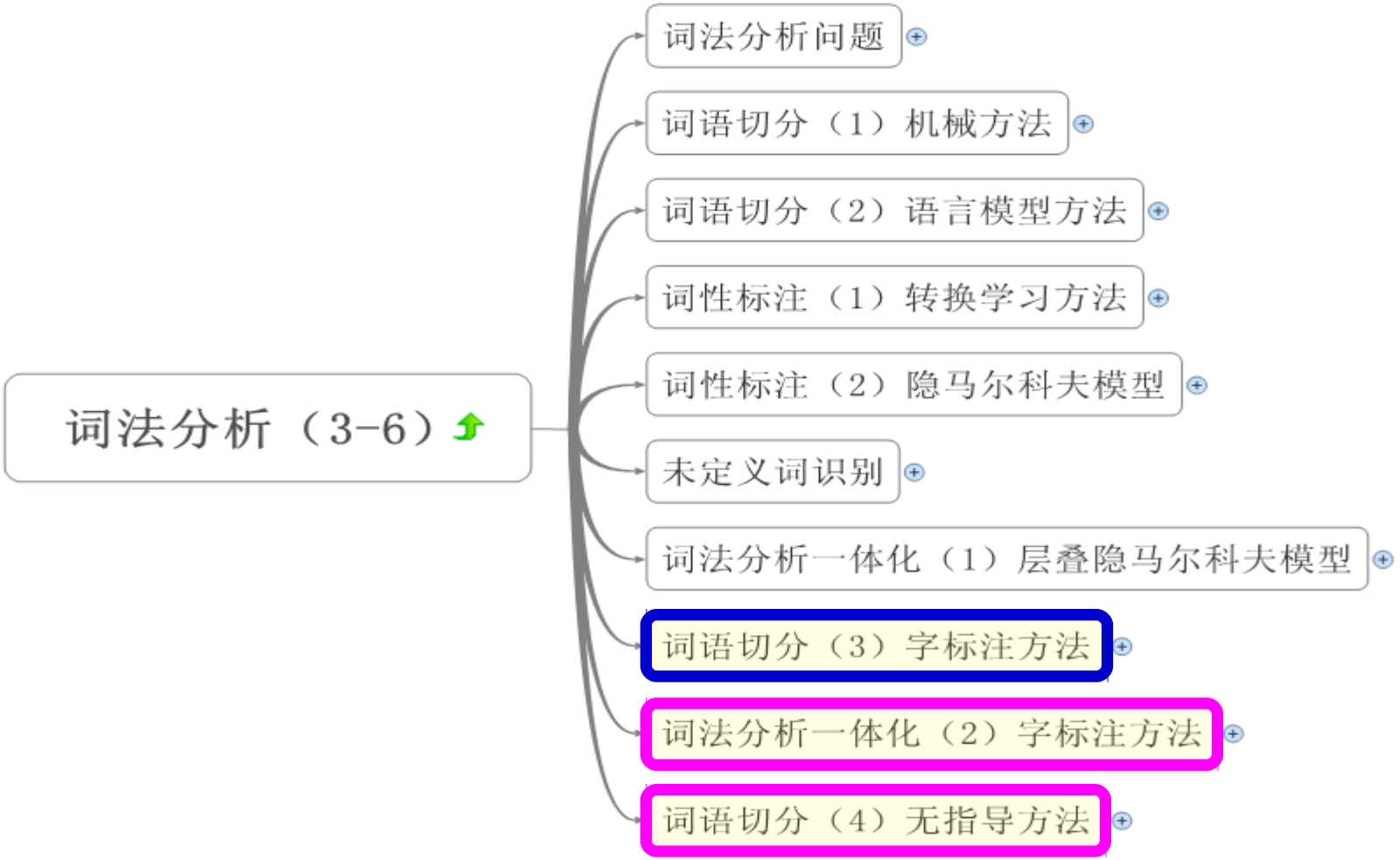
计算语言学

第 6 讲　词法分析（四）

刘群
中国科学院计算技术研究所
liuqun@ict.ac.cn

中国科学院研究生院 2012 年春季课程讲义

内容提要


词法分析（3-6）
词法分析问题
词语切分（1）机械方法
词语切分（2）语言模型方法
词性标注（1）转换学习方法
词性标注（2）隐马尔科夫模型
未定义词识别
词法分析一体化（1）层叠隐马尔科夫模型
词语切分（3）字标注方法
词法分析一体化（2）字标注方法
词语切分（4）无指导方法

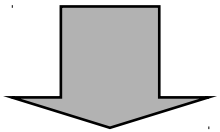
基于字标注的中文词法分析

- Nianwen Xue and Libin Shen. 2003. Chinese word segmentation as LMR tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL'03*, pages 176–179, Sapporo, Japan.

空挡标注

- 最简单的分词方案，可以理解为：
  对句子中每两个汉字之间的空挡判断是否进行切分

费 0 孝 0 通 1 向 1 人 0 大 1 报 0 告

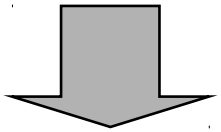


费孝通 向 人大 报告

字标注

- 对每一个汉字进行标注 {B,M,E,S} ：
  - B ： 词首字
  - M ： 词中字
  - E ： 词尾字
  - S ： 单字词

费 /B 孝 /M 通 /E 向 /S 人 /B 大 /E 报 /B 告 /E



费孝通 向 人大 报告

空挡标注与字标注的转换

- 上述两种标注是可以转换的：
  - 字标注可以通过该字左右的空挡标注得到：
    - B➔10
    - M➔00
    - E➔01
    - S➔11

更复杂的字标注

- Hai Zhao, Chang-Ning Huang, and Mu Li, An Improved Chinese Word Segmentation System with Conditional Random Field, Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing (SIGHAN-5), pp.162-165, Sydney, Australia, July 22-23, 2006
- 采用基于字的六标注集合： B 、 B_1 、 B_2 、 M 、 E 、 S
  - 单字词： S
  - 两字词： BE
  - 三字词： BB_1E
  - 四字词： BB_1B_2E
  - 五字词： BB_1B_2ME
  - 六字词： BB_1B_2MME
- 问题：六字标注集如何表示为空挡标注？

字标注模型

- 字标注（或空挡标注）都是序列标注问题
- 理论上，字标注问题也可以采用语言模型或者隐马尔科夫模型来解决
- 但由于标记集太小，采用语言模型和隐马尔科夫模型很难取得很好的效果：语言模型和隐马尔科夫模型的区分能力太弱

更复杂的字标注模型

- 最大熵模型
- 最大熵马尔科夫模型
- 条件随机场模型
- 感知机模型

最大熵原理

- Berger, A.L., Della Pietra, S.A., Della Pietra, V.J., (1996), A Maximum Entropy Approach to Natural Language Processing, Computational Linguistics, Volume 22, No. 1
- 自然语言处理的最大熵模型，常宝宝，北京大学
- 自然语言处理中的最大熵方法（ PPT 讲义），马金山，哈尔滨工业大学信息检索研究室

  （本讲义部分内容源自马金山 PPT ，特此感谢）

什么是熵

- 什么是熵？ 没有什么问题在科学史的进程中曾被更为频繁地讨论过

普里高津

- 熵定律是自然界一切定律中的最高定律

里夫金 & 霍华德

熵的提出

- 热力学第二定律（ Second Law of Thermodynamics ）认为：物理过程总是自发地从有序走向无序，最后达到“热寂” 。
- 德国物理学家克劳修斯（ Rudolph J.E clausius ）从热力学第二定律出发，于 1865 提出熵的概念用来描述一个系统的无序度（ la degré de désordre ）。 因此热力学第二定律又被称为“增熵原理”，即系统的演进总是指向熵增加的方向。
- 克劳修斯的熵概念这是在热力学角度提出的，之后被 Boltzmann 通过统计物理学的角度重新诠释。

熵与信息

- 熵表示了一个系统的不确定性
- 信息可以理解为事件不确定性的减少
  - 原来不确定的事情现在确定下来，就是获得了信息
  - 原来不确定性越大的事情发生了，获得的信息越多
    - 狗咬人不是新闻，人咬狗才是新闻

信息熵

- 1948 年电气工程师香农 (Shannon) 创立了信息论，将信息量与熵联系起来。
- 他用非常简洁的数学公式定义了信息时代的基本概念：信息熵

随机事件的熵

- 熵定量的描述事件的不确定性
- 设随机变量 ξ，它有 A_1， A_2，…， A_n 共 n 个可能的结局，每个结局出现的机率分别为 p_1, p_2， … ，p_n 则 ξ 的不确定程度，即信息熵为：

$$H(\xi) = -\sum_{i=1}^{n} p_i \log p_i$$

- 熵越大，越不确定
- 熵等于 0，事件是确定的
- 通常对数底取 2，熵的单位为比特（ bit ）

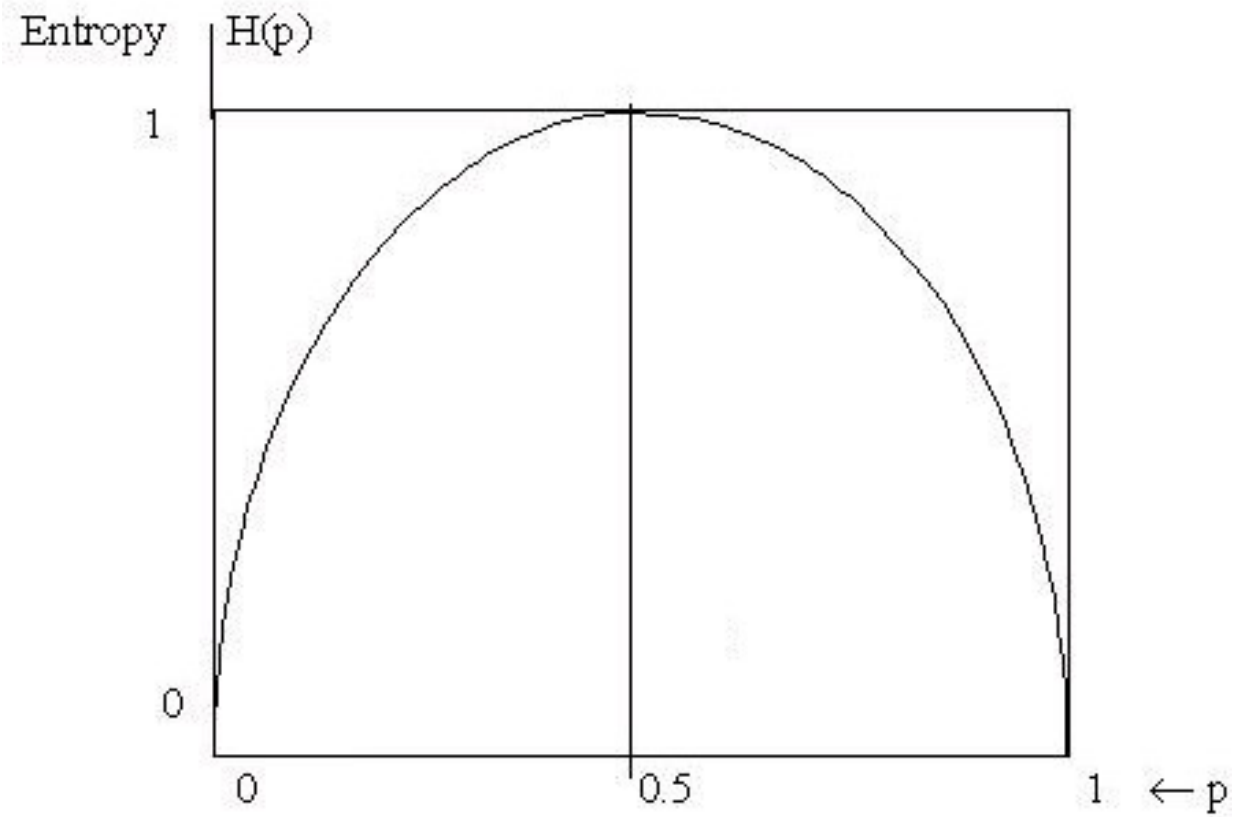
例子

- 抛硬币
  - X={ 正面，反面 }
  - p(正面)=p(反面)=0.5
  - H(X)=-(0.5log0.5+0.5log0.5)=1
- 掷骰子
  - X={1， 2， 3， 4， 5， 6}
  - p(1)=p(2)=p(3)=p(4)=p(5)=p(6)=1/6
  - H(X)=log(6)=2.585
- 不对称的硬币
  - p(正面)=0.3, p(反面)=0.7
  - H(X)=-(0.3log0.3+0.7log0.7)=0.881

通信中的熵

- 表示“是” 和 “否”
- 1 = 是 0 = 否
- 表示“是” 、“否”和“可能是”
- 11 = 是 00 = 否 10(01) = 可能是
- 一条消息的熵就是编码这条消息所需二进制位即比特的个数。

二元事件的熵



信息熵的意义

- 信息熵概念为测试信息的多少找到了一个统一的科学定量计量方法，是信息论的基础。

- 信息熵将数学方法和语言学相结合

最大熵理论

- 熵增原理
- 在无外力作用下，事物总是朝着最混乱的方向发展
- 事物是约束和自由的统一体
- 事物总是在约束下争取最大的自由权，这其实也是自然界的根本原则。
- 在已知条件下，熵最大的事物，最可能接近它的真实状态

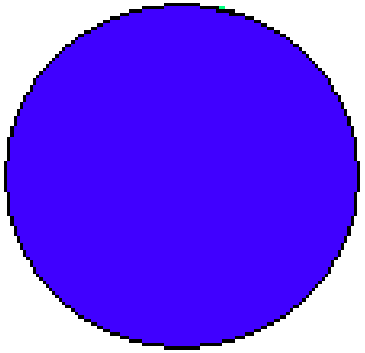
基于最大熵原理选择模型

- 研究某个随机事件，根据已知信息，预测其未来行为。
- 当无法获得随机事件的真实分布时，构造统计模型对随机事件进行模拟。
- 满足已知信息要求的模型可能有很多个，用那个模型来预测最合适呢？

基于最大熵原理选择模型

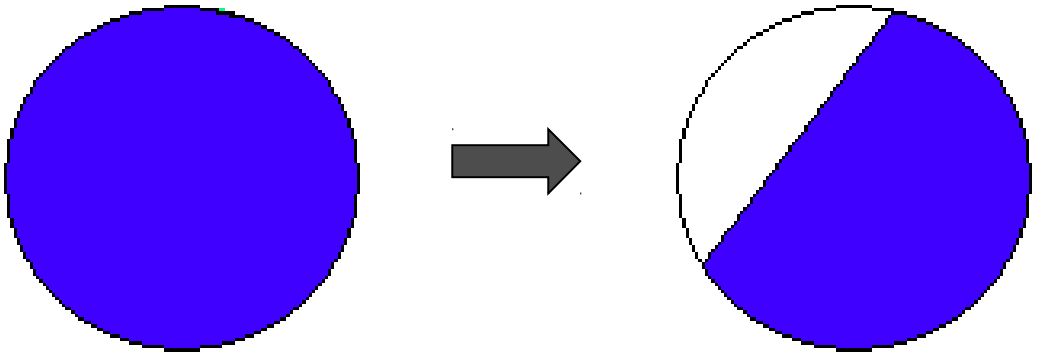
- 选择熵最大的模型
- Jaynes 证明：对随机事件的所有相容的预测中，熵最大的预测出现的概率占绝对优势
- Tribus 证明，正态分布、伽玛分布、指数分布等，都是最大熵原理的特殊情况

最大熵原则下点的分布



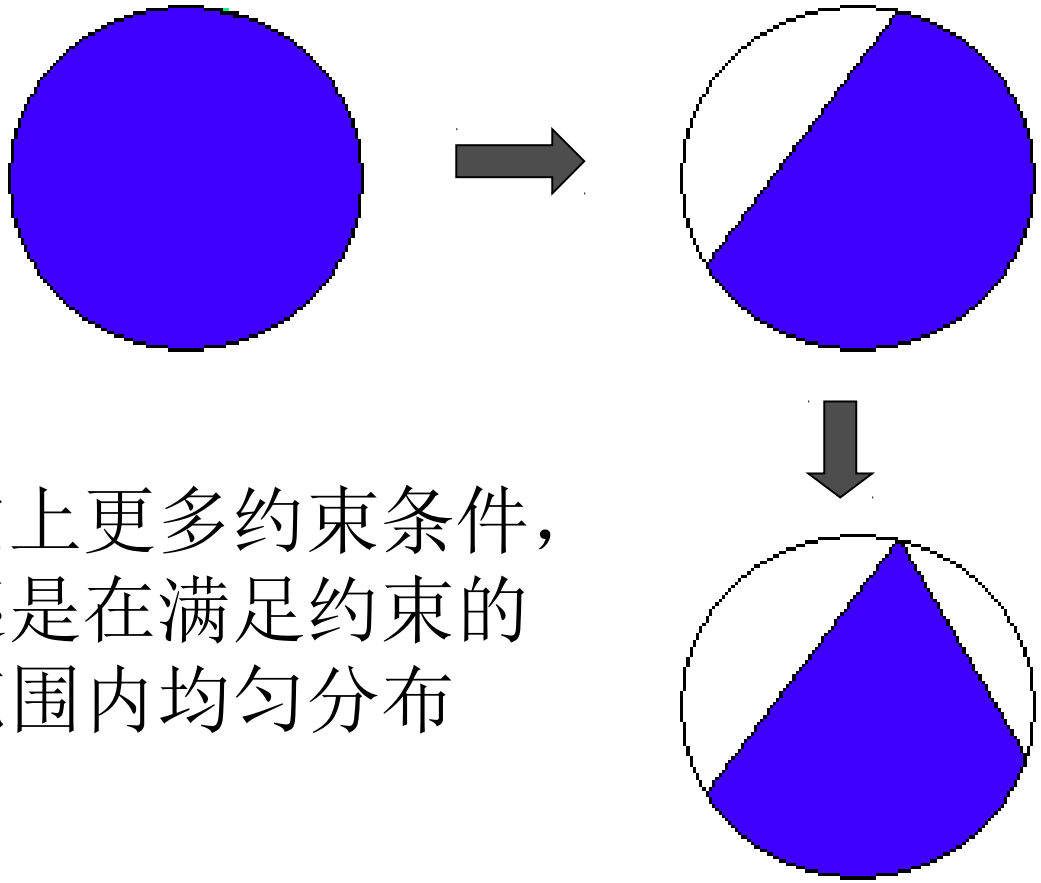
对一随机过程，如果没有任何观测量，
既没有任何约束，则解为均匀分布

最大熵原则下点的分布



加上一个约束条件，
则在满足约束的范围内均匀分布

最大熵原则下点的分布



加上更多约束条件，
还是在满足约束的
范围内均匀分布

一个简单的例子

- 给定一个随机变量 X={A,B,C,D}
- 如果不给定任何约束，那么 X 的熵最大的概率分布形式为平均分布：

  p([0.25, 0.25, 0.25, 0.25]) = 2.000
- 如果给定约束条件 p(A)+p(B)=0.6 ，那么熵最大的分布为满足这个条件下的均匀分布：

  p([0.30, 0.30, 0.20, 0.20]) = 1.971
- 如果给定两个约束条件： p(A)+p(B)=0.6 & p(A)+p(C)=0.6 那么熵最大的分布应该是怎样的呢？

  p([0.30, 0.30, 0.30, 0.10]) = 1.895

  p([0.35, 0.25, 0.25, 0.15]) = 1.941

  p([0.40, 0.20, 0.20, 0.20]) = 1.922

  ……

天气预报

- 假设要用今天的天气预测明天的天气
  - 天气∈{晴、阴、雨}
  - 风向∈{无、南、北}
  - 已知今天的天气和风向，要预测明天的天气
- 已知历史天气数据（见下页表）
- 假设要用今天的天气预测明天的天气
- 已知今天天气为阴，风向北，明天天气最可能是什么？

天气预报：样本

昨天风向	昨天天气	今天天气
无	雨	晴
南	阴	阴
北	晴	晴
北	晴	晴
南	阴	雨
无	晴	晴
北	晴	阴
无	雨	晴
南	阴	雨
北	晴	阴

问题定义

- 用 A 表示条件集合， B 表示结论集合。随机事件 $X=(\alpha,\beta)\in\xi=A\times B,\ \alpha\in A,\ \beta\in B$
- 我们已经有一些 X 的样本：

  $X_1=\{\alpha_1, \beta_1\}, X_2=\{\alpha_2, \beta_2\}, \ldots, X_n=\{\alpha_n, \beta_n\}$
- 假设我们已知 α_{n+1} ， 如何预测 β_{n+1} ？
- 求解一个 X 的概率分布，使得：
  - X 的分布与已知的样本分布一致
  - X 的熵最大

天气预报：建模

- 建立天气预报模型：给出所有可能的条件下结论的概率 $p(\beta|\alpha)$，$\alpha \in A$，$\beta \in B$
- 为简化问题起见，我们通常用联合概率模型取代上述的条件概率模型 $p(\alpha,\beta)$
- 二者关系：$p(\beta|\alpha)= p(\alpha,\beta)/p(\alpha)$
- 由于我们考虑的条件和结论都是离散量，理论上，所有的可能性是可以穷举的，我们只要给出所有可能性的概率即可

天气预报：模型

所有的可能性

昨天风向	昨天天气	今天天气	概率
无	晴	晴	0.011
无	晴	阴	0.003
无	晴	雨	0.002
无	阴	晴	0.008
无	阴	阴	0.013
无	阴	雨	0.005
……	……	……	
北	雨	晴	0.009
北	雨	阴	0.001
北	雨	雨	0.003

概率之和为一

天气预报：建模

- 我们希望得到的上述模型满足以下条件：
  - 模型的概率分布应尽可能与样本一致
    - 模型中，“晴、阴、雨”天的概率分布应与样本一直
    - 模型中，昨天为“晴”、“阴”、“雨”时第二天的天气概率分布应与样本一致
    - 模型中，昨天为“无风”、“南风”、“北风”时第二条的天气概率分布应与样本一致
    - 模型中，昨天天气与风向组合给定的情况下，第二条的天气概率分布应与样本一致
    - ……
  - 模型的熵最大！

问题定义：优化目标

- X 熵最大：可以表示为：

$$\hat{p} = \underset{p}{\operatorname{argmax}} H(p)$$

- X 的分布与已知的样本分布一致：如何表示？
- 实际上，由于 A （预测的条件）的可能性数量可能极其巨大，穷举所有可能性也是不现实的，如何对任意的条件和结论给出其概率？

问题定义：引入特征

- 在实际问题中，由于条件 α 和结果 β 取值比较多样化，为模型表示方便起见，通常我们将条件 α 和结果 β 表示为一些二制特征
- 特征 $f_1,f_2,\ldots,f_n$ 定义如下：

  f_i :ε→ {0,1}, ε=A(预测条件)×B(结论)
- 这样每个事件都可以表示为一个由特征值 {0,1} 组成的 n 维向量，不再直接用条件和结论来描述。

问题定义：引入约束

- “模型分布与样本分布一致”可以描述为：

$E_p[f_i]=e_{\hat{p}}[f_i], i=1...n$, 这里 p 为样本分布， $\hat{p}$ 为模型分布

- 这个公式含义为：对于任何一个特征，模型和样本应该具有相同的均值。由于特征取值只有 {0,1}，因此这个公式实质上可以理解为：模型中任何一个特征为 1 的概率与样本应相同

天气预报：简单特征

- 为天气预报引入以下简单特征：
  - f1：昨天天气 = 晴
  - f2：昨天天气 = 阴
  - f3：昨天天气 = 雨
  - f4：昨天风向 = 无
  - f5：昨天风向 = 南
  - f6：昨天风向 = 北
  - f7：今天天气 = 晴
  - f8：今天天气 = 晴
  - f9：今天天气 = 晴

天气预报：用特征表示样本

昨天风向			昨天天气			今天天气		
无	南	北	晴	阴	雨	晴	阴	雨
1	0	0	0	0	1	1	0	0
0	1	0	0	1	0	0	1	0
0	0	1	1	0	0	1	0	0
0	0	1	1	0	0	1	0	0
0	1	0	0	1	0	0	0	1
1	0	0	1	0	0	1	0	0
0	0	1	1	0	0	0	1	0
1	0	0	0	0	1	1	0	0
0	1	0	0	1	0	0	0	1
0	0	1	1	0	0	0	1	0

天气预报：用特征表示模型

昨天风向			昨天天气			今天天气			概率
无	南	北	晴	阴	雨	晴	阴	雨	
0	0	0	0	0	0	0	0	0	0.005
0	0	0	0	0	0	0	0	1	0.003
0	0	0	0	0	0	0	1	0	0.004
0	0	0	0	0	0	0	1	1	0
0	0	0	0	0	0	1	0	0	0.001
0	0	0	0	0	0	1	0	1	0
…	…	…	…	…	…	…	…	…	……
1	1	1	1	1	1	1	0	1	0
1	1	1	1	1	1	1	1	0	0
1	1	1	1	1	1	1	1	1	0

所有的可能性

概率之和为一

天气预报：用简单特征建模

- 利用上述简单特征，采用最大熵原理为天气预报建模，我们得到的模型将是：
  - 模型中所有特征出现概率与样本一致：
    - 模型中“昨天天气为晴”概率与样本一致
    - 模型中“昨天天气为阴”概率与样本一致
    - ……
    - 模型中“今天天气为雨”概率与样本一致
  - 模型熵最大
- 这个模型显然不是我们想要的模型：
  - 我们希望：
    - 模型中“昨天天气为晴”的情况下今天各种天气的概率与样本一致
    - 模型中“昨天风向为南”的情况下今天各种天气的概率与样本一致
    - 模型中昨天天气和风向给定情况下今天各种天气的概率与样本一致
    - ……

天气预报：复合特征

- 为了使得到的模型满足我们所期望的约束条件，我们可以重新定义特征：
  - f1: 昨天天气为晴，且今天天气为晴
  - f2: 昨天天气为晴，且今天天气为阴
  - ……
  - fn: 昨天风向为北，天气为雨，且今天天气为雨
- 按照这种方式定义特征，那么所有的特征都反映了某种预测条件和结果之间的依赖关系，这样得到的模型才是我们所期望的模型。

天气预报：用特征表示模型

所有的可能性

f_1	f_2	f_3	f_4	f_5	…	f_{n-2}	f_{n-1}	f_n	
0	0	0	0	0	…	0	0	0	0.005
0	0	0	0	0	…	0	0	1	0.003
0	0	0	0	0	…	0	1	0	0.004
0	0	0	0	0	…	0	1	1	0.002
0	0	0	0	0	…	1	0	0	0.001
0	0	0	0	0	…	1	0	1	0.003
…	…	…	…	…	…	…	…	…	……
1	1	1	1	1	…	1	0	1	0.003
1	1	1	1	1	…	1	1	0	0.007
1	1	1	1	1	…	1	1	1	0.010

概率之和为一

天气预报：模型应用

- 假设我们已经得到一个满足上述约束条件且熵最大的模型，如何预测天气？
- 给定预测条件 α （今天风向和今天天气），需要预测明天天气 β ∈{晴，阴，雨}，我们只要根据模型分别计算（α，晴）/（α，阴）/（α，雨）情况下的所有特征值，根据模型计算出相应的概率，取概率最大者即可。

问题定义：最优化

- 根据前面的定义，最大熵模型的参数估计可以表示为一个约束条件下的极值问题；
  - 在以下约束条件下：

$E_p[f_i]=e_{\hat{p}}[f_i], i=1...n$, 这里$p$为样本分布，$\hat{p}$为模型分布

  - 求解熵最大的模型：

$\hat{p}=\arg\max_{p\in C} H(p)$, C为所有可能的模型组成的集合

问题定义：最优化

- 经推导（过程略），满足上述约束条件的最大熵模型具有如下形式：

$$p(\alpha,\beta)=\frac{\exp\left(\sum_{(i=1)}^{n}\lambda_i f_i(\alpha,\beta)\right)}{\Pi}$$

这里， Π 是一个归一化参数，是个常量。

λ_i 是一组参数，其中每个参数对应于一个特征

最大熵模型用于预测

- 给定条件 α，结论为 β 的概率可以表示为：

$$p(\beta|\alpha)=\frac{\exp(\sum_{(i=1)}^{n}\lambda_i f_i(\alpha,\beta))}{Z}$$

- 给定条件 α，最优的 β 可以表示为：

$$\hat{\beta}=\arg\max_{\beta}(\beta|\alpha)=\arg\max_{\beta}\frac{\exp(\sum_{i=1}^{n}\lambda_i f_i(\alpha,\beta))}{Z}$$

$$=\arg\max_{\beta}\sum_{i=1}^{n}\lambda_i f_i(\alpha,\beta)$$

- 可以看出，一旦得到这组参数 λ_i，那么对于给定的条件 α，对所有的结论 β，只要将其所有值为 1 特征 f_i 对应的 λ_i 加起来，取和最大的 β 即可

最大熵模型参数估计

- GIS （ Generalized Iterative Scaling ）
- IIS （ Improved Iterative Scaling ）

最大熵模型的特征选择

- 利用特征集合的信息增益来选择特征集合
- 由于选择合适的特征集合是一个 NP 问题（背包问题），我们通常采用简化的贪心算法来解决
- 更简单的做法：利用样本中该特征出现的频度，选择所有频度大于某个阈值的特征

最大熵工具

- 最普遍使用的工具：
  - 名称：
    Maximum Entropy Modeling Toolkit for Python and C++
  - 作者：张乐（东北大学博士生）
  - 主页：
    http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

判别式模型 vs. 生成式模型

- 生成式模型：
  - N 元语法模型、隐马尔科夫模型、 PCFG 模型……
  - 特征和特征之间的关系直接体现在公式中
  - 特征是给定的
  - 特征之间的权重是固定的
- 判别式模型：
  - 最大熵、条件随机场、感知机……
  - 特征在公式中表现为任意的特征函数，特征直接可以有参数加权，通过对参数进行调节来逼近优化目标
  - 特征可以任意选择
  - 特征之间的权重是可以调节的
- 判别式模型更灵活，通常效果更好
- 判别式模型需要有指导训练，生成式模型可以进行无指导训练

基于最大熵模型字标注的汉语词语切分

- 采用最大熵模型对每个汉字标注 BMES 标记
- 假设当前字 C_0 ，当前标记是 T_0
- 常用最大熵特征模板（当前字是 C_0 ）：
  - $C_n T_0 (n = -2, -1, 0, 1, 2)$ ：汉字
  - $C_n C_{n+1} T_0 (n = -2, -1, 0, 1)$ ：两字组
  - $C_{-1} C_1 T_0$ ： 当前字左右两个字
  - $D(C_0) T_0$ ： 当前字是否数字
  - $A(C_0) T_0$ ： 当前字是否字母
  - $P(C_0) T_0$ ： 当前字是否标点

生成最大熵训练实例

- 假设给定训练样本：

  <s> 今天　是　星期三　。 </s>

- 生成对应标注序列：

  今 |B　天 |E　是 |S　星 |B　期 |M　三 |E　。 |S

- 对于每一个字生成一个训练实例。

生成最大熵训练实例

- 上例中“天”字生成的训练实例：

$C_{-2}T_0$=<s>E	$C_{-1}T_0$= 今 E	C_0T_0= 天 E	C_1T_0= 是 E
C_2T_0= 星 E	$C_{-2}C_{-1}T_0$=<s> 今 E	$C_{-1}C_0T_0$= 今天 E	
$C_0C_1T_0$= 天是 E	$C_1C_2T_0$= 是星 E	$C_{-1}C_1T_0$= 今是 E	
$D(C_0)T_0$= 否	$A(C_0)T_0$= 否	$P(C_0)T_0$= 否	

- 这里列出了该实例中被激活（值为 1 ）的所有特征，其他所有未列出的特征均未被激活（值为 0 ）

最大熵模型的参数训练

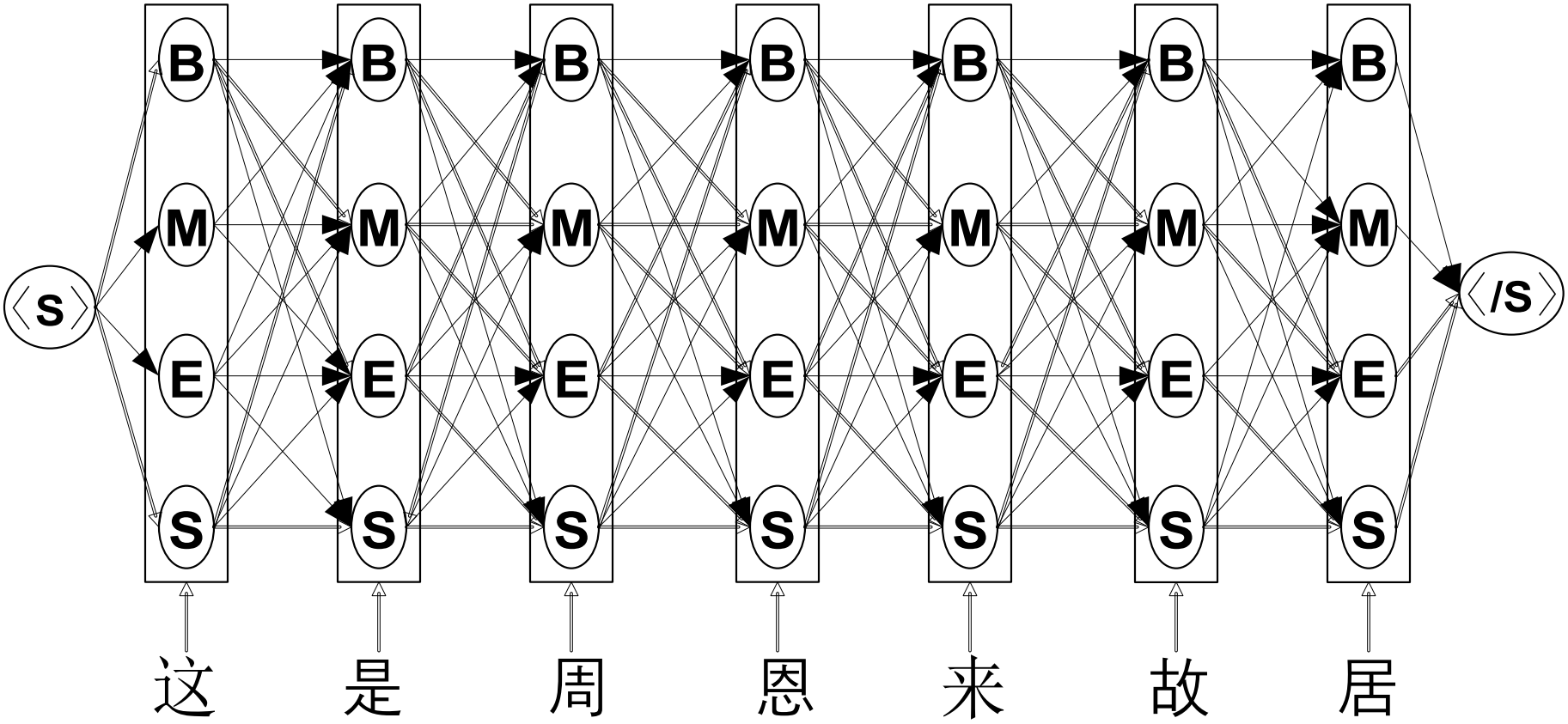
- 对训练语料库中每个汉字生成一个实例
- 把所有实例的列表送给最大熵模型的训练工具
- 最大熵模型的训练工具将为每一个训练实例生成一个参数 λ

最大熵模型的使用

- 输入的汉字串，如“昨天是星期五”
- 对于输入串中每一个汉字（这里假设是“天”字），以及该汉字的每一个可能的标记，生成一个实例，这样对每个汉字就生成了四个实例，如 E(天→ B),E(天→ M),E(天→ E),E(天→ S) 。
- 对这每个实例 E(天→ X) ，生成其所有激活的特征，计算这些特征所对应的参数 λ 之和 N(天→ X)
- 可能性最大的标记为：

$$X=\max_{X'\in\{B,M,E,S\}} N(\text{天}\rightarrow X')$$

搜索最优标注路径

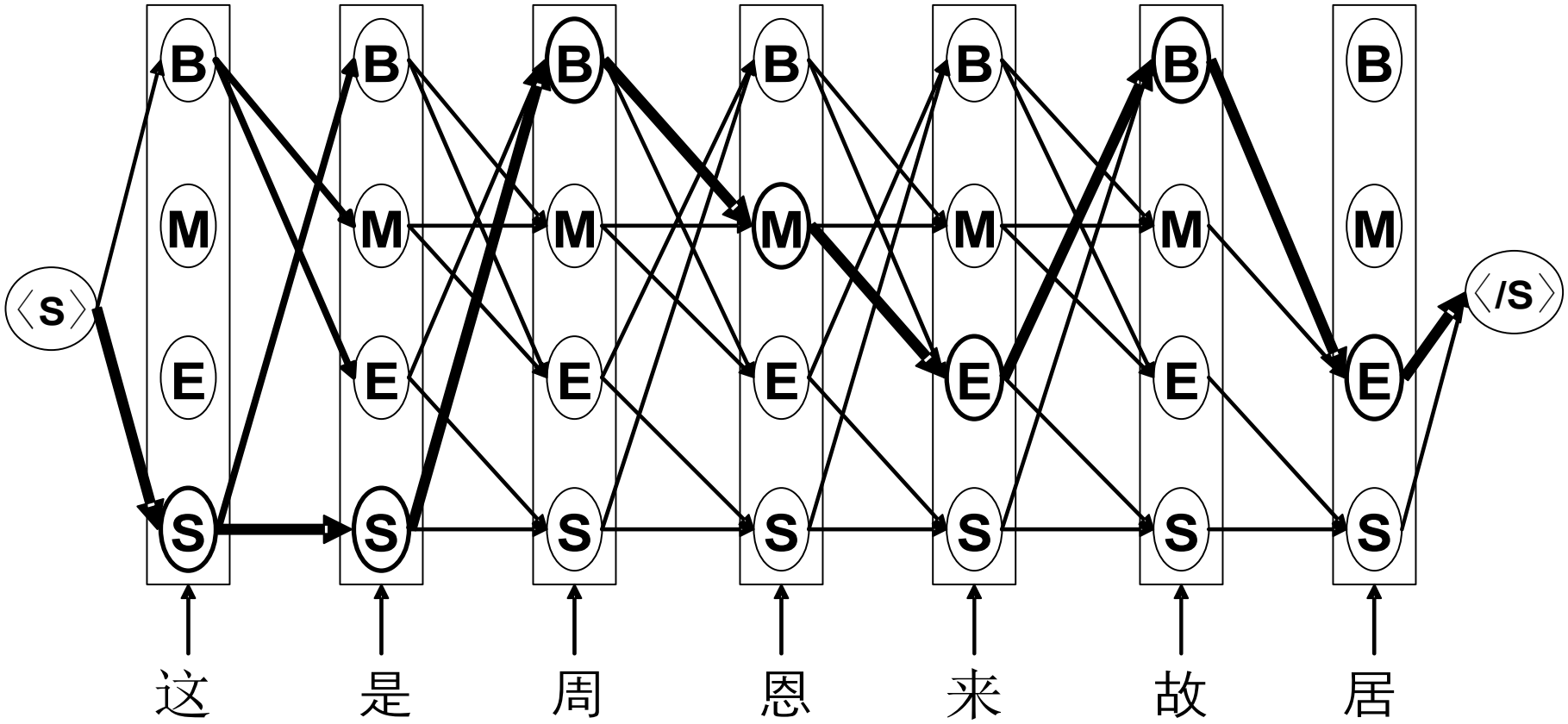


搜索最优标注路径

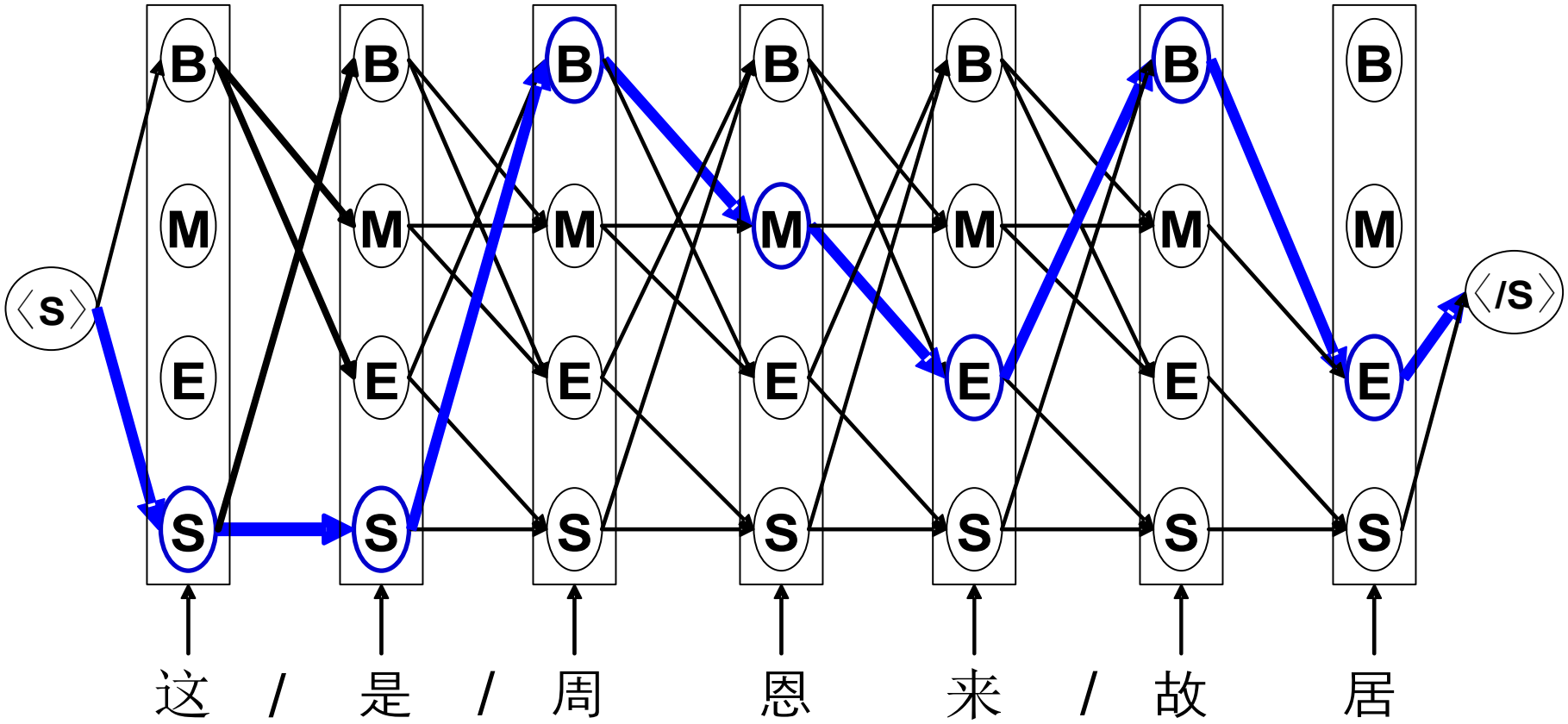
- 有些边是非法的，比如 M→B， E→M， E→E，S→M， S→E，等等

  注意：这里的非法边和未定义词识别的 BMEO 标记转移图的非法边略有不同，想一想为什么
- 每条路径的概率是路径上各节点标注概率 P(标记 | 汉字) 之积，边上没有转移概率
- 搜索算法与 HMM 模型解码的 Viterbi 算法相同

搜索最优标注路径



搜索最优标注路径



NUS 的工作

- Low, Jin Kiat, & Ng, Hwee Tou, & Guo, Wenyuan (2005). A Maximum Entropy Approach to Chinese Word Segmentation. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.* (pp. 161-164). Jeju Island, Korea

NUS 在 SIGHAN05 上的工作

- 最大熵方法
- 基本特征（每个特征都省略了当前字标记）：
  - $C_n(n = -2, -1, 0, 1, 2)$：汉字
  - $C_nC_{n+1}(n = -2, -1, 0, 1)$：两字组
  - $C_{-1}C_1$：当前字左右两个字
  - $P_u(C_0)$：当前字标点符号为 1
  - $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$：字符类别
    - 数字
    - 日期（“年”、“月”、“日”）
    - 英文字母
    - 其他字符

NUS 在 SIGHAN05 上的工作

- 词典特征（每个特征都省略了当前字标记）：
  - Lt_0
  - $C_n t_0$ $(n = -1, 0, 1)$
- 说明：
  - 引入一部外部词典，查找当前字周围最大匹配词 W
  - t_0 为当前字在 W 中的位置标记（ BMES ）
  - L 为 W 的长度（字数）
  - 上述特征为分别为以上符号的组合
- 例子：

  “ 新华社北京……”，当前字为“华”，词典中查到最大匹配的词为“新华社”，于是 L 为 3， t_0 为 M，得到特征为：

  Lt_0=3M　$C_{-1}t_0$= 新 M　$C_{-1}t_0$= 华 M　$C_{-1}t_0$= 社 M

NUS 在 SIGHAN05 上的工作

- 语料的扩充：
  - 基本思想：
    - SIGHAN 评测提供不同切分标注标准的多套训练语料和测试语料，作为不同的评测任务
    - 利用其他标准的训练语料来扩充当前标准的训练语料
  - 具体做法：
    - 用当前标准语料训练好的标注工具标注其他标注的已标注语料
    - 取那些与人工标注一致的语料来扩充当前标准的训练语料

NUS 在 SIGHAN05 上的工作

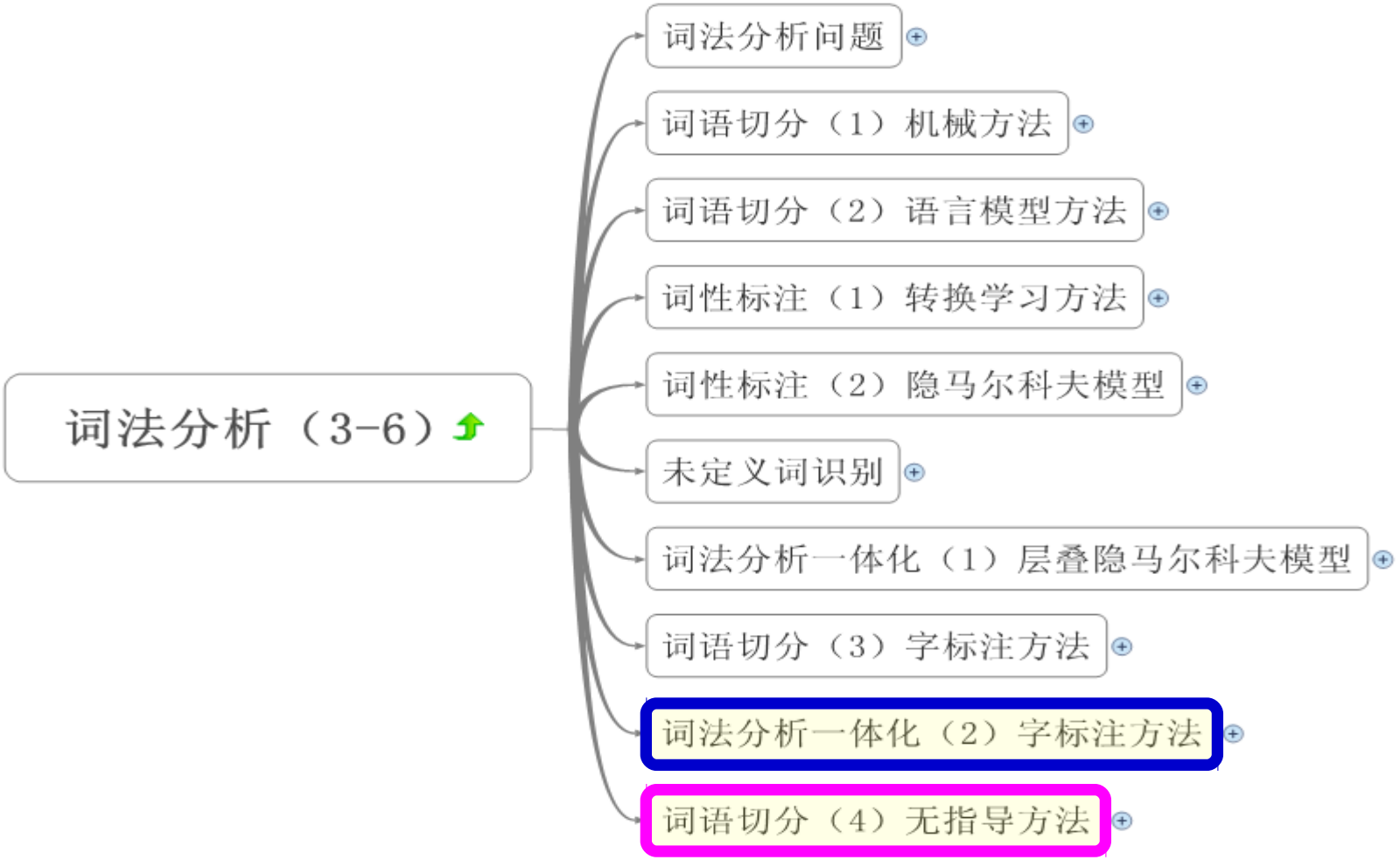
Corpus	R	P	F	R_{OOV}	R_{IV}
AS	0.962	0.950	0.956	0.684	0.975
CITYU	0.967	0.956	0.962	0.806	0.980
MSR	0.969	0.968	0.968	0.736	0.975
PKU	0.968	0.969	0.969	0.838	0.976

- NUS 参加了所有四个开放训练项目评测，由于使用了训练数据以外的词典和语料库，没有参加封闭训练项目的评测
- NUS 在 AS 、 CITYU 、 PKU 三个项目中均获第一名，在 MSR 项目中获第二

基于字标注方法的特点

- 模型简单：单一模型解决所有问题：词语切分、未定义词识别，甚至词性标注也可以用这个模型解决
- 符合直觉：用字直接组词，符合对汉语的直观理解
- 模型功能强大：可以任意选择特征、可以调节特征直接的权重
- 性能高：对 OOV 识别能力比生成式模型大大提高，对 IV 识别性能比生成式模型稍差，总体 F 值高于生成式模型

内容提要


词法分析（3-6）
词法分析问题
词语切分（1）机械方法
词语切分（2）语言模型方法
词性标注（1）转换学习方法
词性标注（2）隐马尔科夫模型
未定义词识别
词法分析一体化（1）层叠隐马尔科夫模型
词语切分（3）字标注方法
词法分析一体化（2）字标注方法
词语切分（4）无指导方法

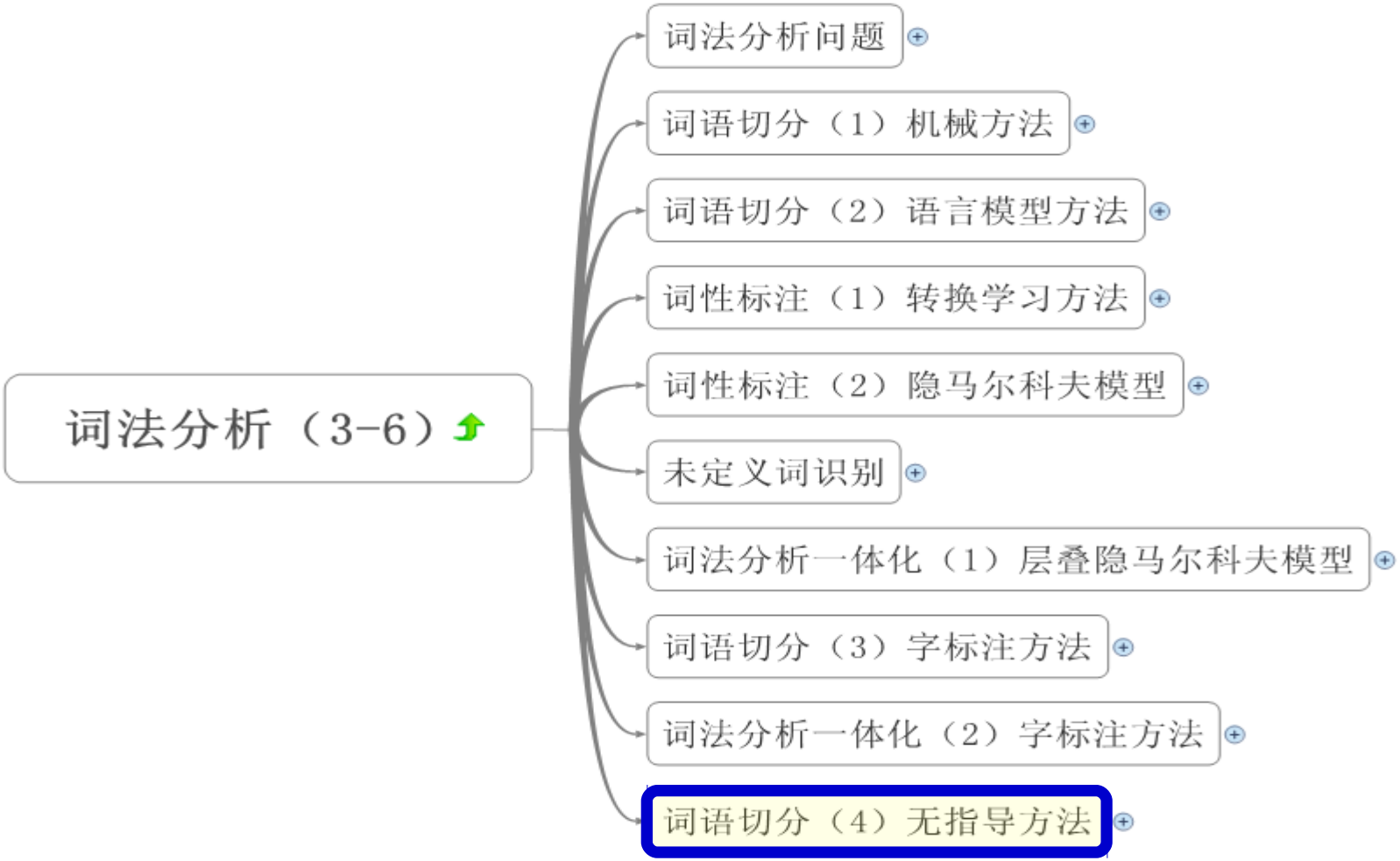
NUS 的工作

- Ng, Hwee Tou & Low, Jin Kiat (2004). Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*. (pp. 277-284). Barcelona, Spain.

基于字标注方法的进一步改进

- 融入词性标注：
  - 扩充标记集
  - 为每一个词性定义 BMES 四个标记
- 更强大的序列标注模型：
  - 最大熵模型：不求解全局最优
  - 最大熵马尔科夫模型：求解全局最优
  - 条件随机场模型：解决标记偏置问题
  - 感知机模型：性能与效率的平衡

内容提要


词法分析（3-6）
词法分析问题
词语切分（1）机械方法
词语切分（2）语言模型方法
词性标注（1）转换学习方法
词性标注（2）隐马尔科夫模型
未定义词识别
词法分析一体化（1）层叠隐马尔科夫模型
词语切分（3）字标注方法
词法分析一体化（2）字标注方法
词语切分（4）无指导方法

无指导的汉语词语切分

- 基于汉字连接强度的方法
  - Sproat R., Shih C.L., A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Languages, 1993,4(4):336-249
  - 孙茂松，肖明，邹嘉彦，基于无指导学习策略的无词表条件下的汉语自动分词，计算机学报，27(6):736-742, 2004.6
- 基于最小描述长度的方法
  - W.J.Teahan, Yingying Wen, Rodger McNab, Ian H. Witten, A compression-based algorithm for Chinese word segmentation, *computational linguistics*, 26(3):375-393, 2000

基于互信息的汉语词语切分

- 训练：
  - 计算任意两个汉字之间的互信息

$$I(X,Y)=-\log_2\frac{p(X)p(Y)}{p(X,Y)}$$

- 切分：对于给定的汉字串 $c_1c_2\ldots c_n$
  - 循环
    - 寻找相邻的两个汉字，使其互信息大于给定的阈值
    - 如果没有找到这样两个相邻的汉子，则返回
    - 将这两个汉字组合成一个词

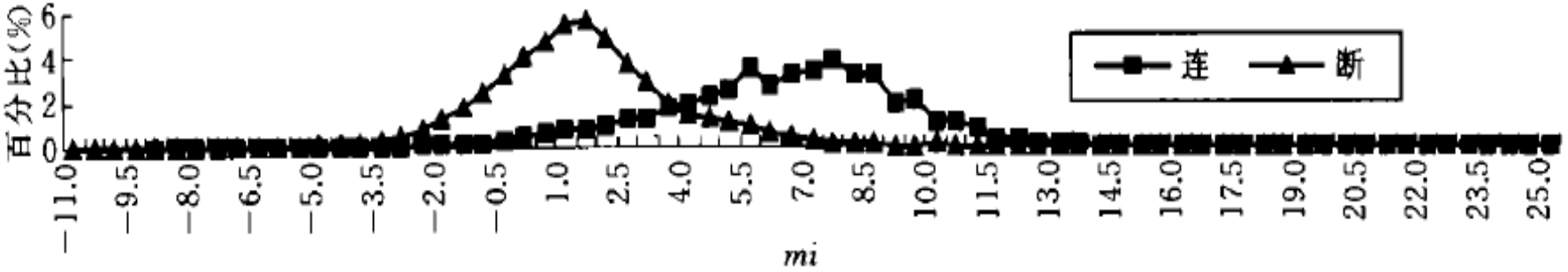
基于互信息的汉语词语切分

	我	弟	弟	現	在	要	坐	火	車	回	家
MI	0	10.4	0	4.2	-2.8	0	0	7.3	2.1	4.7	
1	我	**弟**	**弟**	現	在	要	坐	火	車	回	家
2	我	[弟	弟]	現	在	要	坐	**火**	**車**	回	家
3	我	[弟	弟]	現	在	要	坐	[火	車]	**回**	**家**
4	我	[弟	弟]	**現**	**在**	要	坐	[火	車]	[回	家]
5	我	[弟	弟]	[現	在]	要	坐	[火	車]	[回	家]

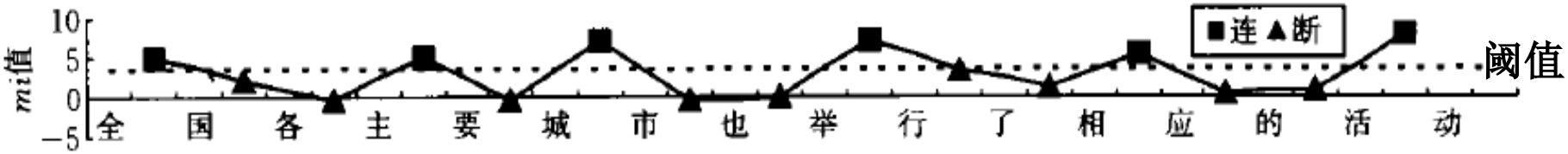
我　弟弟　现在　要　坐　火车　回家

(Sproat & Shin, 1993)

基于互信息的汉语词语切分



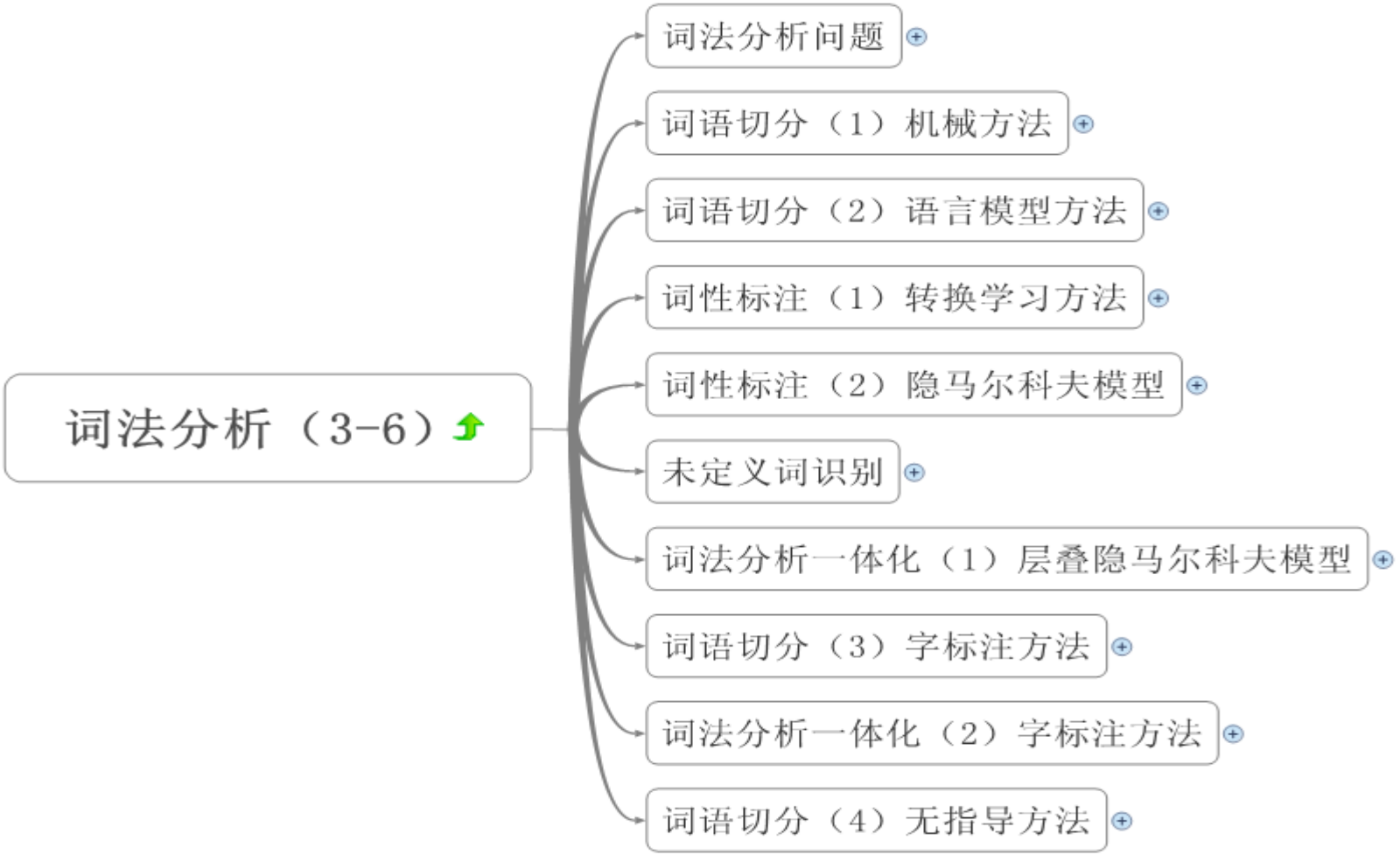
互信息关于“连”、“断”位置的分布



例子：全国　各　主要　城市　也　举行　了　相应　的　活动

（孙茂松等，2004）

词法分析小结


词法分析（3-6）
词法分析问题
词语切分（1）机械方法
词语切分（2）语言模型方法
词性标注（1）转换学习方法
词性标注（2）隐马尔科夫模型
未定义词识别
词法分析一体化（1）层叠隐马尔科夫模型
词语切分（3）字标注方法
词法分析一体化（2）字标注方法
词语切分（4）无指导方法