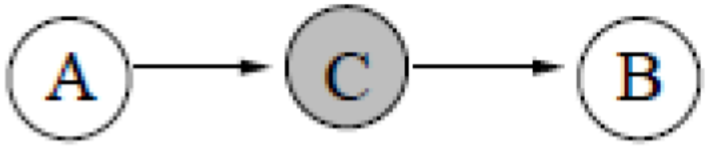
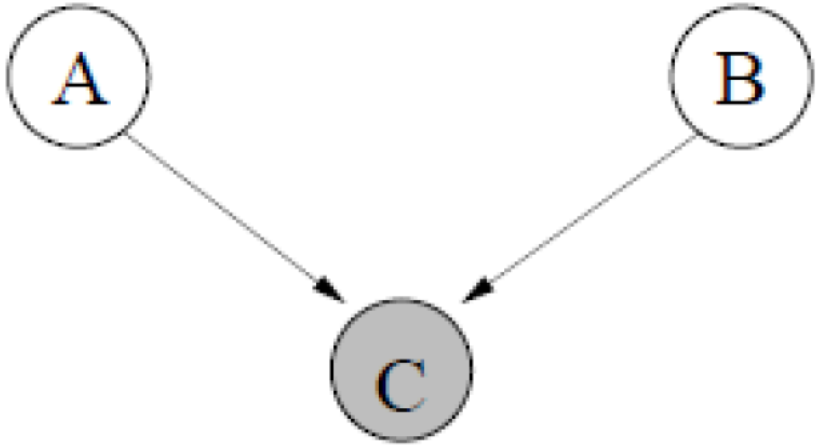
CNN-LSTM-CRF and Joint Training withWord Segmentation

张升涛

2019-05-29

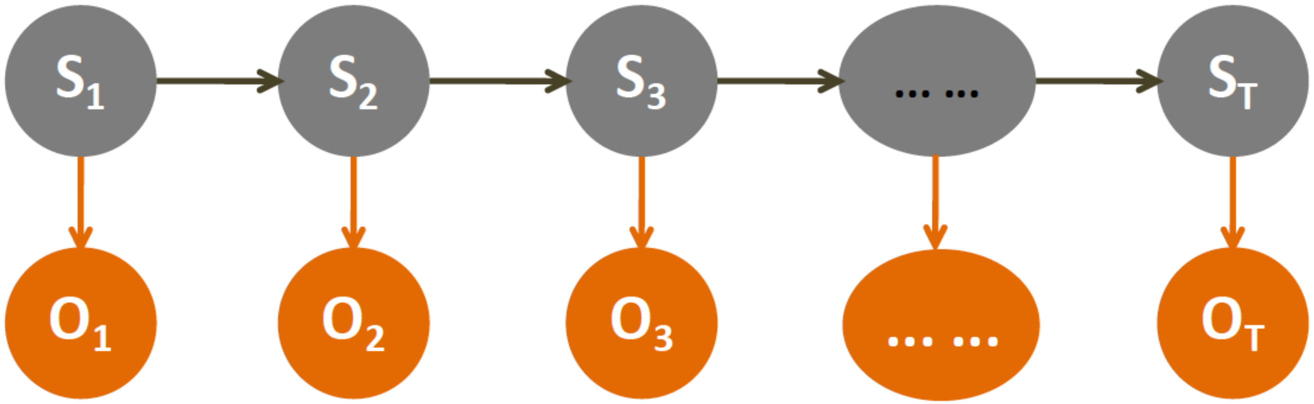
概率图模型 -- D - separation





隐马尔科夫模型

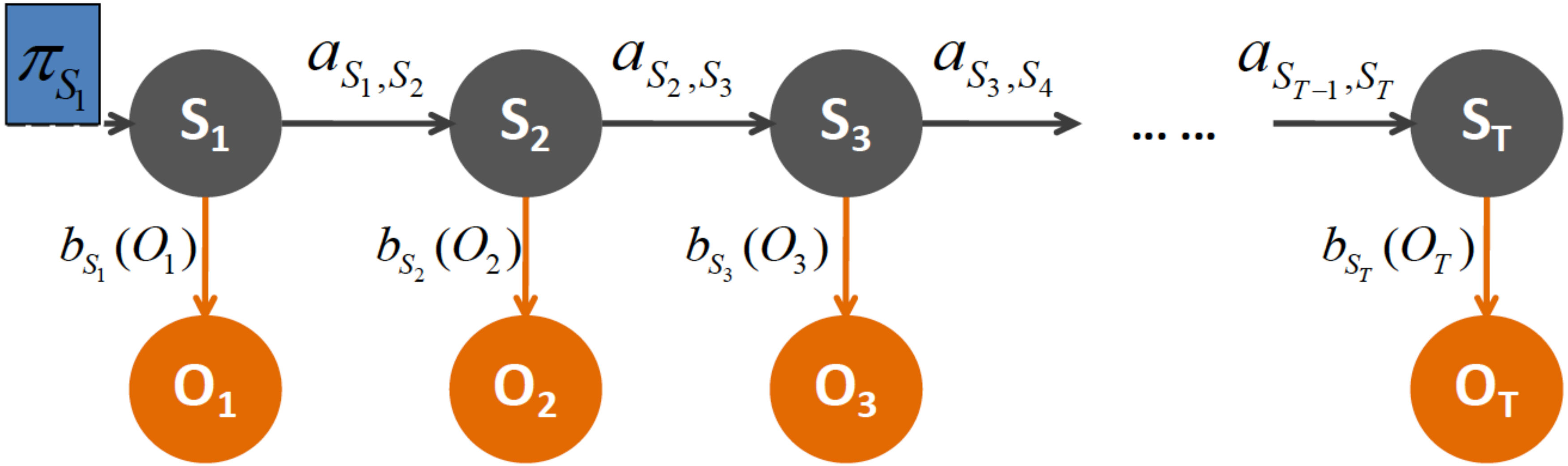
- 观察和隐藏序列共同构成了隐马尔科夫模型(Hidden Markov Model)



  - $O(o_1o_2...o_T)$：观测序列，o_t 只依赖于 s_t
  - $S(s_1 s_2 ... s_T)$：状态序列(隐藏序列)；S是Markov序列，假设1阶markov序列，则 s_{t+1} 只依赖于 s_t，

参数估计

- 先生成第一个状态，然后依次由当前状态生成下一个状态，最后每个状态发射出一个观察值

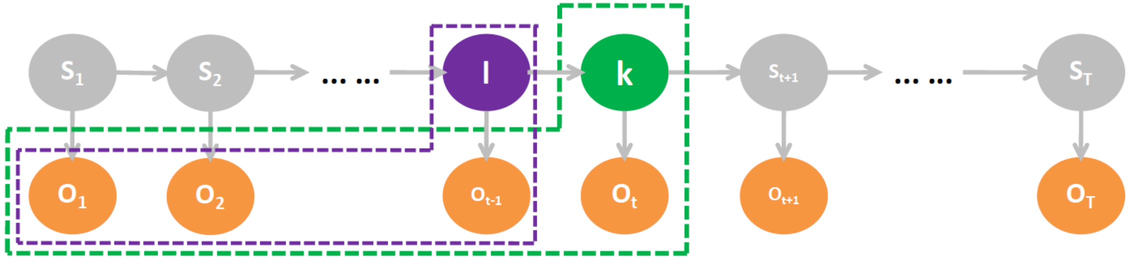


$$P(o_{1:t}, s_{1:t}) = P(s_1)P(o_1 \mid s_1)P(s_2 \mid s_1)\ldots\ldots P(s_t \mid s_{t-1})P(o_t \mid s_t)$$

$$= \pi_{s_1} \prod_{i=1}^{t-1} a_{s_i s_{i+1}} \prod_{i=1}^{t} b_{s_i}(o_i)$$

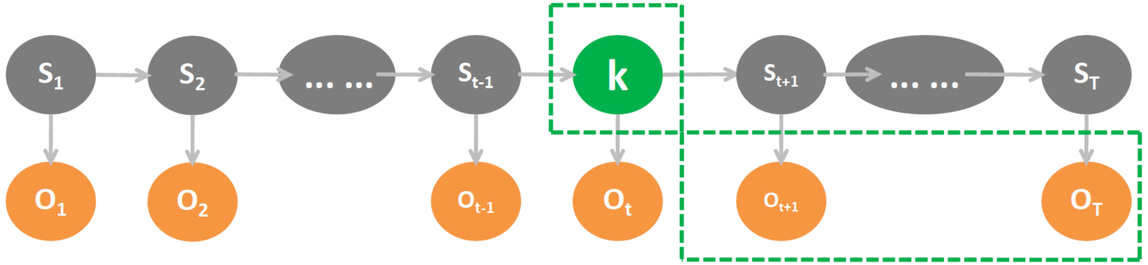
前向概率&后向概率

- 前向概率计算公式



$$\alpha_t(k) = \begin{cases} \pi_k \cdot b_k(o_1) & \text{if t = 1} \\ \sum_{l=1}^{M} b_k(o_t) \cdot a_{l,k} \cdot \alpha_{t-1}(l) & \text{otherwise} \end{cases}$$

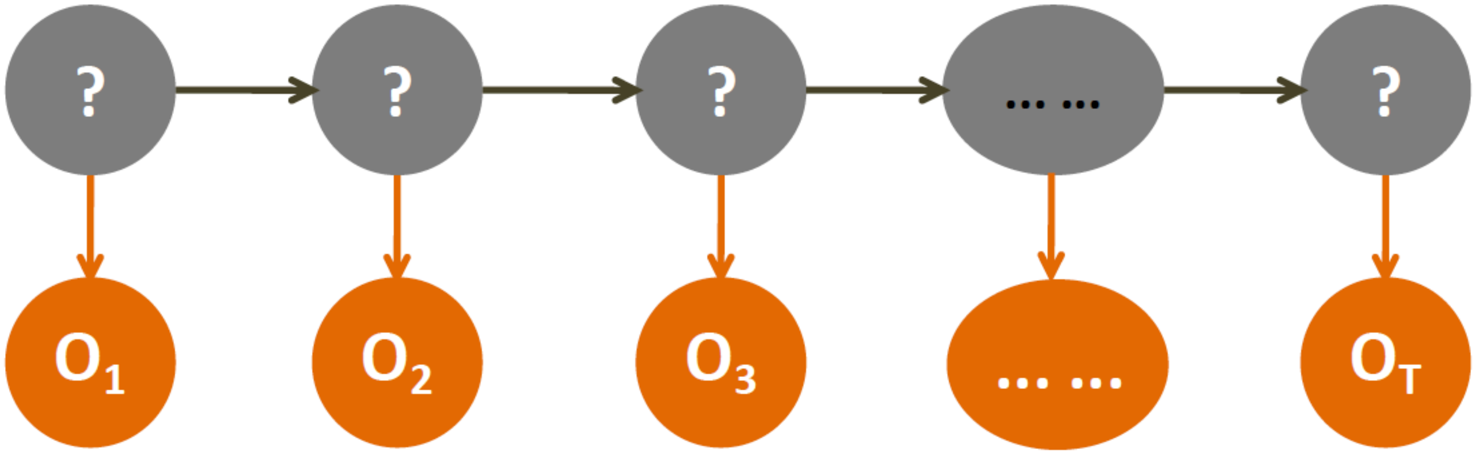
- 后向概率计算公式



$$\beta_t(l) = \begin{cases} 1 & \text{if t = T} \\ \sum_{l=1}^{M} \beta_{t+1}(l) \cdot b_l(o_{t+1}) \cdot a_{k,l} & \text{otherwise} \end{cases}$$

隐马模型的应用

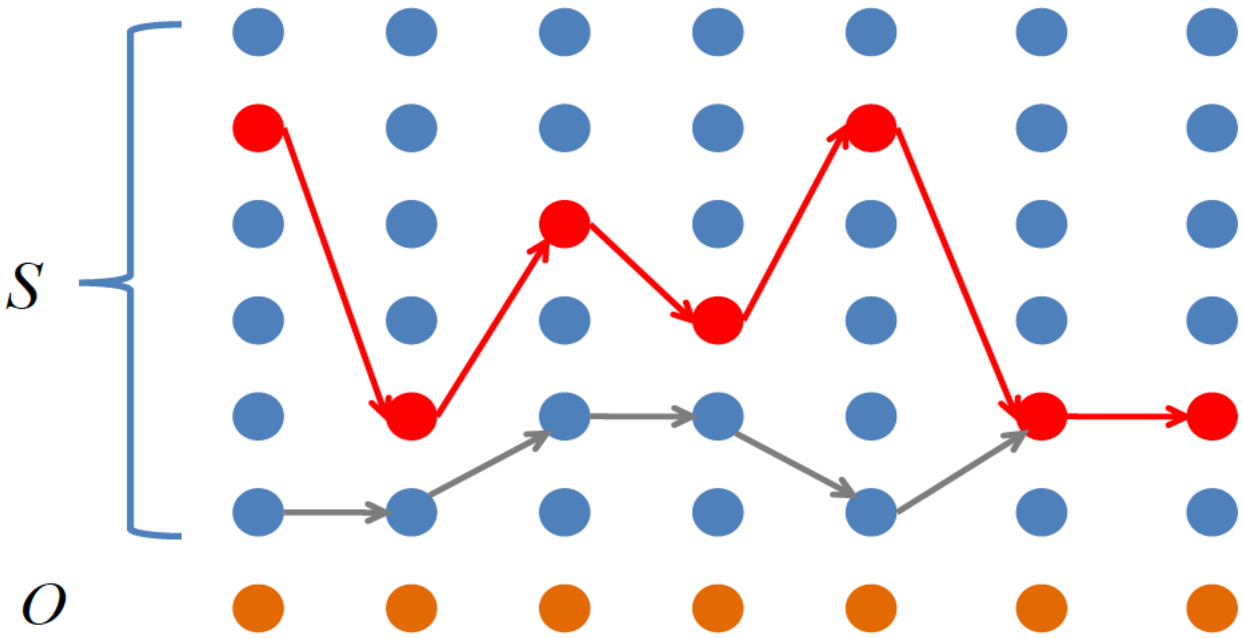
- 给定O，寻找最优的S



$$S^* = \arg\max_S P(S \mid O) = \arg\max_S \frac{P(S,O)}{P(O)} = \arg\max_S P(S,O)$$

隐马模型的应用 - Viterbi算法

- 动态规划(Dynamic Programming)，在 t+1 位置重用 t 的结果
  - 任一到达 t+1 状态的路径必然经过 t 的某一状态，并且 t+1 处的状态只依赖于 t 位置的状态，因此到 t+1 状态的最优路径必然经过 t 的某一状态的最优路径

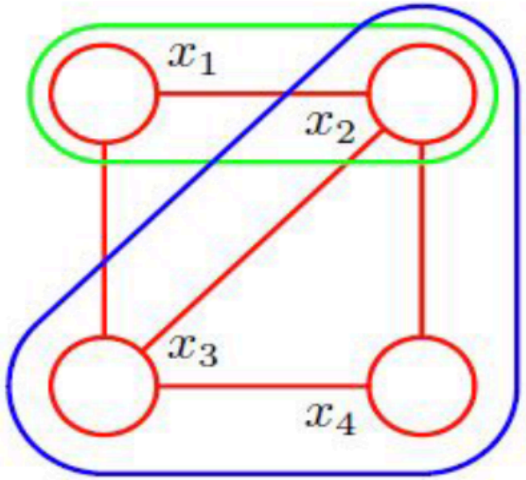


CRF – 无向图模型

- 无向图模型也叫马尔科夫随机场（Markov Random Fields）
  - 定义无向图中的一个团是一个全连通的节点集
  - 定义最大团C上的非负势函数 $\psi_C : X_C \mapsto \mathbb{R}_+$
- 无向图可以由下式表示

$$p(X) = \frac{1}{Z} \prod_C \psi_C(X_C)$$

$$Z = \int \prod_C \psi_C(X_C) dX$$



  - Z是分配函数（partition function）

CRF

- 假设G = (V, E)是一个无向图，Y是G节点上的变量，X也是随机变量
- 若Y构成一个由图G表示的马尔科夫随机场，则称P(Y|X)为条件随机场:

$$p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v),$$

其中w~v表示w和v在G中是邻居

CRF - 模型

- 假如G的结构是树形结构（包括链式），则概率模型可以表示为

$$
p_\theta(\mathbf{y} \mid \mathbf{x}) \propto \exp\left(\sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}|_e, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}|_v, \mathbf{x})\right)
$$

$$
\psi(\mathbf{h}_i, y_i, y_{i-1}) = \exp(y_i^T \mathbf{W}^T \mathbf{h}_i + y_{i-1}^T \mathbf{T} y_i),
$$

Question

- 阿里 entity
  - 拳王阿里是个传奇　the name of person entity
  - 西藏阿里美不胜收 location entity
  - 杭州阿里吸引了很多人才 organization entity
- 习近平常与特朗普通电话
  - 习近/平常/与/特朗/普通/电话

CNN-LSTM-CRF jointly train CNER and CWS

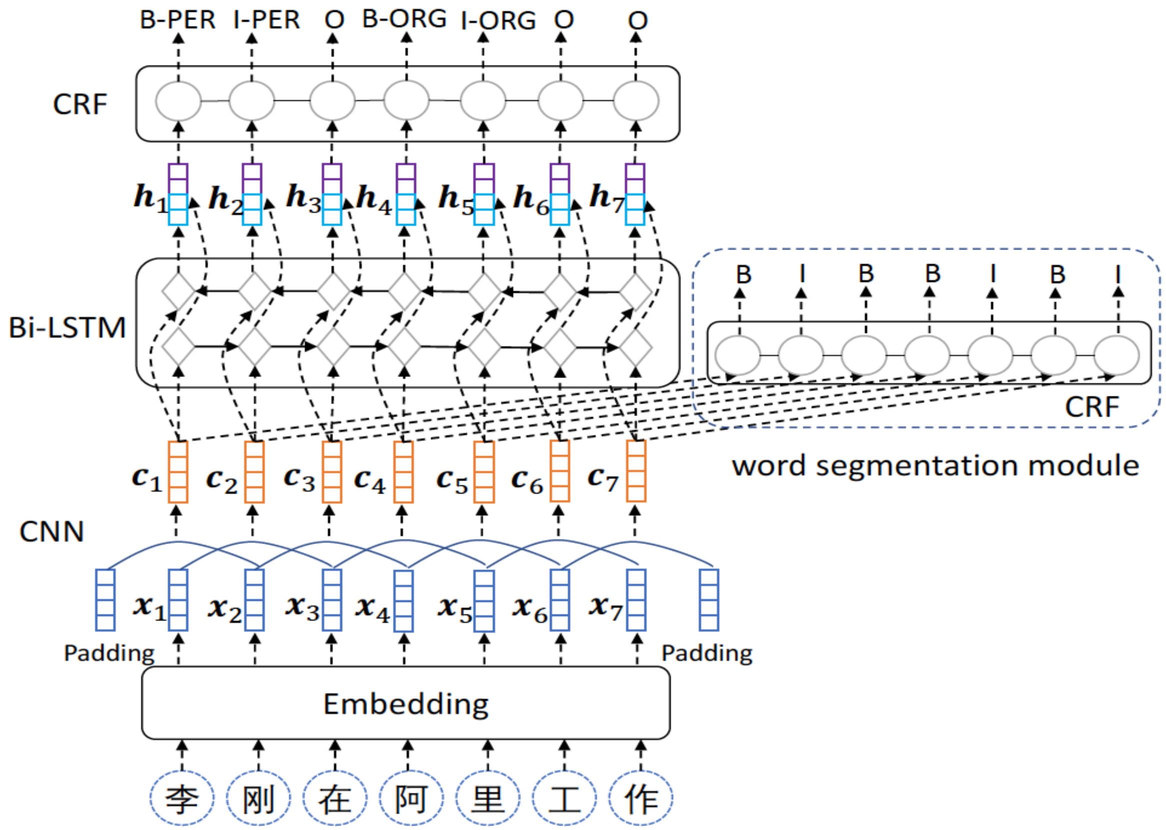


Figure 1: The framework of our approach.

$$p(\mathbf{y}|\mathbf{h};\theta) = \frac{\prod_{i=1}^{N} \psi(\mathbf{h}_i, y_i, y_{i-1})}{\sum_{\mathbf{y}' \in \mathcal{Y}(s)} \prod_{i=1}^{N} \psi(\mathbf{h}_i, y'_i, y'_{i-1})},$$

$$\mathbf{h} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N], \text{ where } \mathbf{h}_i \in \mathcal{R}^{2S}$$

$$\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N], \text{ where } \mathbf{c}_i \in \mathcal{R}^{M}.$$

$$c_i = f(\mathbf{w}^T \times \mathbf{x}_{\lfloor i-\frac{K-1}{2} \rfloor : \lfloor i+\frac{K-1}{2} \rfloor}),$$

$\mathbf{w} \in \mathcal{R}^{KD}$ 卷积核 M个，不同的K

embedding matrix $\mathbf{E} \in \mathcal{R}^{D \times V}$

vectors $[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, where $\mathbf{x}_i = \mathbf{E}w_i \in \mathcal{R}^D$.

Loss Function

$$\mathcal{L}_{NER} = -\sum_{s \in \mathcal{S}} \log(p(\mathbf{y}_s | \mathbf{h}_s; \theta)),$$

$$\mathcal{L}_{CWS} = -\sum_{s \in \mathcal{S}} \log(p(\mathbf{y}_s^{seg} | \mathbf{c}_s; \theta^{seg})),$$

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{NER} + \lambda \mathcal{L}_{CWS},$$

Pseudo Labeled Data Generation

- 李刚在阿里工作
  - 李刚 person entity
  - 阿里 company entity
- 王小超在谷歌工作 | new data
  - 王小超 person entity
  - 谷歌 company entity

replace each entity with the same concept