条件随机场理论综述

韩雪冬[1]，周彩根[2]
[1]北京邮电大学计算机学院 ，北京（100876）
[2]中国人民解放军总参谋部第五十四研究所，北京（100083）
E-mail: hanxuedong@bupt.cn

摘 要：条件随机场理论可以用于序列标记、数据分割、组块分析等自然语言处理任务。在中文分词、中文人名识别、歧义消解等汉语自然语言处理任务中都有应用，表现很好。与一般介绍条件随机场理论的论文有所不同，本文给出了条件随机场理论的概率模型的推导，参数估计的函数形式为对数似然函数的原因及条件随机场矩阵计算的图例说明，能使读者掌握条件随机场理论的依据和整体。

关键词：条件随机场；CRFs；最大熵模型；中文分词
中图分类号：TP301.6

1 引言

条件随机场理论（CRFs）可以用于序列标记、数据分割、组块分析等自然语言处理任务中。在中文分词、中文人名识别、歧义消解等汉语自然语言处理任务中都有应用，表现很好。目前基于 CRFs 的主要系统实现有 CRF，FlexCRF，CRF++[1]。

本文主要介绍条件随机场理论。因为条件随机场理论与它先前的基于统计方法的模型有着联系，所以先是介绍了隐马尔可夫模型，而后介绍了最大熵模型，给出了概率模型的推导过程和其参数估计函数形式。最后重点介绍了条件随机场模型。

2 隐马尔可夫模型

隐马尔可夫模型(Hidden Markov Models，HMMs)研究始于 1966，基于统计方法的隐马尔可夫模型在若干年后变得很受欢迎，原因有二个，一是该模型有丰富的数学理论结构，能被广泛的应用；二是在若干重要应用上经恰当的运用表现的很出色。在讲述隐马尔可夫模型之前，我们先简单介绍以下几个模型用到的马尔可夫随机过程。

2.1 离散马尔可夫过程

设有 N 个不同状态 $\{S_1, S_2, \cdots, S_n\}$ 的随机过程，令 t=1,2,…表示不同的时间点，q_t 表示 t 时刻随机过程所处的状态，a_{ij} 表示状态 S_i 到 S_j 的转移概率。当随机过程满足：当前所处的状态仅与它之前的一个状态有关，即

$$P[q_t = S_i \mid q_{t-1} = S_j, q_{t-2} = S_k, \cdots] = P[q_t = S_i \mid q_{t-1} = S_j] \qquad (1)$$

时，该随机过程为马尔可夫随机过程。而且我们考虑（1）式右边的随机过程是独立于时间的，从而得到状态间的转移概率 a_{ij}，

$$a_{ij} = P[q_t = S_i \mid q_{t-1} = S_j], \qquad 1 \le i, j \le N \qquad (2)$$

转移概率 a_{ij} 具有两个属性：$a_{ij} \ge 0$ 和 $\sum_{j=1}^{N} a_{ij} = 1$ ，因此 a_{ij} 服从概率约束。

以上介绍了离散马尔可夫随机过程，下面我们先介绍隐马尔可夫模型的要素，而后介绍隐马尔可夫模型面临的三个基本问题及解决方法。

2.2 隐马尔可夫要素

隐马尔可夫模型有五个要素[2]组成：

1）N，表示模型中的状态数。模型中的各个状态是相互连结的，任何状态能从其它状态到达。我们用 S 表示各个状态的集合，$S=\{s_1,s_2,\cdots,s_N\}$，q_t 表示 t 时刻的状态。

2）M，表示模型中每个状态不同的观察符号，即输出字符的个数。我们用 V 表示各个字符的集合，$V=\{v_1,v_2,\cdots,v_M\}$。

3）A，状态转移概率分布。$A=\{a_{ij}\}$，其中，$a_{ij}=P[q_t=S_i\,|\,q_{t-1}=S_j]$，$1\le i,j\le N$，当从状态 S_i 经一步到达 S_j 时，$a_{ij}>0$，否则 $a_{ij}=0$。

4）B，观察字符在状态 j 时的概率分布，$B=\{b_j(k)\}$，其中 $b_j(k)=P[v_k\,|\,q_t=S_j]$，$1\le j\le N$，$1\le k\le M$。

5）π，表示初始状态分布，$\pi=\{\pi_j\}$，其中 $\pi_j=P[q_1=S_j]$，$1\le j\le N$。

给定 N,M,A,B,π，HMMs 能输出一个观察字符的序列 $\mathrm{O}=\mathrm{O}_1\mathrm{O}_2\cdots\mathrm{O}_T$，其中 $\mathrm{O}_t\in V$，T 是观察序列的字符个数。

从以上的讨论可知，一个完整的隐马尔可夫模型要求两个具体的模型参数 N 和 M，和三个概率矩阵 A,B,π，也即隐马尔可夫模型可形式化定义为一个五元组 (N,M,A,B,π)。

以上介绍了隐马尔可夫模型的五个要素，下面我们介绍隐马尔可夫模型的三个基本问题及相应的解决方法。

2.3 隐马尔可夫模型的三个基本问题

1）给定一个模型 $\lambda=(N,M,A,B,\pi)$，如何高效的计算某一输出字符序列 $\mathrm{O}=\mathrm{O}_1\mathrm{O}_2\cdots\mathrm{O}_T$ 的概率 $P(\mathrm{O}\,|\,\lambda)$。

2）给定一个模型 $\lambda=(N,M,A,B,\pi)$和一个输出字符序列 $\mathrm{O}=\mathrm{O}_1\mathrm{O}_2\cdots\mathrm{O}_T$，如何找到产生这一序列概率最大的状态序列 $Q=s_is_j\cdots s_T$。

3）给定一个模型 $\lambda=(N,M,A,B,\pi)$和一个输出字符序列 $\mathrm{O}=\mathrm{O}_1\mathrm{O}_2\cdots\mathrm{O}_T$，如何调整模型的参数使得产生这一序列的概率最大。

为了方便分析问题和给出解决方案，这里先介绍一下隐马尔可夫模型的条件独立性假设。隐马可尔可夫模型是一个生成模型，给定一个观察序列，HMMs 模型隐含一个与观察序列对应的状态序列。隐马可夫模型图示如下，图中清楚的表示出了隐马尔可夫模型内部的条件独立关系，有三个独立性假设：一是 t 时刻的状态 $q_t=s_i$ 只依赖于 $t-1$ 时刻的状态 $q_{t-1}=s_j$，即 $P(q_t\,|\,q_{t-1}\cdots q_1,\lambda)=P(q_t\,|\,q_{t-1},\lambda)$。二是 t 时刻所生成的值 $b_i(\mathrm{O}_t)$ 只依赖于 t 时刻的状态 $q_t=s_i$，即 $P(\mathrm{O}_1\mathrm{O}_2\cdots\mathrm{O}_T\,|\,q_1q_2\cdots q_T,\lambda)=\prod_{t=1}^{T}P(\mathrm{O}_t\,|\,q_t)$。三是状态与具体时间无关，即对任意的 i 和 j 都有 $P(q_i\,|\,q_{i-1},\lambda)=P(q_j\,|\,q_{j-1},\lambda)$。

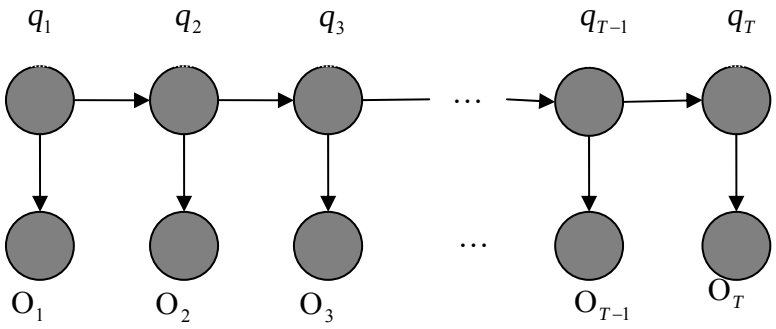

图 1 HMMs

下面我们给出三个基本问题的解决方法。

2.4 forward-backward 算法

问题 1 是一个评价问题，即给定一个模型 λ 和一个观察序列 $O=O_1O_2\cdots O_T$，如何计算由模型产生这一观察序列的概率 $P(O|\lambda)$。最直接的方法是枚举所有长度为 T，输出观察序列为 O 的可能的状态序列。假设状态数为 N，时枚举方法的计算量为 $2T\cdot N^T$，使该方法的在计算上不可行。目前可采用 forward-backward 算法解决这个问题。

forward-backward 过程[3][4]：定义 forward 变量 $\alpha_t(i)$ 为

$$\alpha_t(i)=P(O_1O_2\cdots O_t,q_t=s_i|\lambda) \qquad (3)$$

即对于模型 λ，在 t 时刻，状态为 S_i 时的部分观察序列 $O_1O_2\cdots O_t$ 的概率记为 $\alpha_t(i)$，$\alpha_t(i)$ 为部分观察序列 $O_1O_2\cdots O_t$ 和 t 时刻的状态 S_i 的联合分布概率，则 $\alpha_t(i)$ 可递归得到。由各个观察字符的输出状态相互独立，下面给出 $\alpha_j(t+1)$ 的递推过程：

$$\begin{aligned}
&\alpha_j(t+1)=P(O_1O_2\cdots O_{t+1},q_{t+1}=s_j|\lambda)\\
&=P(O_1O_2\cdots O_{t+1}|q_{t+1}=s_j,\lambda)P(q_{t+1}=s_j|\lambda)\\
&=P(O_1O_2\cdots O_t|q_{t+1}=s_j,\lambda)P(O_{t+1}|q_{t+1}=s_j,\lambda)P(q_{t+1}=s_j|\lambda)\\
&=\sum_{i=1}^{N}P(O_1O_2\cdots O_t,q_t=s_i,q_{t+1}=s_j|\lambda)P(O_{t+1}|q_{t+1}=s_j,\lambda)\\
&=\sum_{i=1}^{N}P(O_1O_2\cdots O_t,q_{t+1}=s_j|q_t=s_i,\lambda)P(q_t=s_i|\lambda)P(O_{t+1}|q_{t+1}=s_j,\lambda)\\
&=\sum_{i=1}^{N}P(O_1O_2\cdots O_t|q_t=s_i,\lambda)P(q_{t+1}=s_j|q_t=s_i,\lambda)P(q_t=s_i|\lambda)P(O_{t+1}|q_{t+1}=s_j,\lambda)\\
&=\sum_{i=1}^{N}P(O_1O_2\cdots O_t,q_t=s_i|\lambda)P(q_{t+1}=s_j|q_t=s_i,\lambda)P(O_{t+1}|q_{t+1}=s_j,\lambda)\\
&=\left[\sum_{i=1}^{N}\alpha_i(t)a_{ij}\right]b_j(O_{t+1})
\end{aligned}$$

则观察序列所有可能的状态序列的概率为，

$$P(\mathrm{O}\mid\lambda)=\sum_{i=1}^{N}\alpha_T(i). \tag{4}$$

下图说明了在t时刻从N个状态S_i，$1\le i\le N$到达$t+1$时刻的状态S_j的forward过程，

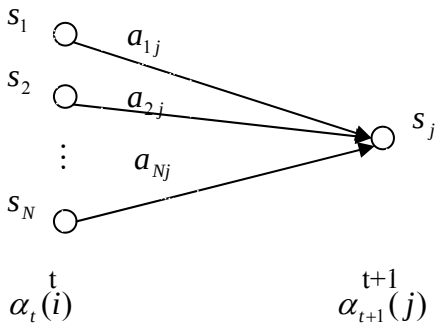

图 2 forward 计算

由以上可知$\alpha_t(i)$是观察序列$\mathrm{O}_1\mathrm{O}_2\cdots\mathrm{O}_t$和$t$时刻所处的状态$s_i$的联合概率，$\alpha_t(i)a_{ij}$是观察序列$\mathrm{O}_1\mathrm{O}_2\cdots\mathrm{O}_t$的在$t$时刻的输出状态序列和在$t+1$时刻经$s_i$到达$s_j$的联合概率，从$t$时刻所有$N$个可能的状态$s_i$，$1\le i\le N$到达$t+1$的$s_j$状态的概率和，而后乘以$t+1$时刻观察字符$\mathrm{O}_{t+1}$在状态$s_j$的概率$b_j(\mathrm{O}_{t+1})$为$t+1$时刻在$s_j$状态的概率，即得到$\alpha_{t+1}(j)$，$1\le j\le N$,$t=1,2,\cdots,T-1$。最终的 forward 变量$\alpha_T(i)$为，

$$\alpha_T(i)=P(\mathrm{O}_1\mathrm{O}_2\cdots\mathrm{O}_T,q_T=s_i\mid\lambda) \tag{5}$$

因此$P(\mathrm{O}\mid\lambda)$等于各个$\alpha_T(i)$的和，其中$1\le i\le N$。forward 算法计算$P(\mathrm{O}\mid\lambda)$需要计算$\alpha_t(j)$，而$1\le j\le N$，$1\le t\le T$，所以总计算量为N^2T，远小于直接计算$P(\mathrm{O}\mid\lambda)$所用的$2T\cdot N^T$的计算量。

与定义 forward 变量类似，我们可以定义 backward 变量$\beta_t(i)$为

$$\beta_t(i)=P(\mathrm{O}_{t+1}\mathrm{O}_{t+2}\cdots\mathrm{O}_T\mid q_t=s_i,\lambda) \tag{6}$$

即t时刻，部分观察序列从O_{t+1}到O_T,给定模型λ和状态s_i条件下的概率，由 HMMs 的特性，我们可以递推$\beta_t(i)$：

$$\begin{aligned}
&\beta_t(i)=P(\mathrm{O}_{t+1}\mathrm{O}_{t+2}\cdots\mathrm{O}_T\mid q_t=s_i,\lambda)\\
&=\sum_{j=1}^{N}P(\mathrm{O}_{t+1}\mathrm{O}_{t+2}\cdots\mathrm{O}_T,q_{t+1}=s_j\mid q_t=s_i,\lambda)\\
&=\sum_{j=1}^{N}P(\mathrm{O}_{t+1}\mathrm{O}_{t+2}\cdots\mathrm{O}_T\mid q_{t+1}=s_j,\lambda)P(q_{t+1}=s_j\mid q_t=s_i,\lambda)\\
&=\sum_{j=1}^{N}P(\mathrm{O}_{t+2}\cdots\mathrm{O}_T\mid q_{t+1}=s_j,\lambda)P\big(\mathrm{O}_{t+1}\mid q_{t+1}=s_j,\lambda\big)P(q_{t+1}=s_j\mid q_t=s_i,\lambda)\\
&=\sum_{j=1}^{N}a_{ij}b_j(\mathrm{O}_{t+1})\beta_{t+1}(j)
\end{aligned}$$

给定观察序列O和模型λ条件下，t时刻状态为$q_t = s_i$时的状态序列的概率定义为$P(q_t = s_i \mid \mathrm{O}, \lambda)$。由$\alpha_t(i)$，$\beta_t(i)$可知观察序列和$t$时刻状态为$q_t = s_i$时的联合概率$P(q_t = s_i, \mathrm{O} \mid \lambda)$和观察序列的概率$P(\mathrm{O} \mid \lambda)$分别为：

$$P(q_t = s_i, \mathrm{O} \mid \lambda) = \alpha_t(i)\beta_t(i) \tag{7}$$

$$P(\mathrm{O} \mid \lambda) = \sum_{i=1}^{N} P(q_t = s_i, \mathrm{O} \mid \lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \tag{8}$$

由上面两式可知：

$$P(q_t = s_i \mid \mathrm{O}, \lambda) = \frac{P(q_t = s_i, \mathrm{O} \mid \lambda)}{P(\mathrm{O} \mid \lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{9}$$

图示如下：

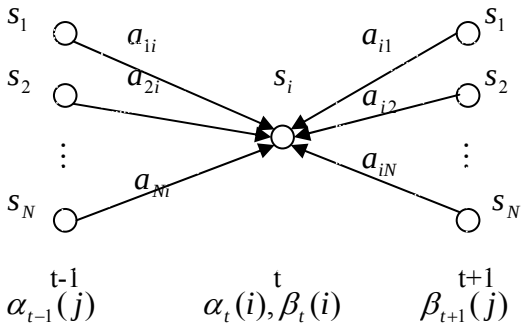

图 3 节点概率计算

2.5 Viterbi 算法

问题 2 是一个解码问题，即从N^T个可能的状态序列中找到一个“最优”的状态序列，其中N是 HMMs 模型中状态的个数，T是观察序列的长度。不像问题 1 能给出一个确定的解决方案，对于问题 2 根据“最优”的标准不同，可以有若干个解决方案，所以给定观察序列，找出“最优”状态序列的困难是最优状态序列的定义，即最优标准的选择。例如一个最优标准是去选择在t时刻，状态为q_t的单个概率为最大，这个最优标准是使状态序列中正确的单个状态的数学期望值最大。但最广泛应用的标准是考虑使整个状态序列最优，也即是最优路径问题，这种方法试图找到最大的$P(Q \mid \mathrm{O}, \lambda)$，由于$P(\mathrm{O} \mid \lambda)$的概率对找到最大的$P(Q \mid \mathrm{O}, \lambda)$没有影响，实际上等于找到最大的$P(Q, \mathrm{O} \mid \lambda)$。一个有效的查找最优路径的算法是 Viterbi 算法，它基于动态规划方法。

Viterbi 算法[5][6]：给定观察序列$\mathrm{O} = \mathrm{O}_1\mathrm{O}_2 \cdots \mathrm{O}_T$，利用 Viterbi 算法可以有效率的找到一个最优的状态序列$Q = q_1 q_2 \cdots q_T$，计算量为NT^2，我们定义$\delta_t(i)$表示t时刻状态$q_t = s_i$时的最优状态序列和前t个观察序列的联合概率：

$$\delta_t(i) = \max_{q_1, q_2, \cdots q_{t-1}} P(q_1 q_2 \cdots q_t = s_i, \mathrm{O}_1\mathrm{O}_2 \cdots \mathrm{O}_t \mid \lambda) \tag{10}$$

由t时刻状态$q_t = s_i$时的最优状态序列和前t个观察序列的联合概率$\delta_t(i)$可递推得到$t+1$时刻状态$q_{t+1} = s_j$时的最优状态序列和前$t+1$个观察序列的联合概率$\delta_{t+1}(j)$：

$$\delta_{t+1}(j)=[\max_{1\le i\le N}\delta_t(i)a_{ij}]\cdot b_j(\mathrm{O}_{t+1}). \tag{11}$$

在实际应用中，为了由当前最优路径中的最后一个状态检索出最优状态序列，我们需要用数组 $\varphi_t(j)$ 保存 t 时刻的各个状态处于最优路径时的前一个状态索引。发现最优状态序列的完整过程如下：

1）初始化， $\delta_1(i)=\pi_i b_i(\mathrm{O}_1)$ ， $\varphi_1(i)=0$ ， $1\le i\le N$ ；

2)递归得到：

$$\delta_t(j)=[\max_{1\le i\le N}\delta_{t-1}(i)a_{ij}]\cdot b_j(\mathrm{O}_t),\qquad 2\le t\le T,1\le j\le N \tag{12a}$$

$$\varphi_t(j)=\arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}],\qquad 2\le t\le T,1\le j\le N \tag{12b}$$

3)最终完整状态序列的最大概率 P^* 和最大概率的状态序列的最后一个状态记为 q_T^*，则

$$P^*=\max_{1\le i\le N}[\delta_T(i)] \tag{13a}$$

$$q_T^*=\arg\max_{1\le i\le N}[\delta_T(i)] \tag{13b}$$

4）让 q_t^* 表示最优状态序列 t 时刻的状态索引，则最优路径或状态序列的逆向索引为：

$$q_t^*=\varphi_{t+1}(q_{t+1}^*),\quad t=T-1,T-2,\cdots,1. \tag{14}$$

除了逆向找出状态索引的步骤不同，Viterbi 算法与 forward 计算过程很相似，其最主要区别是 Viterbi 算法是在根据先前的状态序列找出转到当前状态有最大概率的那个状态，而 forward 算法是根据先前的状态序列计算转移到当前状态的概率和。它们都可以采用格子(Lattice or trellis)的结构形式实现有效的计算。

2.6 参数估计

问题 3 是模型参数估计问题。给定观察序列 O，调整模型的参数使得在给定模型 λ 的条件下该观察序列的概率 $P(\mathrm{O}\mid\lambda)$ 最大。无法用解析方法求解，事实上，给定任意有限的观察序列作为训练数据，不存在一个最优的方法去估计模型参数，然而我们可以用 Baum-Welch 方法（等价于 EM(Expectation-Modification)方法）或者用梯度技术，通过不断循环迭代更新参数的方法，设法使 $P(\mathrm{O}\mid\lambda)$ 达到最优。Baum-Welch 方法是 EM 算法的一种实现，因采用爬山法往往得到的是局部最优。

2.7 隐马尔可夫模型的局限性

应用 HMMs 模型，在序列标记任务中，我们的目标是找到一个给定观察序列 $\mathrm{O}=\mathrm{O}_1\mathrm{O}_2\cdots\mathrm{O}_T$ 的条件下，使标记序列 $Q=q_1q_2\cdots q_T$ 的条件概率最大的那个标记序列 $Q_{\max}$，即 $Q_{\max}=\arg\max_{allQ}P(Q\mid\mathrm{O})$ 。隐马尔可夫模型定义的是观察序列和状态序列的联合概率 $P(\mathrm{O},Q)$，由贝叶斯公式：

$$P(\mathrm{O}\mid Q)P(Q)=P(\mathrm{O},Q)=P(Q\mid\mathrm{O})P(\mathrm{O}) \tag{15}$$

可知 $Q_{\max}=\arg\max_{allQ}\dfrac{P(\mathrm{O},Q)}{P(\mathrm{O})}$，可以看出生成模型定义的是观察序列和标记序列的联合概率分布 $P(\mathrm{O},Q)$。但在标记数据时模型关心的是在给定观察序列 O 的条件下，标记序列 Q 的

条件分布 $P(Q|O)$。定义观察序列和标记序列的联合分布意味着所有可能的观察序列必须是可枚举的，如果观察序列中存在长距离依赖，枚举所有可能的观察序列是十分困难的，因此，为了便于模型处理问题，生成模型给出了一个严格的输出独立性假设。例如在隐马尔可夫模型中，我们假设 t 时刻的观察值只依赖于 t 时刻的状态，这确保了序列中的所有观察值互相独立。但事实上，数据序列并不能完全地表示为一组独立的单元。当序列中的数据元素存在长距离依赖时，允许这种长距离依赖并且使观察序列可以表示为非独立的交叉特征的模型才是比较合适的。

下面所提到的条件模型或判别模型克服了生成模型所要求的严格的独立假设，它定义了一个在给定观察序列 O 的条件下，状态序列 Q 的条件分布 $P(Q|O)$。下一小节我们将介绍最大熵模型，主要是其条件概率模型的推导和参数估计的函数形式。

3 最大熵模型

最大熵模型（Maximum Entropy Models, MEMs）[7][8][9]是基于最大熵理论的统计模型，广泛应用于模式识别和统计评估中。最大熵原理有一个很长的历史，其中最大熵理论方面的先驱 E.T.Jaynes 在 1990 年给出了最大熵原理的基本属性[10]：最大熵概率分布服从我们已知的不完整信息的约束。主要思想是，在用有限知识预测未知时，不做任何有偏的假设。根据熵的定义，一个随机变量的不确定性是由熵体现的，熵最大时随机变量最不确定，对其行为做准确预测最困难。最大熵原理的实质是，在已知部分知识前提下，关于未知分布最合理的推断是符合已知知识的最不确定或最随机的推断，这是我们可以做出的唯一不偏不倚的选择。最大熵的原理可以概括为，将已知事件作为约束条件，求得可使熵最大化的概率分布作为正确的概率分布。熵的计算公式如下[11]：

$$H(X) \equiv -\sum_{x \in X} p(x) \log p(x) \tag{16}$$

熵有如下的性质：

$$0 \le H(X) \le \log |X| \tag{17}$$

其中 $|X|$ 在离散分布时是随机变量的个数。当 X 为确定值，即没有变化的可能时上式左边的等式成立。由条件：$\sum_{x \in X} p(x) = 1$，对熵的计算公式（16）求条件极值，可知当随机变量 X 服从均匀分布时，$H(X) = \log |X|$ 成立，即均匀分布时熵最大。

最大熵模型用到熵的计算公式是有条件约束的，如 $p(x_1) + p(x_2) = 0.3$ 条件约束下的最大熵，或者更多约束条件时的最大熵。我们的目的是找到一个能同时满足这些约束条件的最均匀的模型。由此最大熵模型面临两个问题，一是如何确定模型是均匀的，二是根据一个约束集如何找到一个最优的均匀分布。由上面熵取得最大值时分布可知，当熵模型在满足约束条件下取得最大值时，熵模型是均匀的。下面我们介绍这二个问题的解决过程。

自然语言处理中的很多问题可以归结为统计分类问题，因此可以将自然语言处理任务的所有输出值构成一个类别有限集 Y，对于每个 $y \in Y$，其生成均受上下文信息 x 的影响和约束。已知与 y 有关的所有上下文信息组成的集合为 X，则模型的目标是，在给定上下文信息 $x \in X$，计算输出为 $y \in Y$ 的条件概率 $p(y|x)$，模型的输入为人工标注后训练数据样本集 $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_\Gamma, y_\Gamma)\}$。我们可以从训练样本归结出随机变量 x 和 y 的联合经验概率分布 $\tilde{p}(x, y)$，

$$\tilde{p}(x,y)=\frac{(x,y)在样本中同时出现的次数}{\Gamma} \tag{18}$$

其中Γ为整个样本空间D的大小。特殊情况，样本空间中(x,y)根本没有同时出现，或者在一些上下文中出现了多次。

3.1 最大熵模型的约束条件

我们的目标是构造一个能生成训练样本分布$\tilde{p}(x,y)$的统计模型，建立特征方程。该特征必须能较完整地表达训练样本中数据的特性。例如，在中文分词任务中，可以引入特征函数$f(x,y)$，

$$f(x,y)=\begin{cases}1 & if \quad y=\sin gle \quad and \quad x=',' \\ 0 & otherwise\end{cases}$$

设$\tilde{p}(f)$是相对于经验分布$\tilde{p}(x,y)$，特征函数f的数学期望，称为经验期望，公式为：

$$\tilde{p}(f)=\sum_{x,y}\tilde{p}(x,y)f(x,y) \tag{19}$$

$p(f)$是相对于由模型确定的概率分布$p(x,y)$的数学期望，称为模型期望，公式如下：

$$p(f)=\sum_{x,y}p(x,y)f(x,y)=\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y) \tag{20}$$

其中$\tilde{p}(x)$是随机变量x在训练样本中的经验分布，即在样本中出现的频率。我们约束由模型得到的特征函数的数学期望等于由训练样本得到的特征函数的经验数学期望，即：

$$p(f)=\tilde{p}(f) \tag{21}$$

由上面的三个式子（19）（20）（21）可以得到下面的等式：

$$\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y)=\sum_{x,y}\tilde{p}(x,y)f(x,y) \tag{22}$$

我们把等式（21）称为模型的约束等式，或者简单的称为约束。

我们现在可以用$\tilde{p}(f)$描述训练样本的统计现象的内在属性，同时也要求由模型得到的$p(f)$能完整的展现这些统计现象的内在属性，即$p(f)=\tilde{p}(f)$。

3.2 最大熵模型的原则

下面介绍最大熵模型的原则，先定义P为所有条件分布的集合，根据$p(y|x)$的定义，有$p(y|x)\in \mathrm{P}$。假设我们选择m个对模型真正有用的特征函数f_i，用以体现统计数据的特性。约束条件下所产生的集合C是P的一个子集，即$C\subset \mathrm{P}$，定义如下：

$$C=\left\{p\in \mathrm{P}\mid p(f_i)=\tilde{p}(f_i),\quad i\in\{1,2,\cdots,m\}\right\} \tag{23}$$

满足约束条件的模型很多。模型的目标是产生在约束集下具有最均匀分布的模型，条件熵$H(Y|X)$是作为条件概率$p(y|x)$均匀性的一种数学测度方法。为了强调熵对概率分布p的依赖，我们用$H(p)$代替$H(Y|X)$，得条件分布的熵公式如下：

$$H(p)\equiv-\sum\tilde{p}(x)p(y|x)\log p(y|x) \tag{24}$$

对于任意给定的约束集C，能找到唯一的$p*\in C$使$H(p)$取得最大值，如何找到p^*，是一个约束最优化问题。我们给出p^*的等式，

$$p^* = \arg\max_{p \in C} H(p) \tag{25}$$

对于简单的约束条件，我们能用解析的方法找到最优的概率分布，但对于一般性问题，这种方法是不可行的。为解决一般性约束最优化问题，我们应用了约束最优化理论中的 Lagrange 乘子定理解决这个问题。首先对模型中的每一个特征 f_i 都引入一个参数 λ_i，即 Lagrange 乘子。由条件熵定义 $H(p)$ 和约束条件 $p(f) = \tilde{p}(f)$，我们定义 Lagrange 函数为 $\Lambda(p, \lambda)$，

$$\Lambda(p, \lambda) = H(p) + \sum_{i=1}^{m} \lambda_i (p(f_i) - \tilde{p}(f_i)) \tag{26}$$

假设 Lagrange 函数 $\Lambda(p, \lambda)$ 中的变量 λ 固定，我们可以求出无约束的 Lagrange 函数 $\Lambda(p, \lambda)$ 的最大值时的 p， $p \in \mathrm{P}$。我们定义 λ 固定时 Lagrange 函数 $\Lambda(p, \lambda)$ 的最大值记为 $\Psi(\lambda)$， $\Lambda(p, \lambda)$ 为最大值时的 p 记为 p_λ，即有：

$$p_\lambda \equiv agr \max_{p \in \mathrm{P}} \Lambda(p, \lambda) \tag{27a}$$

$$\Psi(\lambda) \equiv \Lambda(p_\lambda, \lambda) \tag{27b}$$

$\Psi(\lambda)$ 是对偶函数，即 $\lambda^* = \arg\max_{i \in \{1,2,\cdots,m\}} \Psi(\lambda_i)$ 与 $p^* = \arg\max_{p \in C} H(p)$ 是对偶关系。称 $p^* = \arg\max_{p \in C} H(p)$ 为原问题，$\lambda^* = \arg\max_{i \in \{1,2,\cdots,m\}} \Psi(\lambda_i)$ 为对偶问题，利用 KT(Kuhn-Tucker) 对偶定理：在合适条件下，初始问题和对偶问题是紧密联系的。如果 λ^* 是对偶问题的解，那么 p^* 就是初始化问题的解， $p^* = p_\lambda$。下面给出求 p_λ 和 $\Psi(\lambda)$ 的过程：

$$\frac{\partial \Lambda(p, \lambda)}{\partial p} = -\sum_{x,y} \tilde{p}(x)(\log^{p(y|x)} + 1) + \sum_{x,y} \tilde{p}(x) \sum_i \lambda_i f_i(x, y) \tag{28}$$

使上式 $\Lambda(p, \lambda)$ 对 p 的一阶偏导数等于零，可得到：

$$p_\lambda^* = \frac{1}{e} \exp(\sum_i \lambda_i f_i(x, y)) \tag{29}$$

为使 p_λ^* 满足经典概率规则 $\sum_y p_\lambda^*(y \mid x) = 1$，需要引入归一化因子 $Z_\lambda^*(x)$，则有，

$$Z_\lambda^* = \frac{1}{e} \sum_y \exp(\sum_i \lambda_i f_i(x, y)) \tag{30}$$

常数 $1/e$ 对概率分布和求最值没有影响，所以概率 p_λ 和归一化因子 $Z_\lambda(x)$ 可写为下列形式：

$$p_\lambda = \frac{1}{Z_\lambda(x)} \exp(\sum_i \lambda_i f_i(x, y)) \tag{31}$$

$$Z_\lambda(x) = \sum_y \exp(\sum_i \lambda_i f_i(x, y)) \tag{32}$$

$\Psi(\lambda) \equiv \Lambda(p_\lambda, \lambda)$ 和 p_λ 可知：

$$\Psi(\lambda) = -\sum_{x} \tilde{p}(x)\log Z_{\lambda}(x) + \sum_{i} \lambda_i \tilde{p}(f_i) \tag{33}$$

由对偶最优化理论可知，我们的最终问题求解 $\lambda^* = \arg\max_{i\in\{1,2,\cdots,m\}} \Psi(\lambda_i)$，可以用无约束最优化方法找到 $\Psi(\lambda)$ 最大时的 λ。在求解对偶问题之前，我们先给出 $\Psi(\lambda)$ 与对数似然函数的关系。

设 $L_{\tilde{p}}(p)$ 是由模型估计的经验分布 $\tilde{p}$ 的对数似然，则：

$$L_{\tilde{p}}(p) = \log \prod_{x,y} p(y \mid x)^{\tilde{p}(x,y)} = \sum_{x,y} \tilde{p}(x,y)\log p(y \mid x) \tag{34}$$

事实上可以推知，指数分布模型 p_{λ} 的对数最大似然与 $\Psi(\lambda)$ 存在相等关系，即：

$$\Psi(\lambda) = L_{\tilde{p}}(p_{\lambda}) \tag{35}$$

求解 λ^* 转化为求解最大对数似然时的 λ，对数似然函数 $L_{\tilde{p}}(p)$ 是平滑的凸函数，存在最优解 λ^*。可以用坐标爬升法、梯度爬升法和共轭梯度法等数学方法来计算 λ^*。在实现最大熵的估计时利用了 IIS(Improved Iterative Scaling)迭代算法来求解。在此不作介绍。我们介绍最大熵模型的目的有两个，一是得到条件概率的指数形式，二是说明求最大熵的实质是求对数似然函数的最大值。

利用最大熵模型建模时，我们只需集中精力选择特征，而不需要花费精力考虑如何使用这些特征。该模型的另一个优点是特征选择灵活，且不需要额外的独立性假设或内在约束。但最大熵模型时空开销大，存在严重的数据稀疏问题，需要进行平滑处理，且对语料库的依赖性大。下面一节介绍条件随机场理论。

4 条件随机场理论

条件随机场(Conditional Random Fields, CRFs)[12]最早由 Lafferty 等人于 2001 年提出的，其模型思想的主要来源是最大熵模型，模型的三个基本问题的解决用到了 HMMs 模型中提到的方法如 forward-backward 和 Viterbi。我们可以把条件随机场看成是一个无向图模型或马尔可夫随机场，它是一种用来标记和切分序列化数据的统计模型。该模型是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率，而不是在给定当前状态条件下，定义下一个状态的分布。标记序列(Label Sequence)的分布条件属性，可以让 CRFs 很好的拟和现实数据，而在这些数据中，标记序列的条件概率信赖于观察序列中非独立的、相互作用的特征，并通过赋予特征以不同权值来表示特征的重要程度。

4.1 条件随机场定义

在以下的条件随机场模型介绍中，随机变量 X 表示需要标记的观察序列集。随机变量 Y 表示相应的表示标记序列集。所有的 $\mathrm{Y}_i \in \mathrm{Y}$ 被假设在一个大小为 N 的有限字符集内。随机变量 X 和 Y 是联合分布，但在判别式模型中我们构造一个关于观察序列和标记序列的条件概率模型 $p(Y \mid X)$ 和一个隐含的边缘概率模型 $p(X)$。下面给出条件随机场定义：

条件随机场定义：令 $G = (V, E)$ 表示一个无向图，$\mathrm{Y} = (\mathrm{Y}_v)_{v\in V}$，$\mathrm{Y}$ 中元素与无向图 G 中的顶点一一对应。当在条件 X 下，随机变量 Y_v 的条件概率分布服从图的马尔可夫属性：$p(\mathrm{Y}_v \mid \mathrm{X}, \mathrm{Y}_w, w \neq v) = p(\mathrm{Y}_v \mid \mathrm{X}, \mathrm{Y}_w, w \sim v)$，其中 $w \sim v$ 表示 (w, v) 是无向图 G 的边。这

时我们称 (X,Y) 是一个条件随机场。

4.2 势函数

尽管在给定每个节点的条件下，分配给该节点一个条件概率是可能的，但条件随机场的无向性很难保证每个节点在给定它的邻接点条件下得到的条件概率和以图中其它节点为条件得到的条件概率一致。因此导致我们不能用条件概率参数化表示联合概率，而要从一组条件独立的原则中找出一系列局部函数的乘积来表示联合概率。选择局部函数时，必须保证能够通过分解联合概率使没有边的两个节点不出现在同一局部函数中。最简单的局部函数是定义在图结构中的最大团(clique)上的势函数(Potential function)，并且是严格正实值的函数形式。但是一组正实数函数的乘积并不能满足概率公理，则必须引入一个归一化因子 Z，这样可以确保势函数的乘积满足概率公理，且是 G 中节点所表示的随机变量的联合概率分布。

$$Z=\sum_{v_i}\prod_{c\in C}\Phi_{v_c}(v_c) \tag{36}$$

其中 C 为最大团集合，利用 Hammersley-Clifford 定理，可以得到联合概率公式如下：

$$p(v_1,v_2,\cdots,v_N)=\frac{1}{Z}\prod_{c\in C}\Phi_{v_c}(v_c) \tag{37}$$

基于条件独立的概念，条件随机场的无向图结构可以用来把关于 $\mathrm{Y}_v\in\mathrm{Y}$ 的联合分布因式化正的和实值的势函数的乘积，每个势函数操作在一个由 G 中顶点组成的随机变量子集上。根据无向图模型条件独立的定义，如果两个顶点间没有边，则意味着这顶点这些顶点对应的随机变量在给定图中其它顶点条件下是条件独立的。所以在因式化条件独立的随机变量联合概率时，必须确保这些随机变量不在同一个势函数中。满足这个要求的最容易的方法是要求每个势函数操作在一个图 G 的最大团上，这些最大团由随机变量相应顶点组成。这确保了没有边的顶点在不同的势函数中，在同一个最大团中的顶点都是有边相连的。在无向图中，任何一个全连通（任意两个顶点间都有边相连）的子图称为一个团(clique)，而称不能被其它团所包含的才为最大团(maximal clique)。

理论上讲，图 G 的结构为任意，然而，在构造模型时，CRFs 采用了最简单和最重要的一阶链式结构。如下图所示，条件随机场 (X,Y) 以观察序列 X 作为全局条件，并且不对 X 做任何假设。这种简单结构可以被用来在标记序列上定义一个联合概率分布 $p(y\mid x)$，我们主要关心的是两个序列：$\mathrm{X}=(\mathrm{X}_1,\mathrm{X}_2,\cdots,\mathrm{X}_T)$ 和 $\mathrm{Y}=(\mathrm{Y}_1,\mathrm{Y}_2,\cdots,\mathrm{Y}_T)$。

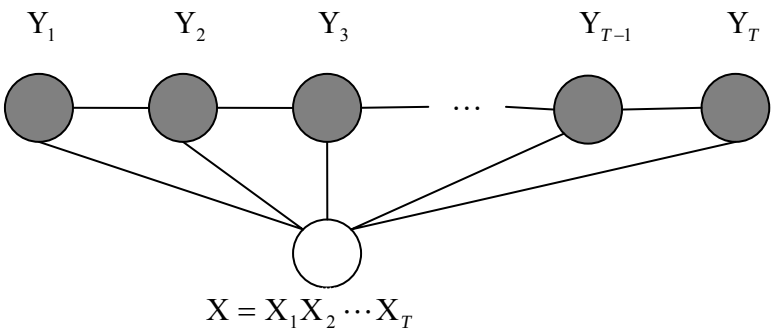

图 4 CRFs

4.3 条件随机场概率模型的形式

Lafferty 对 CRFs 势函数的选择很大程度上受最大熵模型的影响[13]。定义每个势函数的形式如下：

$$\Phi_{y_c}(y_c)=\exp(\sum_k \lambda_k f_k(c,y|c,x)) \tag{38}$$

其中 $y|c$ 表示第 c 个团中的节点对应的随机变量， f_k 是一个布尔型的特征函数，则 $p(y|x)$ 为：

$$p(y|x)=\frac{1}{Z(x)}\exp(\sum_{c\in C}\sum_k \lambda_k f_k(c,y_c,x) \tag{39}$$

其中 $Z(x)$ 是归一化因子，

$$Z(x)=\sum_y \exp(\sum_{c\in C}\sum_k \lambda_k f_k(c,y_c,x) \tag{40}$$

在一阶链式结构的图 $G=(V,E)$ 中，最大团仅包含相邻的两个节点，即是图 G 中的边。对于一个最大团中的无向边 $e=(v_{i-1},v_i)$，势函数一般表达形式可扩展为：

$$\Phi_{y_c}(y_c)=\exp(\sum_k \lambda_k t_k(y_{i-1},y_i,x,i)+\sum_k \mu_k s_k(y_i,x,i)) \tag{41}$$

其中 $t_k(y_{i-1},y_i,x,i)$ 是整个观察序列和相应标记序列在 $i-1$ 和 i 时刻的特征，是一个转移函数。而 $s_k(y_i,x,i)$ 是在 i 时刻整个观察序列和标记的特征，是一个状态函数。联合概率的表达形式可以写为：

$$p(y|x)=\frac{1}{Z(x)}\exp(\sum_i\sum_k \lambda_k t_k(y_{i-1},y_i,x,i)+\sum_i\sum_k \mu_k s_k(y_i,x,i)) \tag{42}$$

其中，参数 λ_k 和 μ_k 可以由从训练数据中估计，大的非负参数值意味着优先选择相应的特征事件，大的负值所对应的特征事件不太可能发生。

定义特征函数之前，先构造观察序列的实数值特征 $b(x,i)$ 集合来描述训练数据的经验分布特征，这些特征与模型同分布。例如：

$$b(x,i)=\begin{cases}1 & if \quad x_i=september\\ 0 & otherwise\end{cases} \tag{43}$$

每个特征函数表示为观察序列的实数值特征 $b(x,i)$ 集合中的一个元素，如果当前状态（状态函数）或前一个状态和当前状态（转移函数）具有特定的值，则所有的特征函数都是实数值。例如转移函数：

$$t_k(y_{i-1},y_i,x,i)=\begin{cases}b(x,i) & if \quad y_{i-1}=B,y_i=M\\ 0 & otherwise\end{cases} \tag{44}$$

为了统一转移函数和状态函数的表达形式，我们可以把状态函数写为下式：

$$s_k(y_i,x,i)=s_k(y_{i-1},y_i,x,i) \tag{45}$$

并用 $f_k(y_{i-1},y_i,x,i)$ 统一表示， f_k 可能是状态函数 $s_k(y_{i-1},y_i,x,i)$ 或转移函数 $t_k(y_{i-1},y_i,x,i)$，又令：

$$F_k(y,x)=\sum_{i=1}^{T} f_k(y_{i-1},y_i,x,i) \tag{46}$$

从而结定观察序列 x 条件下，相应的标记序列为 y 的概率可以写为：

$$p(y\mid x)=\frac{1}{Z(x)}\exp(\sum_k \lambda_k F_k(y,x)) \tag{47}$$

其中 $Z(x)$ 是归一化因子。

4.4 条件随机场模型的参数估计

以上的 CRFs 理论介绍给出了 CRFs 的概率形式公式，主要是基于最大熵理论，下面将介绍的是 CRFs 模型的参数估计，由最大熵模型可知参数估计的实质是对概率的对数最大似然函数求最值，即运用最优化理论循环迭代，直到函数收敛或达到给定的迭代次数。

假设给定训练集 $D=\{(\mathrm{X}_1,\mathrm{Y}_1),(\mathrm{X}_2,\mathrm{Y}_2),\cdots,(\mathrm{X}_\Gamma,\mathrm{Y}_\Gamma)\}$，根据最大熵模型对参数 λ 估计采用最大似然估计法。条件概率 $p(y\mid x,\lambda)$ 的对数似然函数形式为：

$$L(\lambda)=\log\prod_{x,y} p(y\mid x,\lambda)^{\tilde{p}(x,y)}=\sum_{x,y}\tilde{p}(x,y)\log p(y\mid x,\lambda) \tag{48}$$

已知条件概率 $p(y\mid x,\lambda)$ 的形式化公式为：

$$p(y\mid x,\lambda)=\frac{1}{Z(x)}\exp(\sum_k \lambda_k F_k(y,x))$$

其中归一化因子 $Z(x)$ 的表达式为：

$$Z(x)=\sum_y \exp(\sum_k \lambda_k F_k(y,x))$$

对于该 CRFs 概率模型来说，对数最大似然参数估计的任务是从相互独立的训练数据中估计参数 $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_n)$ 的值，则对数似然函数可写为下式：

$$L(\lambda)=\sum_{x,y}\tilde{p}(x,y)\sum_k \lambda_k F_k(y,x)-\sum_x \tilde{p}(x)\log Z(x) \tag{49}$$

为了表达，假设链式结构的无向图分别有一个特殊的起始节点和终止节点，分别用 Y_0 和 Y_{n+1} 表示。则经验分布概率和由模型得到的概率的数学期望为：

$$E_{\tilde{p}}[f_k]\overset{def}{=}\sum_{x,y}\tilde{p}(x,y)\sum_{i=1}^{n+1} f_k(y_{i-1},y_i,x,i)=\sum_{x,y}\tilde{p}(x,y)F_k(x,y)=E_{\tilde{p}}[F_k] \tag{50}$$

$$E_p[f_k]\overset{def}{=}\sum_{x,y}\tilde{p}(x)p(y\mid x,\lambda)\sum_{i=1}^{n+1} f_k(y_{i-1},y_i,x,i)=\sum_{x,y}\tilde{p}(x)p(y\mid x,\lambda)F_k(x,y)=E_p[F_k] \tag{51}$$

我们根据对数似然函数对相应的参数 λ_k 求一阶偏导数：

$$\frac{\partial L(\lambda)}{\partial\lambda_k}=\sum_{x,y}\tilde{p}(x,y)F_k(y,x)-\sum_x\left[\tilde{p}(x)\frac{\sum_y\left[\exp(\sum_k F_k(x,y))\cdot F_k(x,y)\right]}{Z(x)}\right]$$

$$= E_{\tilde{p}}[F_k] - \sum_{x,y} \tilde{p}(x) \frac{\exp(\sum_k F_k(x,y)) \cdot F_k(x,y)}{Z(x)}$$
$$= E_{\tilde{p}}[F_k] - \sum_{x,y} \tilde{p}(x) p(y \mid x, \lambda) F_k(x,y)$$
$$= E_{\tilde{p}}[F_k] - E_p[F_k] \quad (52)$$

通过梯度为零来求解参数 λ 并不一定总是得到一个近似解，因而需要利用一些迭代技术来选择参数，使对数似然函数最大化。通常采用的方法是改进的迭代缩放（Improved Iterative Scaling, IIS）或者基于梯度的方法来计算参数。以上的介绍中，我们给出了对数似然函数 $L(\lambda)$ 梯度的计算表达形式，即经验分布 $\tilde{p}(x,y)$ 的数学期望与由模型得到的条件概率 $p(y \mid x, \lambda)$ 的数学期望的差。而经验分布的数学期望为训练数据集中随机变量 (x,y) 满足特征约束的个数，模型的条件概率的数学期望的计算实质上是计算条件概率 $p(y \mid x, \lambda)$，在下一节中我们将介绍条件概率的有效计算方法。

4.5 条件概率的矩阵计算

建立条件随机场模型的主要任务是从训练数据中估计特征的权重 λ。下面主要对 CRFs 用到最大似然估计方法进行介绍。

由上面可知，条件随机场对数似然函数的梯度公式如下：

$$\frac{\partial L(\lambda)}{\partial \lambda_k} = E_{\tilde{p}(x,y)}[F_k] - E_{p(y|x,\lambda)}[F_k]$$

如果直接使用对数最大似然估计，可能会发生过度学习问题，通常引入罚函数的方法解决这一问题。如使用惩罚项 $\frac{\sum_k \lambda_k^2}{2\sigma^2}$，则对数似然函数和对数似然梯度公式变为：

$$L(\lambda) = \sum_{x,y} \tilde{p}(x,y) \sum_k \lambda_k F_k(y,x) - \sum_x \tilde{p}(x) \log Z(x) - \frac{\sum_k \lambda_k^2}{2\sigma^2} \quad (53)$$

$$\frac{\partial L(\lambda)}{\partial \lambda_k} = E_{\tilde{p}(x,y)}[F_k] - E_{p(y|x,\lambda)}[F_k] - \frac{\lambda_k}{\sigma^2} \quad (54)$$

由此，参数估计问题可以用最优化方法解决，可以使用迭代方法或 L-BFGS 算法。

对于一个链式条件随机场，我们在图的模型中添加一个开始状态 Y_0 和一个结束状态 Y_{n+1}。Y 为按字母排序的标记列表，$y_{i-1} = y'$ 和 $y_i = y$ 是取自该列表中的标记。我们定义一组矩阵 $\{M_i(x) \mid i = 1,2,\cdots,n+1\}$，其中每个 $M_i(x)$ 是 Y×Y 阶的随机变量矩阵。$M_i(x)$ 中的每个元素 $M_i(y_{i-1}, y_i \mid x)$ 定义为：

$$M_i(y_{i-1} = y', y_i = y \mid x) = \exp(\sum_k \lambda_k f_k(y_{i-1}, y_i, x, i))$$
$$= \exp(\sum_k \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i)) \quad (55)$$

其中 y_{i-1} 为 y_i 的前一个标记，$M_i(y_{i-1}, y_i \mid x)$ 是前一个状态到当前状态的转移概率与当前状态以观察序列为条件的概率的乘积，图示如下：

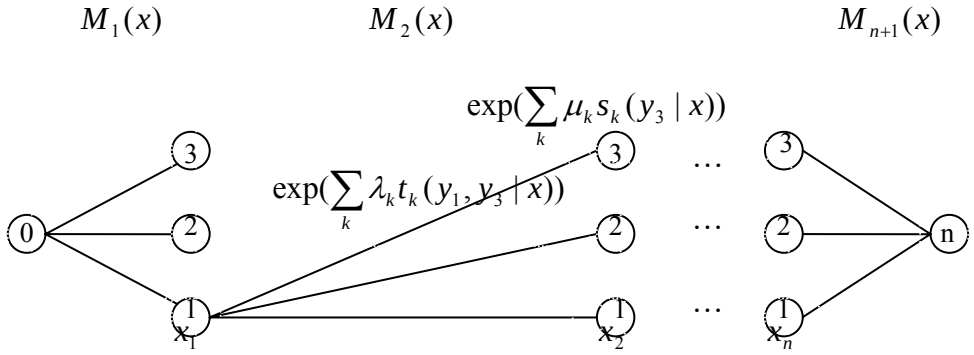

图 5 M 矩阵计算

如上图显示 $M_2(y_1, y_3 \mid x) = \exp(\sum_k \lambda_k t_k(y_1, y_3 \mid x) + \sum_k \lambda_k s_k(y_3 \mid x))$，且有：

$$M_1(x) = \begin{bmatrix} M_1(y_0, y_1 \mid x) & M_1(y_0, y_2 \mid x) & M_1(y_0, y_3 \mid x) \end{bmatrix}$$

$$M_2(x) = \begin{bmatrix} M_2(y_1, y_1 \mid x) & M_2(y_1, y_2 \mid x) & M_2(y_1, y_3 \mid x) \\ M_2(y_2, y_1 \mid x) & M_2(y_2, y_2 \mid x) & M_2(y_2, y_3 \mid x) \\ M_2(y_3, y_1 \mid x) & M_2(y_3, y_2 \mid x) & M_2(y_3, y_3 \mid x) \end{bmatrix}$$

$$M_{n+1}(x) = \begin{bmatrix} M_{n+1}(y_1, y_n \mid x) \\ M_{n+1}(y_2, y_n \mid x) \\ M_{n+1}(y_3, y_n \mid x) \end{bmatrix}$$

因为 $p(y \mid x, \lambda)$ 实际上是从开始节点到结点节点的一条路径的概率，所以有：

$$p(y \mid x, \lambda) = \frac{1}{Z(x)} \prod_{i=1}^{n+1} M_i(y_{i-1}, y_i \mid x) \tag{56}$$

其中 $Z(x)$ 为归一化因子，为所有路径概率的和，表达式如下：

$$Z(x) = \prod_{i=1}^{n+1} M_i(x) \tag{57}$$

不论是使用迭代缩放还是 L-BFGS 算法进行参数估计与训练，为了计算最大似然参数值，就需要对训练数据中的每个观察值 X 对应的标记序列的条件概率相对特征函数的数学期望进行有效的计算。枚举计算是不可行的，Lafferty 提出了动态规划方法来计算 $E_{p(y|x,\lambda)}[f_k]$。

$$E_p[f_k] \overset{def}{=} \sum_{x,y} \tilde{p}(x) p(y \mid x, \lambda) \sum_{i=1}^{n+1} f_k(y_{i-1}, y_i, x, i)$$

上式的右边可以改写为：

$$\sum_x \tilde{p}(x) \sum_{i=1}^{n+1} \sum_{y',y} p(y_{i-1} = y', y_i = y \mid x, \lambda) f_k(y_{i-1} = y', y_i = y, x, i) \tag{58}$$

而后，可使用动态规划方法计算 $p(y_{i-1}, y_i \mid x, \lambda)$，该方法与隐马尔可夫模型中介绍的 forward-backward 算法类似，我们分别定义 forward 向量和 backward 向量为 $\alpha_i(x)$ 和 $\beta_i(x)$，则构造步骤如下：

$$\alpha_0(y \mid x) = \begin{cases} 1 & if \quad y = start \\ 0 & otherwise \end{cases}$$

$$\beta_{n+1}(y \mid x) = \begin{cases} 1 & if \quad y = stop \\ 0 & otherwise \end{cases} \tag{59}$$

递归关系表示为：

$$\alpha_i(x)^T = \alpha_{i-1}(x)^T M_i(x)$$

$$\beta_i(x) = M_{i+1}(x)\beta_{i+1}(x) \tag{60}$$

由以上公式可知，在给定观察序列 x 条件下， $y_{i-1} = y'$ 和 $y_i = y$ 的概率为：

$$p(y_{i-1} = y', y_i = y \mid x) = \frac{\alpha_{i-1}(y' \mid x) M_i(y', y \mid x) \beta_i(y \mid x)}{Z(x)} \tag{61}$$

由上式可以有效计算出条件概率的数学期望。

5 总结

本文介绍了中文分词领域广泛应用的统计模型，如隐马尔可夫模型、最大熵模型和条件随机场模型。这三种模型中，HMMs 需要对观察序列建模，且要求严格的输出独立性假设，导致其不能考虑上下文的特征，限制了特征的选择。MEMs 提出了一种条件模型，能够把多种特征信息纳入一个模型中，它用最大熵原理构造分布模型并利用对偶定理得出参数估计的实质是求最大对数似然估计。CRFs 具有其它模型的优点，且不存在标记偏置问题。

参考文献

[1] http://crfpp.sourceforge.net/
[2] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. IEEE, VOL.77,No.2,pp:257-286. 1989,2.
[3] L. E. Baum and J. A. Egon. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. Bull. Amer. Meteorol. Soc.,vol. 73, pp.360-363,1967.
[4] L. E. Baum and G. R. Sell. Growth functions for transformations on manifolds. Pac. J. Math., vol.27, no.2, 1968, pp.211-227.
[5] A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. IEEE Trans. Informat. Theory, vol.IT-13, Apr, 1967, pp.260-269.
[6] G. D. Forney. The Viterbi algorithm. Proc. IEEE, vol.61, Mar, 1973, pp. 268-278.
[7] Adam L. Berger, Vincent J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. Association for Computational Linguistics, vol.22, 1996, pp.39-48.
[8] A. Ratnaparkhi. A simple introduction to maximum entropy models for natural language processing. Technical Report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania, 1997.
[9] Ratnaparkhi, A. A Maximum Entropy Model for Part-of-Speech Tagging. Proc. EMNLP. New Brunswick, New Jersey: Association for Computational Linguistics. 1996.
[10] E. T. Jaynes. Notes on Present Status and Future Prospects. In Maximum Entropy and Bayesian Methods, edited by W. T. Grandy and L. H. Schick. Kluwer, 1990, pp.1-13.
[11] C. E. Shannon. A mathematical theory of communication. Bell System Tech. Journal, 27, 1948, pp.379–423 and 623–656.
[12] J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In International Conference on Machine Learning, 2001.
[13] Hanna M. Wallach. Conditional Random Fields: An Introduction. University of Pennsylvania CIS Technical Report MS-CIS-04-21. 2004.

Conditional Random Fields Theory Review

Xuedong Han[1], Caigen Zhou[2]
[1]Beijing University of Posts and Telecommunications, Beijing, P.R.China (100876)
[2]Institute of People's Liberation Army General Staff, 54th, Beijing, P.R.China (100083)

Abstract

Conditional Random Fields theory can be used to segmenting or labeling sequential data, text chunking and other natural language processing tasks. Chinese word, Chinese name recognition, ambiguity resolution in such as in applications In the Chinese natural language processing tasks where good performance. This Conditional Random Fields Theory and Introduction of paper is different from other papers, that we give probability model derivation, the reasons of parameter estimation function in the form of log-likelihood function, and example of conditional probability in matrix calculation, enabling readers to grasp the Conditional Random Fields theoretical basis and as a whole.

Key Words: Conditional Random Fields; CRFs; MEMs; Chinese word