基于线性条件随机场的命名实体识别方法

李舒敏

Background

命名实体识别是序列标注的一个子问题，

那么，什么是序列标注?

假设，小明是个摄影狂，无时无刻都在照照片，在得到了许多张连续的照片之后，小明想给每一张照片打上标签：例如，吃饭，睡觉，喝水，etc。

- 序列标注

  Y=P(X)　—-X:输入序列；Y：预测序列

命名实体识别

- 目标：找出文本中的命名实体（人名，地名，机构名，etc）
- 输入：文本序列
- 输出：实体标签序列

- text_input:

  预约明天上午9点在八卦岭支行取五百块钱

- label_output:

  [0, 0, 'B_DATE', 'E_DATE', 'B_TIME', 'I_TIME', 'I_TIME', 'E_TIME', 0, 'B_LOCATION', 'I_LOCATION', 'E_LOCATION', 0, 0, 0, 'B_MONEY', 'I_MONEY', 'I_MONEY', 'E_MONEY']

任务驱动的多轮对话系统

- 用户：“我想订一张去北京的机票。”
- 机器人：

  第一步，意图识别 -> 订机票意图

  第二步：找出预先设置的订机票意图的信息槽 ->（出发地， 目的地，日期）

  第三步：提取用户查询里与信息槽有关的信息 ->（location:北京）

  第四步：将信息填入槽中（出发地=None,目的地=北京， 日期=None）

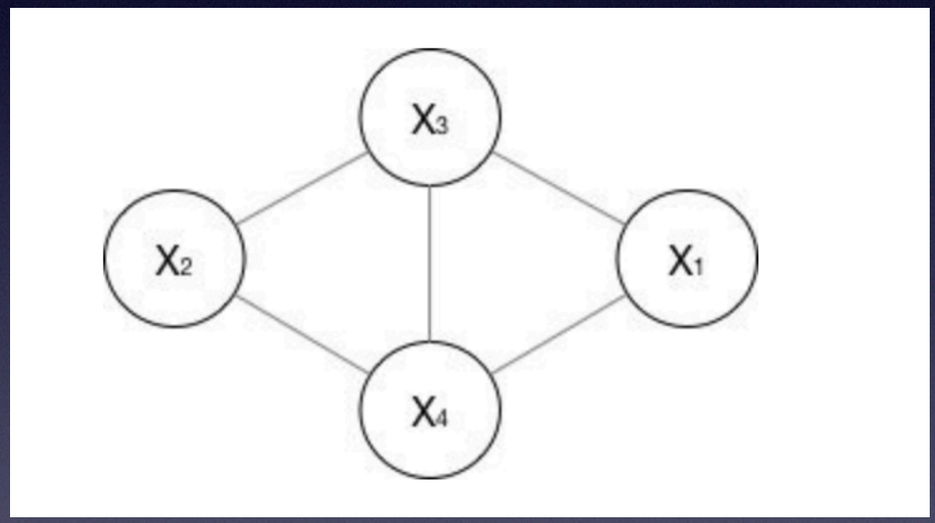
  第五步：根据缺失信息的槽，对用户进行反问。

- 机器人：“请问您想哪一天出发呢？”

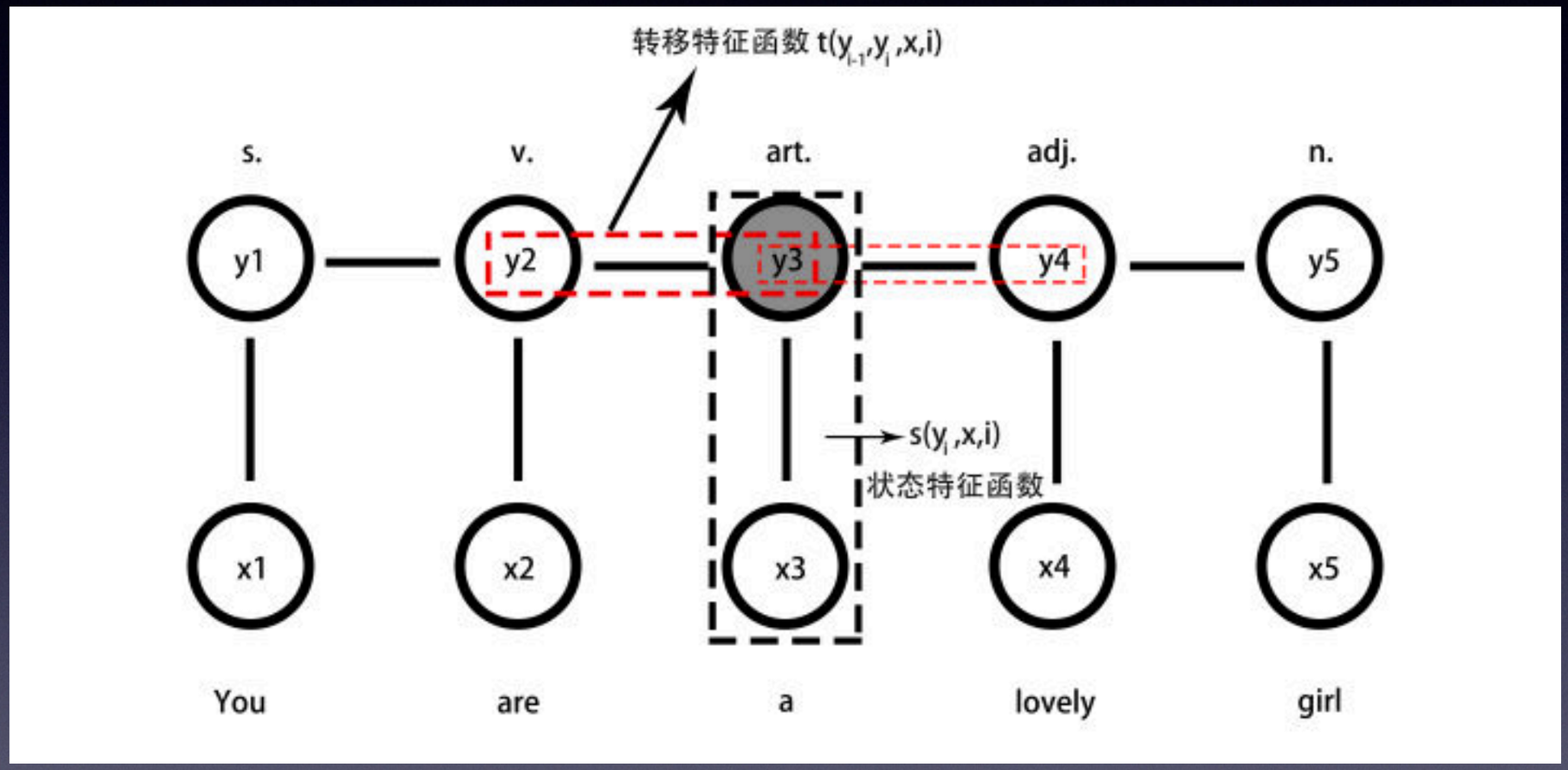
线性条件随机场

概率图模型

概率图模型是一类用图来表达变量相关关系的概率模型



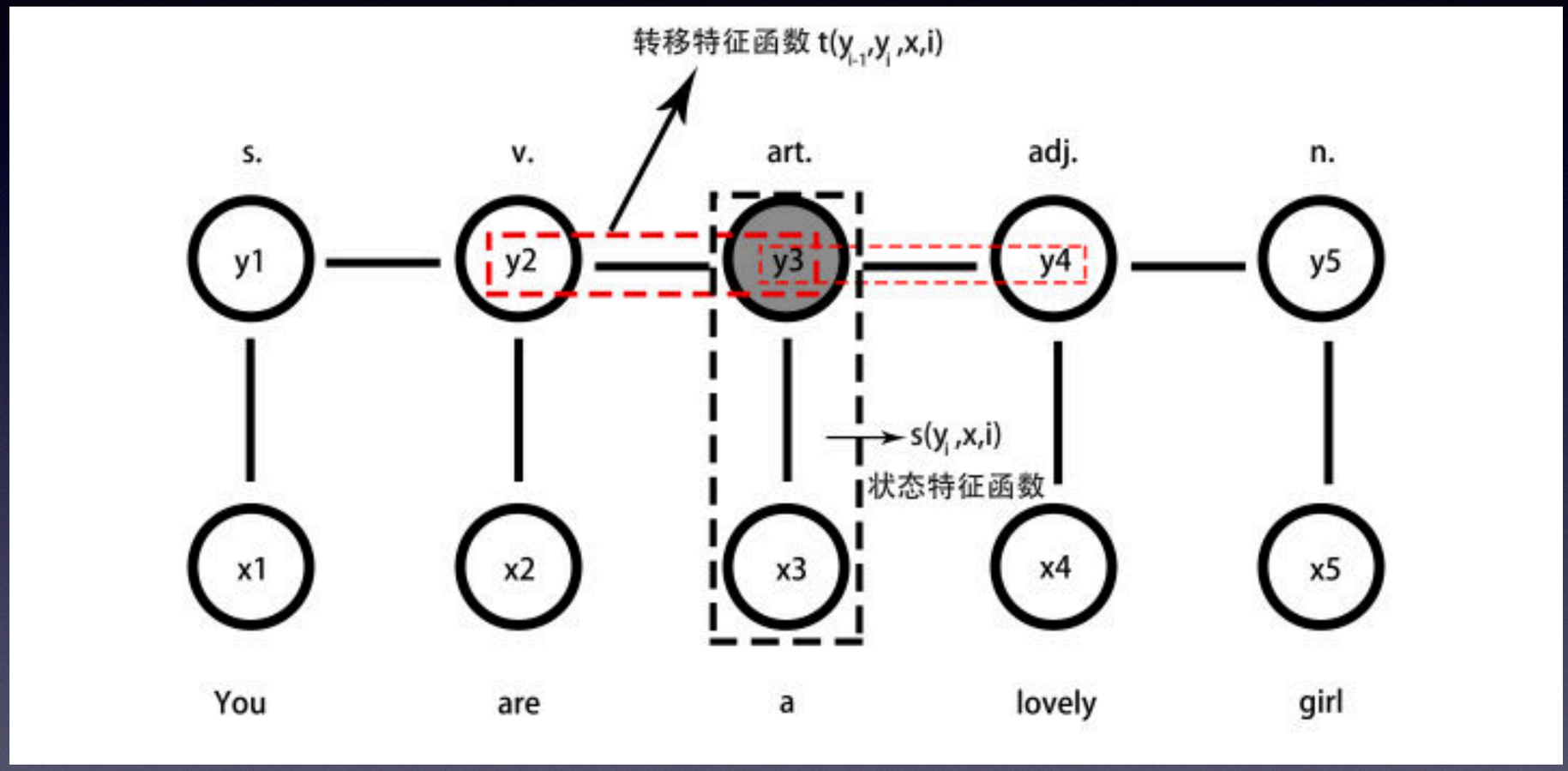
线性条件随机场



目标：计算条件概率P(Y|X)。

X为观测到的序列，Y为所有可能的预测序列。

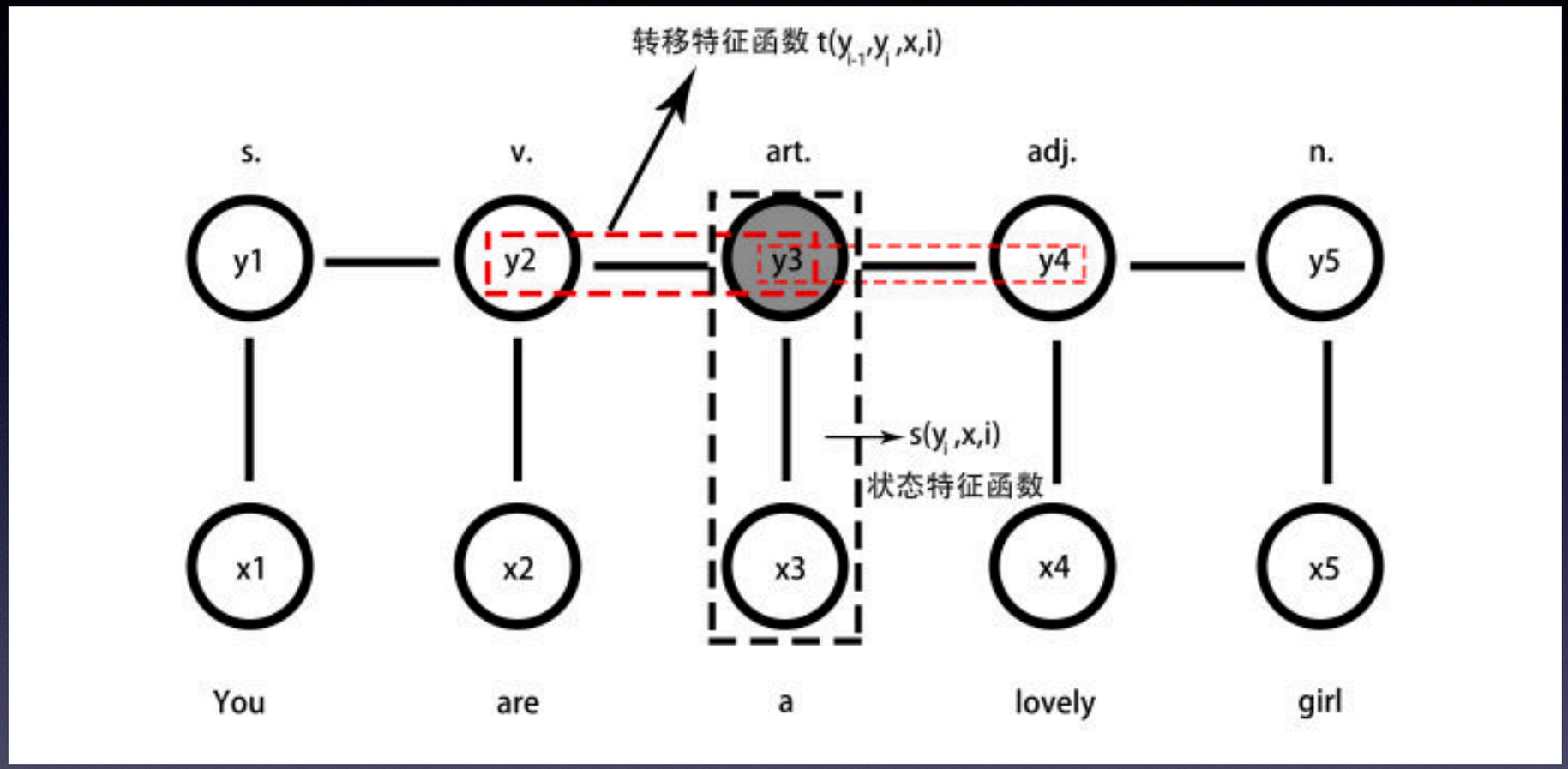
线性条件随机场



$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i))$$

$$Z(x) = \sum_{y} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i))$$

线性条件随机场在命名实体识别上的应用


转移特征函数 t(y_{i-1},y_i,x,i)
s.
v.
art.
adj.
n.
y1
y2
y3
y4
y5
s(y_i,x,i)
状态特征函数
x1
x2
x3
x4
x5
You
are
a
lovely
girl

第一部分：输入

- 如果我们的目标是给文本序列中的每一个字打标签，那么就是基于字；
- 如果我们的目标是给文本序列中每一个词打标签，那么就是基于词。

基于词：每一个x变量的取值范围为词袋的大小

缺陷：分词性能大大影响命名实体识别的性能。

基于字：每一个x变量的取值范围为字袋的大小

第二部分：特征函数

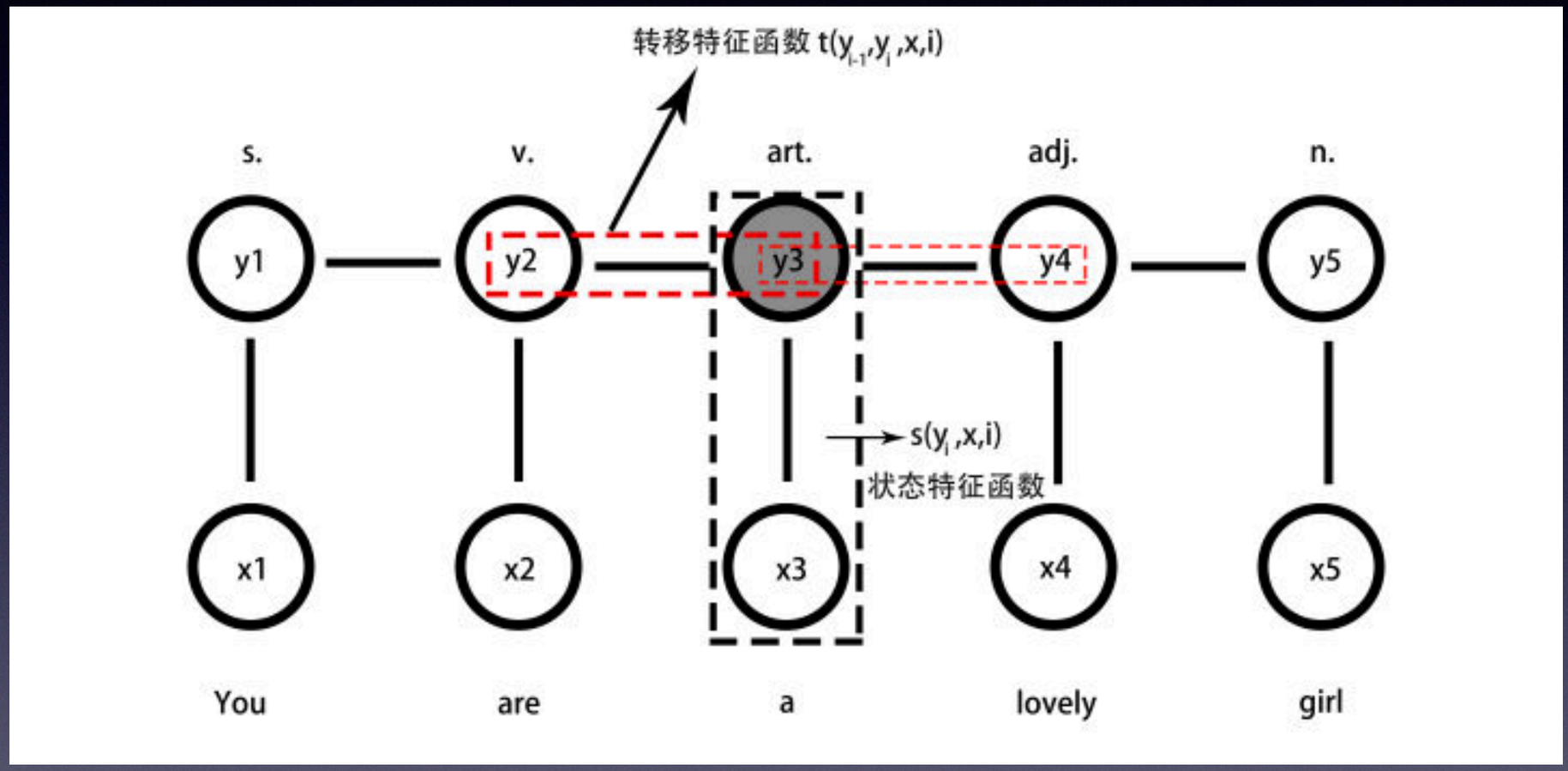
- 什么是特征函数

  特征函数是一个二值函数（值域只有0或者1），特征函数的集合为序列标注进行打分。

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i))$$

$$Z(x) = \sum_{y} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i))$$

线性条件随机场



$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i))$$

$$Z(x) = \sum_{y} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i))$$

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i))$$

$$Z(x) = \sum_{y} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i))$$

- 特征函数例子

$$t_0(y_{i-1}, y_i, \mathrm{X}, i) = \begin{cases} 1, (y_{i-1} = \mathrm{B-Person} \text{ 且} y_i \text{为 I-Person 且} \mathrm{X_i} \text{为‘杰’}) \\ 0 \end{cases}$$

$$t_1(y_{i-1}, y_i, \mathrm{X}, i) = \begin{cases} 1, (y_{i-1} = \mathrm{B-Person} \text{ 且} y_i \text{为 B-Person 且} \mathrm{X_i} \text{为‘惨’}) \\ 0 \end{cases}$$

特征函数

- 基于条件随机场的命名实体识别技术关键在于构建特征函数。

$$t_4(y_{i-1}, y_i, X, i) = \begin{cases} 1, (y_{i-1} = B-Person \text{ 且} y_i \text{为 } B-Person \text{ 且} X_i \text{的词性为名词}) \\ 0 \end{cases}$$

第三部分：预测

- 假设一句文本有L个字，实体类标有M个，那么这句文本预测的可能性有M的L次方个。
- 根据特征函数进行打分，选择打分最大的。

第三部分：预测

- 例如：我今天下午要去哈工大深圳。
- 所有可能性：
- OOOOOOOOOOOO；
- OOOOOOOOOOOB-organization；
- ……
- OB-dateE-dateB-timeE-timeOOB-organizationI-organizationI-organizationI-organizationE-organization

实际实现

- 工具：CRF++；

根据特征模板自动生成特征函数。

- 特征：字，词性，词边界，命名实体列表

Thank you!