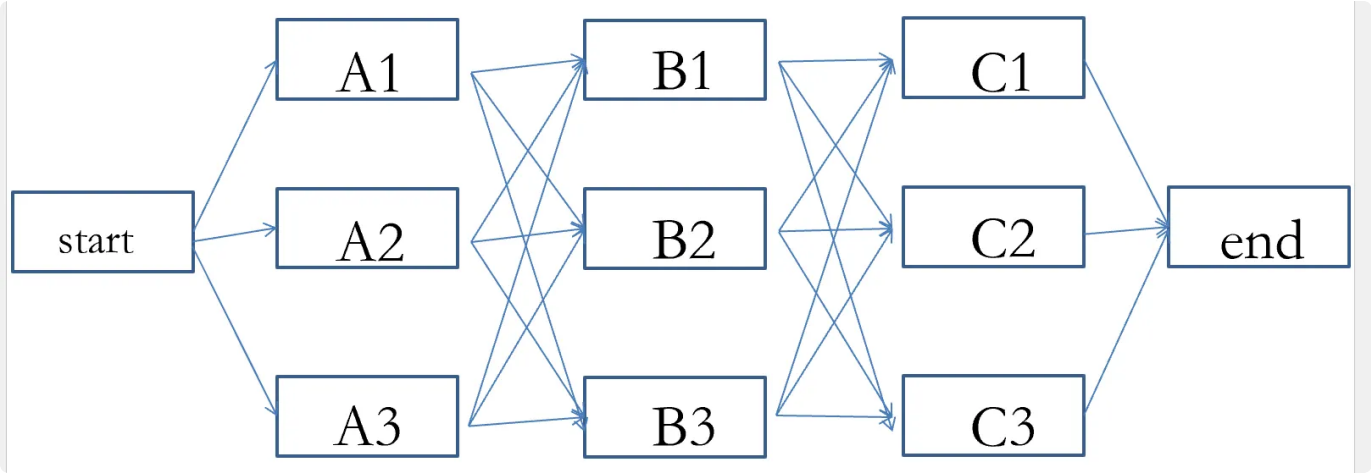
输入序列X，预测输出序列为y的概率：$p(y|X)=\frac{e^{s(X,y)}}{\sum_{\tilde{y}\in Y_X}e^{s(X,\tilde{y})}}$ 相当于上式做softmax

对上式取log，目标为最大化该值：$log(p(y\mid X)=log\frac{e^s(X,y)}{\sum_{y\in Y_X}e^s(X,y')}=S(X,y)-log(\sum_{y'\in Y_X}e^s(X,y'))$

对上式取相反数做loss，目标为最小化该值：$Loss=\log(\sum_{\tilde{y}\in Y_X}e^{S(X,\tilde{y})})-S(X,y)$

实际上，CRF-Loss的计算本质是softmax，最大化真实路径的概率。 Y_X 表示从X到Y所有路径的集合。

CRF在基础模型能力强的情况下，性能提示有限。



线表示转移矩阵的值，方块内的元素表示发射矩阵的值

维特比解码：每次向下一层搜索时，保留最高概率的一条路径，在下一次搜索时，继续向下一层求得最大概率的一条路径。因此每次搜索的复杂度为 $o(D^2)$ ，总复杂度为 $o(nD^2)$ 。维特比解码的核心在于每个保留了到每一层中每个候选者的路径。

Beam Search：每次搜索后，保留B条概率最高的路径，在下一次搜索时，从这B条路径继续出发，再保留B条概率最高的路径。因此每次搜索的复杂度为 $o(BD)$ ，总复杂度为 $o(nBD)$

暴力搜索：复杂度为 D^n

- 在序列标注问题中，一般用维特比解码，因为序列标注的标签比较少，一般 $D<B$
- 在生成任务中，一般有Beam Search，因为D为词表大小， $D>>B$

3. 评价指标

- 序列标注准确率 ≠ 实体挖掘准确率
- 实体需要完整命中才能算正确
- 对于标注序列要进行解码

预测值 / 真实值	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

$$Accuracy = TP + TN/(TP + TN + FP + FN)$$

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F1score = 2 * Precision * Recall/(Precision + Recall)$$

准确率：正确检测的概率

召回率：正确 检测正样本的概率，适用于漏检正类代价很高的场景，如疾病诊断，安全检测

F1值：准确率和召回率的综合考虑

- Macro-F1
- 对所有类别的F1值取平均

- Micro-F1
- 将所有类别的样本合并计算准确率和召回率，之后计算F1

- 区别在于是否考虑类别样本数量的均衡

如果Macro-F1和Micro-F1相差很大时，样本是不均衡的

4. 问题

4.1. 抽取的实体有标签重叠如何解决?

- 我周末去了北京博物馆看展览
- 地点　　B E
- 机构　　B M M M E

- 对于每种实体使用独立模型
- 生成式模型

5. 文本加标点

- 经过语音识别或机器翻译可能会得到没有标点符号的文本此时进行自动文本打标有助于增强文本可读性

- 经过语音识别，或机器翻译，可能会得到没有标点符号的文本。此时进行自动文本打标，有助于增强文本可读性

- 是一种粗粒度的分词

6. 句子级别的序列标注

例如审稿人和作者的邮件对话，需要得到那句话是审稿人提出的第一个意见？

- 对于一个段落中的多句话，对每句话进行分类

- paragraph -> sentence -> token

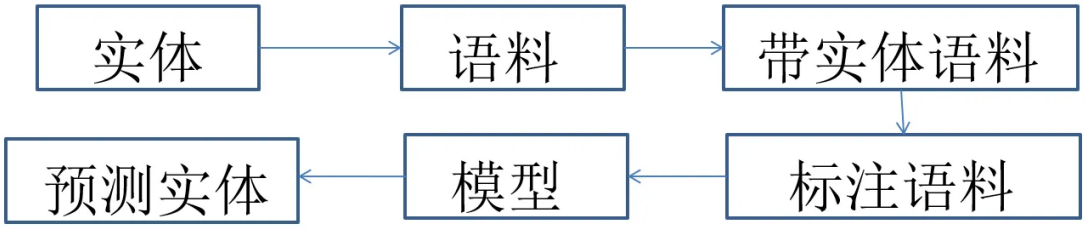
- 将每句话进行向量化，之后仅需进行序列标注

模型：

1. bert进行文本向量化
2. lstm得到句子之间的关联关系（否则就是对独立的句子进行分类，效果并不好）
3. 再进行分类任务

7. 远程监督

- If two entities have a relationship in a known knowledge base, then all sentences that mention these two entities will express that relationship in some way.



1. 我们有实体库，通过实体，可以标注语料。
2. 通过标注语料，我们可以训练模型
3. 模型来预测其他文本，得到新的实体，补充回原实体库

类似一个**滚雪球**的做法

8. 用bert进行实体识别任务

用bert预训练模型需要注意几个地方

1. 由于用的是bert的预训练好的词向量，因此分词也应该相应地使用bert的自带的tokenizer
2. bert的分词器，自动会将句首添加一个起始符<CLS>（编号为101）在句尾添加一个结束符<SEP>（编号102），因此在构建数据集时，应该对应添加无关label