中文信息处理

基于最大熵模型的名实体识别

刘秉权
智能技术与自然语言处理实验室
哈尔滨工业大学
liubq@hit.edu.cn

教学目的

- 学习了解ME模型的原理
- 学习如何使用ME模型解决名实体识别问题
- 结合ME模型，深入了解特征选择、权值偏置、特征融合等问题
- 了解如何使用ME模型解决人名起源问题

主要内容

- 1. 最大熵模型(ME)
- 2. 基于ME的中文名实体识别
- 3. 基于ME的人名起源识别
- 4. 小结

1. 最大熵模型
Maximum Entropy Model

- 1.1 最大熵模型的引入
- 1.2 最大熵模型的形式化描述
- 1.3 参数估计
- 1.4 基于信息增益的特征选择
- 1.5 最大熵马尔科夫模型

1.1 最大熵模型的引入

什么是熵

定义:

$$H(X)=\sum_{i=1}^{k} p(x=x_i)\log\frac{1}{p(x=x_i)}$$

X 的具体内容跟信息量无关，我们只关心概率分布，于是 H(X)可以写成：

$$H(X)=\sum_{x\in X} p(x)\log\frac{1}{p(x)}$$

熵的性质

$$0 \le H(X) \le \log|X|$$

第一个等号在 X 为确定值的时候成立（没有变化的可能）。

第二个等号在 X 均匀分布的时候成立。

均匀分布的时候，熵最大。

条件熵(Conditional Entropy)

有两个变量：x,y。它们不是独立的。已知 y，x 的不确定度又是多少呢?

$$H(X|Y)=\sum_{(x,y)\in X\times Y} p(x,y)\log\frac{1}{p(x|y)}$$

$$H(X|Y)=H(XY)-H(Y)$$

$$H(X|Y)\le H(X)$$

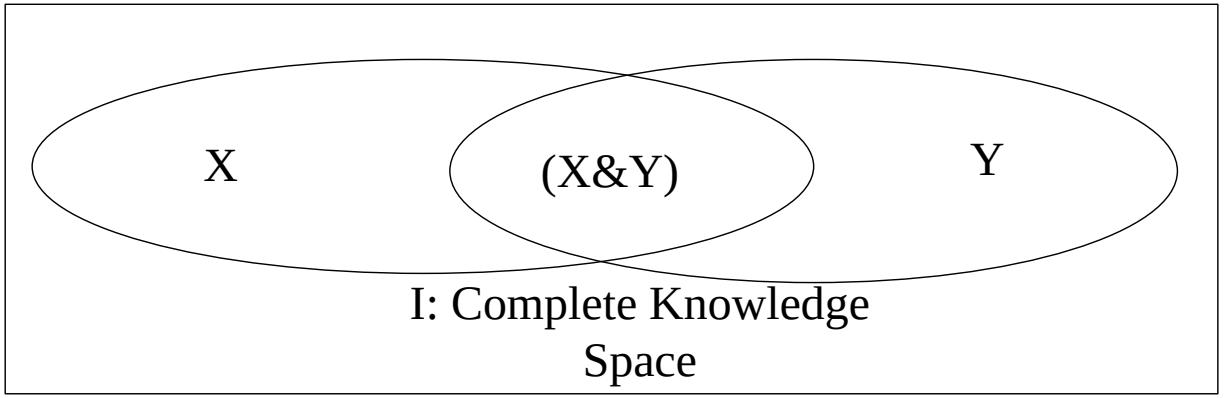
条件熵(Conditional Entropy)

$$H(X \mid Y) \le H(X)$$

Condition Reduces Entropy (C.R.E.)

知识（Y）减少不确定性（X）

用文氏图说明：



已知与未知的关系

对待已知事物和未知事物的原则：

- 承认已知事物（知识）
- 对未知事物不做任何假设，没有任何偏见

“知之为知之，不知为不知”

已知与未知的关系—例子

已知：

“学习”可能是动词，也可能是名词。可以被标为主语、谓语、宾语、定语……

令x_1表示“学习”被标为名词，x_2表示“学习”被标为动词。

令y_1表示“学习”被标为主语，y_2表示被标为谓语，y_3表示宾语，y_4表示定语。得到下面的表示：

$$p(x_1)+p(x_2)=1 \qquad \sum_{i=1}^{4} p(y_i)=1$$

如果仅仅知道这一点，根据无偏见原则，“学习”被标为名词的概率与它被标为动词的概率相等。

$$p(x_1)=p(x_2)=0.5$$

$$p(y_1)=p(y_2)=p(y_3)=p(y_4)=0.25$$

已知与未知的关系—例子

$$p(x_1)+p(x_2)=1 \qquad \sum_{i=1}^{4} p(y_i)=1$$

已知：

“学习”可能是动词，也可能是名词。可以被标为主语、谓语、宾语、定语……

“学习”被标为定语的可能性很小，只有0.05

我们引入这个新的知识：$p(y_4)=0.05$

除此之外，仍然坚持无偏见原则：$p(x_1)=p(x_2)=0.5$

$$p(y_1)=p(y_2)=p(y_3)=\frac{0.95}{3}$$

已知与未知的关系—例子

$$p(x_1)+p(x_2)=1$$

$$\sum_{i=1}^{4} p(y_i)=1$$

已知：

“学习”可能是动词，也可能是名词。可以被标为主语、谓语、宾语、定语……

“学习”被标为定语的可能性很小，只有0.05 $p(y_4)=0.05$

当“学习”被标作动词的时候，它被标作谓语的概率为0.95

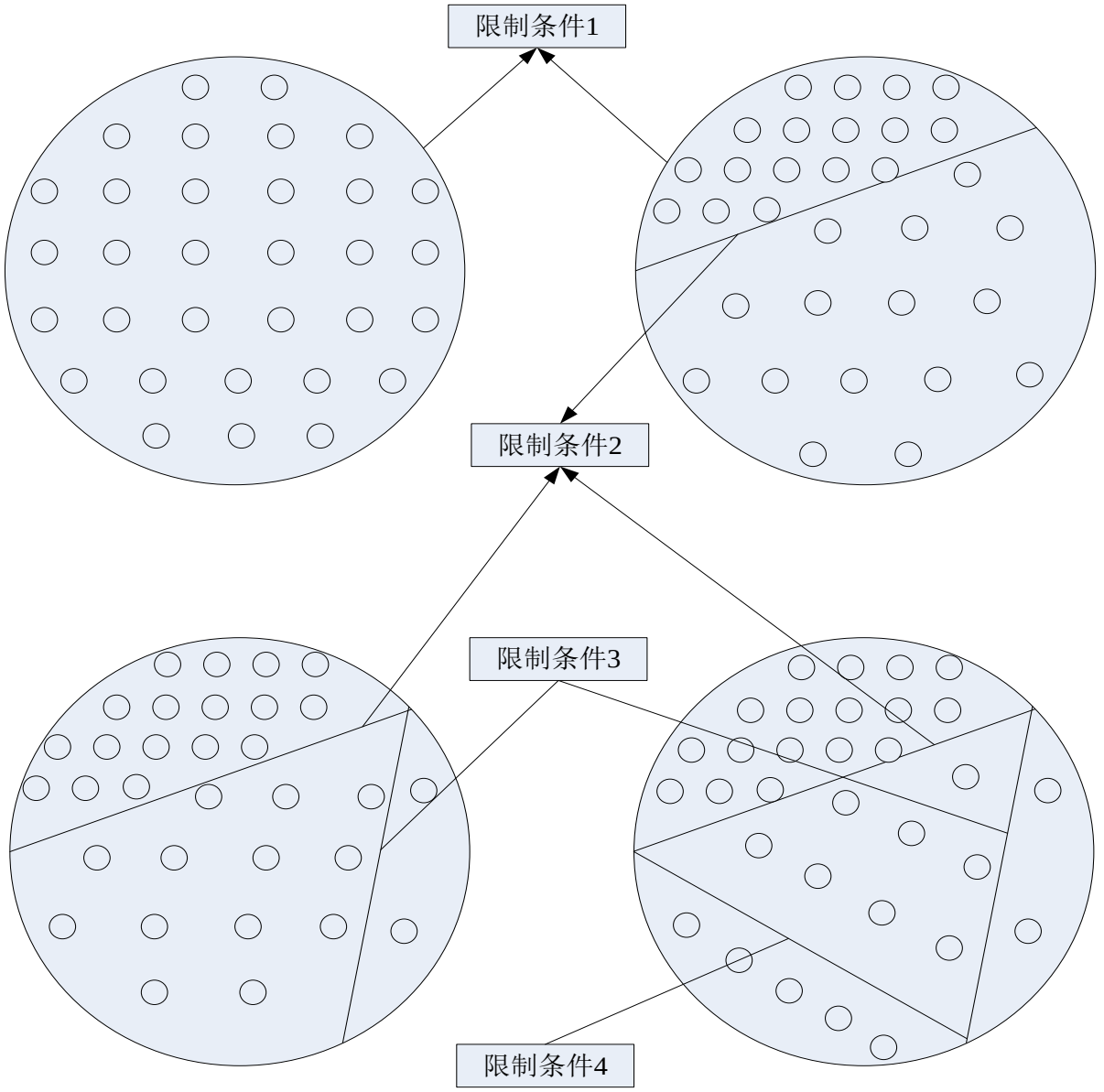
引入这个新的知识：

$$p(y_2 \mid x_1)=0.95$$

除此之外，仍然坚持无偏见原则，我们尽量使概率分布平均。

但问题是：什么是尽量平均的分布？

最大熵原理示意图



最大熵模型(Maximum Entropy)

- 概率平均分布〈=〉熵最大
- 我们要确定一个x和y的分布，满足：

$$p(x_1)+p(x_2)=1 \qquad \sum_{i=1}^{4} p(y_i)=1 \qquad p(y_4)=0.05$$

$$p(y_2 \mid x_1)=0.95$$

- 同时使$H(Y \mid X)$达到最大值

最大熵模型(Maximum Entropy)

$$\max H(Y \mid X) = \sum_{\substack{x \in \{x_1, x_2\} \\ y \in \{y_1, y_2, y_3, y_4\}}} p(x, y) \log \frac{1}{p(y \mid x)}$$

$$p(x_1) + p(x_2) = 1$$

$$p(y_1) + p(y_2) + p(y_3) + p(y_4) = 1$$

$$p(y_4) = 0.05$$

$$p(y_2 \mid x_1) = 0.95$$

最大熵模型(Maximum Entropy)

一般模型：$\max_{p \in P} H(Y \mid X) = \sum_{(x,y)} p(x,y) \log \frac{1}{p(y \mid x)}$

P={p|p是X上满足条件的概率分布}

What is Constraints?

--模型要与已知知识吻合

What is known?

--训练数据集合

特征(Feature)

特征：(x,y)

y:这个特征中需要确定的信息

x:这个特征中的上下文信息

注意一个标注可能在一种情况下是需要确定的信息，在另一种情况下是上下文信息 ：

x1x2…xn	p(y1=a|x1x2…xn)
x1x2…xn y1	p(y2=a|x1x2…xn y1)

样本(Sample)

关于某个特征(x,y)的样本--特征所描述的语言现象在标准集合里的分布：

(x_i,y_i) pairs

y_i是y的一个实例

x_i是y_i的上下文

(x_1,y_1) (x_2,y_2) (x_3,y_3)……

特征与样本

特征函数：对于一个特征(x0,y0)，定义特征函数：

$$f(x,y)=\begin{cases}1: & \text{如果}y=y_0\text{而且}x=x_0\\ 0: & \text{其他情况}\end{cases}$$

特征函数期望值：

对于一个特征(x0,y0)，在样本中的期望值是：

$$\overline{p}(f)=\sum_{(x_i,y_i)}\overline{p}(x,y)\,f(x,y)$$

$\overline{p}(x,y)$ 是(x,y)在样本中出现的概率

条件(Constraints)

条件：

对每一个特征(x,y)，模型所建立的条件概率分布要与训练样本表现出来的分布相同。

假设样本的分布是（已知）：$\overline{p}(x) = x$出现的概率

$\overline{p}(x,y) = xy$出现的概率　　$\overline{p}(f) =$ 特征f在样本中的期望值

特征f在模型中的期望值：

$$p(f) = \sum_{(x_i,y_i)} p(x_i,y_i) f(x_i,y_i)$$
$$= \sum_{(x_i,y_i)} p(y_i \mid x_i) p(x_i) f(x_i,y_i)$$
$$= \sum_{(x_i,y_i)} p(y_i \mid x_i) \overline{p}(x_i) f(x_i,y_i)$$

$$p(f) = \overline{p}(f)$$

最大熵模型

$$p^* = \arg\max_{p \in P} H(Y \mid X)$$

P={p|p是y|x的概率分布并且满足下面的条件}

对训练样本，对任意给定的特征fi：

$$p(f_i) = \overline{p}(f_i)$$

最大熵模型

$$p^* = \arg\max_{p \in P} \sum_{(x,y)} p(y \mid x)\overline{p}(x)\log\frac{1}{p(y \mid x)}$$

$$P = \left\{ p(y \mid x) \;\middle|\; \begin{array}{ll} \forall f_i : & \sum_{(x,y)} p(y \mid x)\overline{p}(x) f_i(x,y) \\ & = \sum_{(x,y)} \overline{p}(x,y) f_i(x,y) \\ \forall x : & \sum_{y} p(y \mid x) = 1 \end{array} \right\}$$

最大熵模型的求解

- 问题：

已知若干条件，求若干变量的值使得目标函数（熵）最大

- 数学本质：

最优化问题（Optimization Problem）

- 条件：线性、等式
- 目标函数：非线性

非线性规划（线性约束）(non-linear programming with linear constraints)

拉格朗日算子(Lagrange Multiplier)

- 一般地，对于k个限制条件的Constrained Optimization问题：

$$\max H(p) \qquad 1 \le i \le k : C_i(p) = b_i$$

- 拉格朗日函数为：

$$L(p,\lambda) = H(p) + \sum_{i=1}^{k} \lambda_i [C_i(p) - b_i]$$

- 其中引入的拉格朗日算子：

$$\lambda = [\lambda_1, ..., \lambda_k]^T$$

最优解（Exponential）

$$p^*(y \mid x) = ce^{\sum_i \lambda_i f_i(x,y)} \qquad c = \frac{1}{\sum_y e^{\sum_i \lambda_i f_i(x,y)}}$$

$$p^*(y \mid x) = \frac{1}{Z(x)} e^{\sum_i \lambda_i f_i(x,y)} \qquad Z(x) = \sum_y e^{\sum_i \lambda_i f_i(x,y)}$$

$$\lambda_i \, ?$$

最优解（Exponential）

λ?

- 几乎不可能有解析解（包含指数函数）
- 近似解不代表接近驻点。

- 能不能找到另一种逼近？比如……

等价成求某个函数 $f(\lambda)$ 的最大/最小值？

非线性规划中的对偶问题

递推公式：

$$\lambda_i^{(n+1)} = \lambda_i^{(n)} + c\left[E_{\tilde{p}} f_i - \sum_{x,y} \tilde{p}(x) p^*(y \mid x) f_j(x,y) \right]$$

1.2 最大熵模型的形式化描述

最大熵模型通过求解一个有条件约束的最优化问题来得到概率分布的表达式。

假设有 n 个学习样本 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$，其中 x_i 是由 k 个属性特征构成的样本向量 $x_i = \{x_{i1}, x_{i2}, \cdots, x_{ik}\}$， y_i 是类别标记 $y_i \in Y$。所要求解的问题是，在给定一个新样本 x 的情况下，其最佳的类别标记是什么

最大熵的目标函数被定义如下：

$$H(p) = -\sum \tilde{p}(x) p(y \mid x) \log p(y \mid x)$$

即最大熵模型要求信息系统的目标状态的条件熵取得最大值，同时要求满足下述条件：

$$P = \{p \mid E_p f_i = E_{\tilde{p}} f_i, 1 \le i \le k\}$$

$$\sum_y p(y \mid x) = 1$$

最大熵模型的形式化描述

式中 f_i 定义在样本集上的特征函数，$E_p f_i$ 表示特征 f_i 在模型中的期望值，$E_{\tilde{p}} f$ 表示特征 f_i 在训练集上的经验期望值。两种期望分别定义如下：

$$\begin{cases} E_p f_i = \sum_{c,h} \tilde{p}(x) p(y \mid x) f_i(y,x) \\ E_{\tilde{p}} f_i = \sum_{c,h} \tilde{p}(y,x) f_i(y,x) = \frac{1}{N} \sum f_i(y,x) \end{cases}$$

$$f_i(y,x) = \begin{cases} 1 & if: y = y' \ and \ h(x) = TRUE \\ 0 & else \end{cases}$$

其中 $h(x)$ 为谓词函数，其类型的个数和系统特征模板的类型个数相等。通过拉格朗日变换，求出满足条件极值的概率如下：

$$p(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_i \lambda_i f_i(y,x) \right)$$，其中 $Z(x) = \sum_c \exp\left(\sum_i \lambda_i f_i(y,x) \right)$

1.3 参数估计

λ_i 是特征 f_i 对应的拉格朗日系数，即权值。求解的目标问题转换为如何估计这些特征的权值。在最大熵模型中最多被使用的参数估计方法是迭代放大算法，GIS 算法，在实践中，为了方便计算，需要把指数形式变换为对数形式，所以最大熵模型也是对数线形模型的一种。

GIS 迭代算法需要满足一个限制条件，每个样本包含的特征数必须相等，即 $C = \max_{x \in X} \sum_{j=1}^{k} f_j(y, h(x))$

为此对于不完整数据的处理，必须引入一个补偿特征，来弥补丢失特征对分类结果的影响：

$$\forall x \in X \quad f_l(c, h(x)) = C - \sum_{j=1}^{k} f_j(c, h(x))$$

GIS参数估计算法

输入：样本集 $S=\{s_1,\cdots,s_n\}$,类别集合 $Y=\{y_1,\cdots,y_l\}$，

输出：模型参数集合，即所有特征的权值 $\lambda=\{\lambda_1,\cdots,\lambda_T\}$

1. 把所有的样本转换为事件，形成事件空间 $EVE=\{e_1,\cdots,e_m\}$　$m\le n$；
2. 计算各种特征，包括补偿特征的经验期望；
3. 设定初始模型 $\lambda^0=\{\lambda_1^0=0,\cdots,\lambda_T^0=0\}$，$p_{old}$= 0.0, p_{new}=1.0;
4. 前后两次模型的精度差(p_{new}-p_{old})是否为零，是则退出，否则继续；
5. p_{old}=p_{new};
6. from i=1 to i=m
   from j=1 to j=k

   （1） 用上一次模型 λ^{old} 代入公式 $p(y\mid x)=\frac{1}{Z(x)}\exp\left(\sum_i \lambda_i f_i(y,x)\right)$，计算每个类别的出现概率

   （2） 选择概率最大的类别作为该事件的输出，并与真实输出做比较，记录分类正确的个数
7. 计算在该模型下的训练精度 p_{new};
8. 利用公式更新各个特征的模型期望；
9. 利用迭代公式 $\lambda_j^{new}=\lambda_j^{old}+\frac{1}{k}(\log E_{\tilde{p}} f_j-\log E_p f_j)$ 计算新的模型；
10. $\lambda^{old}=\lambda^{new}$；
12. 跳转至第 5 步。

1.4 基于信息增益的特征选择

信息增益度：根据 Kullback-Leibler（KL）距离定义：

$$D(p \parallel q) = \sum_x p(x) \ln \frac{p(x)}{q(x)}$$

KL 距离越小说明两个分布越相似。在模型特征空间加入第 n 个特征前后模型分布与样本分布之间的 KL 距离为：

$$D(\tilde{p} \parallel p^{(n-1)}) = \sum_x \tilde{p}(x) \ln \frac{\tilde{p}(x)}{p^{(n-1)}(x)} \qquad D(\tilde{p} \parallel p^{(n)}) = \sum_x \tilde{p}(x) \ln \frac{\tilde{p}(x)}{p^{(n)}(x)}$$

这样，引入第 n 个特征 f_n 后的信息增益值为：

$$G(p, f_n) = D(\tilde{p} \parallel p^{(n-1)}) - D(\tilde{p} \parallel p^{(n)})$$

那么被选择加入特征空间的第 n 个特征为：

$$\hat{f}_n = \arg\max_{f_i} G(p, f_n)$$

1.5 最大熵马尔科夫模型

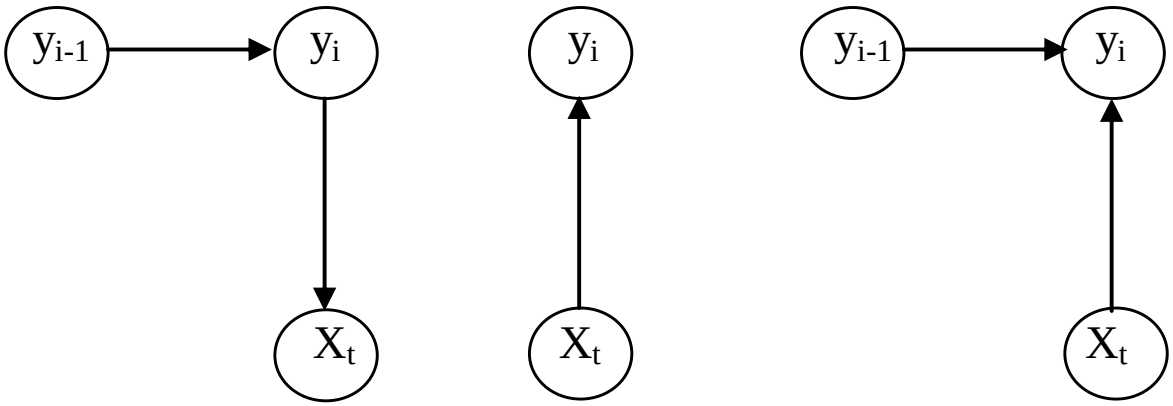
在自然语言处理领域，需求解序列标注问题。其输入都是一个马尔科夫链，即每个随机变量并非独立。为求解这类问题，在最大熵模型中引入马尔科夫过程，即最大熵马尔科夫模型。

最大熵马尔科夫模型（MEMM：Maximum Entropy Markov Model）是最大熵模型的一种延伸模型。

MEMM 求解一个条件概率，通过马尔科夫原理，把一个“标记链”的条件概率分解为多个单个标记链的乘积。HMM 是求解一个联合概率，不仅要对标记链建模，还要对观测链建模，因为在 HMM 中认为两者都是随机变量，而非确定值。

MEMM 的概率形式为指数模型，即可以继承所有 ME 模型的优点，能适用灵活的没有任何独立假设条件的特征，而 HMM 则不具有该优点。

HMM、ME和MEMM的结构比较



HMM　　ME　　MEMM

HMM 是由标记 y_{i-1} 到 y_i 转移，并且产生（发射）出观察变量 x_t 的，而 MEMM 是在 y_{i-1} 和 x_t 的基础上，产生标记 y_i 的。因此 MEMM 也经常被称之为有限状态接收机（Finite-state Acceptor）。

MEMM序列标注

在 MEMM 的框架下给定一个输入序列 $X=\{x_1,x_2,\cdots,x_T\}$，给定标记集 $Y=\{y_1,y_2,\cdots,y_n\}$，最适合 X 的标记序列 $\hat{Y}$ 可以表示为：

$$\hat{Y}=\arg\max_{Y} p(Y\mid Y',X)$$

$$p(Y\mid Y',X)=p(y_0\mid x_0)p(y_1\mid y_0,x_1)\cdots p(y_n\mid y_{n-1},x_n)$$

其中 $p(y_i\mid y_{i-1},x_i)$ 可以被切分为 $|Y|$ 个独立训练的状态转移方程：

$$p_{y_{i-1}}(y_i\mid x)=p(y_i\mid y_{i-1},x)$$

每个函数表示在每一个 y_i 下的条件概率，这些函数每一个都是以最大熵模型的形式，即指数线形模型的形式给出。

MEMM中Viterbi算法

迭代公式：

$$SeqScore(i, j) = SeqScore(k, j-1) \times p(x_j \mid y_j^i)\, p(y_j^i \mid y_{j-1}^k)$$

修正为：

$$SeqScore(i, j) = SeqScore(k, j-1) \times p_{y_{j-1}^k}(y_j^i \mid x_j)$$

MEMM中ME模型求解

MEMM 中的转移概率方程可以转换为多个 ME 模型，但是求解每个ME 模型的方式和前文中介绍的ME略有不同，即：

$$p_{y_{i-1}}(y_i \mid x_i) = \frac{1}{Z(x_i, y_{i-1})} \exp\left(\sum \lambda_i f_i(y_i, x_i)\right)$$

说明了 MEMM 是对前一个状态发射出的所有可能下一个状态进行归一化处理，同 ME 不同的在于：ME 只是对观测 x 所有可能出现的类别概率进行归一。

2. 基于最大熵模型的中文名实体识别

- 2.1 用于名实体识别的ME模型的特征
- 2.2 特征组合中的权值偏置问题
- 2.3 特征融合策略

2.1 用于名实体识别的ME模型的特征

- 把名实体识别的过程看作为一个多分类的过程，用最大熵模型来完成这一任务
- 需要识别的目标类别包括中文姓名、地名、机构名、其他专有名词四种
- 把每一种类别的名实体又细分为开始部分、中间部分、结尾部分和整体四种情况
- 把不属于以上四种类别中任何一种的词语都归为一类，这样类别标记集合总共包含4×4+1=17个标记
- 模型中的特征由一系列的特征模板产生

名实体识别用特征模板

	谓词特征	含　义
0	Current word	当前词
1	First Preceding word	第一个前接词
2	Second Preceding word	第二个前接词
3	First Succeeding word	第一个后接词
4	Second Succeeding word	第二个后接词
5	Preceding class label	前接词的类别标记
6	word Be Number	词是否为数字
7	word Be Chinese surname	词是否为中文姓
8	word Be cityname	词是否为城市名
9	suffix	词的后缀
10	Default	缺省特征

2.2 特征组合中的权值偏置问题

在使用最大熵模型时，首先需要在上下文环境中对要分类或者识别的词语提取出所有与其相关的特征，然后根据训练得到的权向量把这些特征进行线性组合。这种特征组合的方法在通常情况下能比较好地体现各个特征对分类的贡献程度。

但是这种线性组合方法存在特征权值的偏置（Weight Bias）问题。

示例：输入经分词后的语句“张 大庆”，假定被识别的当前词为“大庆”，而且其前接词，即“张”，已经被标记为 nf（name first：姓名开始部分）。通过特征提取算法，我们得到了和当前被标注词“大庆”相关的特征。

特征示例

特征类型 ID 号	特征值	预测类别 ID 号	频度	权值
0	大庆	7	58	1.93494
1	张	1	479	1.57131
1	张	2	223	1.30702
5	nf	1	12631	1.74149
5	nf	2	8328	1.78044
8	1	7	19002	1.49712
8	1	2	248	-1.4277
10	NULL	2	20960	-2.20704
10	NULL	7	43288	-1.2546

问题分析

由上表中的数据，并根据最大熵模型中的特征线性组合的方法，“大庆”被识别为姓名结束和地名的可能性分别为：

$$pro(end\ of\ \ personName) =$$

$$\exp\left(1.30702+1.78044-1.4277-2.20704\right)=\exp(-0.54728)$$

$$pro(\ place)=\exp\left(1.93494+1.49712-1.2546\right)=\exp(2.17745)$$

“大庆”作为城市名的概率要大于作为人名结束的概率，导致最终做出了错误的分类决策，即输出“张/nf 大庆/p”。 “大庆”在训练语料中大都是以地名的类别形式出现，其对应的特征是：

$$f(\text{大庆},\ place)=\begin{cases}1 & currentWord=\text{大庆}\ and\ c=place\\ 0 & else\end{cases}$$

该特征的权值是唯一不变的，即 1.93494，无论它前面的词和前面词的标记是什么。由此发现，在加入词典特征后，由于扩大的信息粒度，即对词语进行了聚类，而导致属于某个特定词典（类别）的词会继承整个类别中所有词的出现特征。体现在模型中就是，某些词的词典特征因为共享权值，而使得权值过度增加，无法正确识别出该词在特定上下文环境中的类别。这种特征权值的唯一性在进行特征的线性组合时，就导致了权值的偏置问题。

2.3 特征融合策略

特征融合：在特征产生之前，对原有特征模板进行组合以产生新的复合模板，然后再在训练数据集上对新的模板集合进行实例化，得到新的特征集合。

首先对前九个特征模板进行两两组合，得到 9×8=72 种新的特征模板；

然后利用特征模板和类别标记之间的平均互信息来选择分别适用于人名、地名和机构名的特征模板；

此外，为了利用 tri-gram 信息，对每一种类别都手工添加由前 5 个模板所组合而成的三种复合模板。用于特征模板选择的模板和类别的互信息被定义如下：

$$I(template, class) = H(template) - H(template \mid class)$$

$$= \sum_{i,j} p(template_i, class_j) \log \frac{p(template_i, class_j)}{p(template_i)\, p(class_j)}$$

$H(template) = -\sum_{i=0}^{16} p(c_i) \log(p(c_i))$ 表示一个模板 template 的信息熵；

$$H(template \mid class) = -\sum_{i,j} p(template_i, class_j) \log p(template_i \mid class_j)$$

表示模板 template 的条件熵。

通过设定一个阈值 ϑ，来决定模板的取舍。ϑ 的大小会影响最终特征集合的规模，进而影响模型的训练时间，同时该阈值对模型的精度和召回率也会有一定影响。

用于人名识别的复合特征

类型	ID	特征	含义
二元特征	11	Pretag_and_currentWord	前接词的标记和当前词的组合
	12	preWordIsNum_and_currentWord	前接词是否为数字和当前词的组合
	13	preWordIsNum_and_currentWordIsFisrtName	前接词是否为数字和当前词是否为中文姓的组合
	14	currentWordIsFisrtName_and_SucWord	当前词是否为中文姓和后接词的组合
	15	preWordIsFirstName_and_currentWordIsCity	前接词是否为中文姓和当前词是否为地名的组合
	16	preWord_and_currentWordIsCity	前接词和当前词是否为地名的组合
三元特征	17	preWord1_and_preWord2_and_currentWord	第一个前接词、第二个前接词和当前词的组合
	18	preWord2_and_currentWord_and_sucWord1	第二个前接词、当前词和第一个后接词的组合
	19	currentWord_and_sucWord1_and_sucWord2	当前词、第一个后接词和第二个后接词的组合

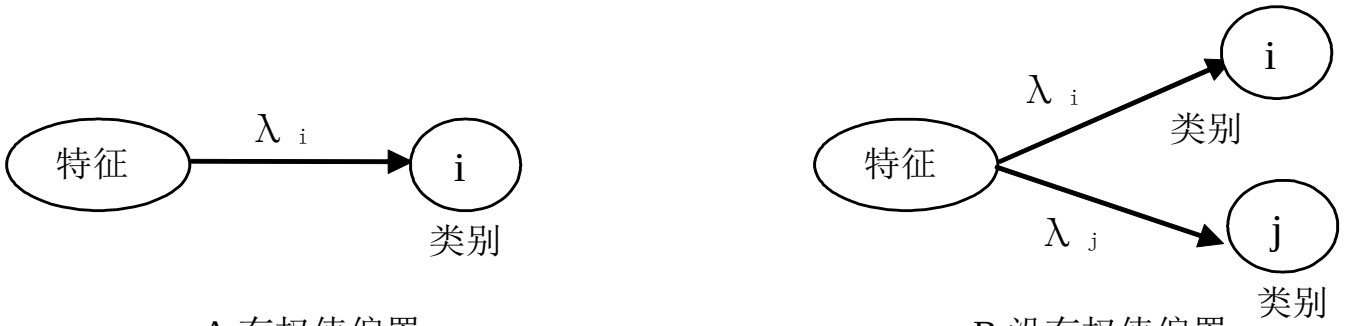
融合后的特征示例

有了以上的复合特征，重新标注输入样本“张 大庆”，多了以下三个特征：

特征类型 ID 号	特征值	预测类别 ID 号	频度	权值
11	nf/大庆	2	4	1.0142
15	1/1	2	15	1.46341
16	张/1	2	6	1.10702

解决权值偏置问题

比较 ID 为 11 的特征和 ID 为 0 的两个特征可以看出，当前词为“大庆”不仅对地名类别有贡献，对人名结束类别也有贡献，也就是对于当前词为“大庆”特征，其输出的权值不再是唯一的，而多了一个候选。如下图所示：



A:有权值偏置　　B:没有权值偏置

图 4-1 权值偏置的表示

Figure4-1 Diagram of weight bias

然后再利用公式 $p(c \mid h) = \frac{1}{Z(h)} \exp\left(\sum_i \lambda_i f_i(c,h) \right)$，可得到最终的正确分类结果，即“张/nf 大庆/ne”。

特征融合对系统的影响

属性	融合前	融合后
特征类型数量	11	23
参数规模	54,000	130,000
事件规模	78.4M	88M
迭代次数	74	65
训练时间	45（小时）	60（小时）

特征融合对模型性能的影响（人名识别）

性能指标	融合前	融合后
精度	72.13%	79.56%
召回率	68.42%	76.25%
F-measure	70.23%	77.87%

问题分析

- 特征融合后模型改进的原因：
  - 消除了原有系统中存在的中国人名和其他类别名实体重叠的错误
- 特征融合后错分的原因及改进措施：
  - 由外国人名的中文译名（包括中国少数民族姓名）的边界识别错误，以及外国地名和外国人名的重叠错误所产生
  - 人名的定义本身造成的，例如模型识别出的结果“范营长/n”在测试语料中并没有被认定为人名
  - 可以通过把中国人名和外国人名的中文译名分为两类来识别，并在模型中增加外国人名常用字的词典信息，来有效地消除第一种错误

3. 基于最大熵模型的名字起源识别

- 3.1 简介
- 3.2 相关工作
- 3.3 最大熵分类器
- 3.4 特征选择
- 3.5 实验分析
- 3.6 结论

3.1 简介

- Name origin refers to the source language of a name (personal/location names) where it originates from.

English:	Richard-理查德(Li-Cha-De) Hackensack-哈肯萨克(Ha-Ken-Sa-Ke)
Chinese:	Wen JiaBao-温家宝(Wen-Jia-Bao) ShenZhen–深圳(Shen-Zhen)
Japanese:	Matsumoto-松本 (Song-Ben) Hokkaido-北海道(Bei-Hai-Dao)
Korean:	Roh MooHyun-卢武铉(Lu-Wu-Xuan) Taejon-大田(Da-Tian)
Vietnamese:	Phan Van Khai-潘文凯(Pan-Wen-Kai) Hanoi-河内(He-Nei)

简介(cont')

- Name origin recognition is very useful in NLP/Search:
  - provide useful semantic information (regional and language information) for common NLP tasks, such as co-reference resolution and name entity recognition.
  - decide the way of name transliteration.
- Two Tasks:
  - recognize the origins of names written in English (**ENOR**)
  - recognize the origins of names written in Chinese (**CNOR**)
- We propose to use Maximum Entropy model and explore diverse features for the tasks.

3.2 相关工作(1)

- Little attention has been given to the study of **ENOR** and no previous work on **CNOR** is reported so far.
- **ENOR:**
  - Rule-based Methods (phonetic information)
    - Left-to-right syllable segmentation (Kuo and Yang, 2004)
    - But not all languages have a finite set of *discriminative* syllable inventory
  - Statistical Methods (phonetic and orthographic information)
    - N-gram Sum (Qu and Grefenstette, 2004)
    - N-gram Perplexity (Li et al., 2007)
    - Data sparseness issue when N>2 and only using single knowledge
- Ours: MaxEnt-based to combine diverse features

3.2 相关工作(2)

- N-gram Summarization Measure
  - select the origin which could make the test name achieves the highest N-gram summarization among all origins
- Bi-gram Based Perplexity Measure
  - select the origin which could make the test name achieves the lowest perplexity among all origins

$$PP_c = 2^{-\frac{1}{N_c}\sum_{i=1}^{N_c}\log p(c_i|c_{i-1})}$$

3.3 最大熵分类器

- MaxEnt-based Classifier:

$$p(c_i \mid x) = \frac{1}{Z} \prod_{j=1}^{K} \alpha_j^{f_j(c_i, x)}$$

  - $Alpha_j$: feature weight
  - f_j: feature value
  - c_i: name origin
  - x: input name
  - Z: normalization factor

3.4 特征选择

- Diverse features
  - N-gram features (N=1,2,3):
    - Character-based N-gram: <*SM*, *MI*, *IT*, *TH*>
    - To capture phonetic and orthographic information
  - Position specific n-gram features:
    - Position-specific character-based N-gram
    - To distinguish surname and given name
  - Phonetic rule-based features:
    - to indicate whether a name is a sequence of Chinese Mandarin or Cantonese *Pinyin*
  - Other features:
    - # of Chinese characters in a name (for CNOR only)
    - Frequency of n-gram in a name (repeated syllable and phonemes)

特征实例

- Feature example
  - The features of name “Smith”
    - S, M, I, T, H
    - SM, MI, IT, TH
    - SMI, MIT, ITH
    - 0S, 1M, 2I, 3T, 4H
    - 0SM, 1MI, 2IT, 3TH
    - 0SMI, 1MIT, 2ITH
    - S=1, M=1, I=1, T=1, H=1
    - SM=1, MI=1, IT=1, TH=1
    - SMI=1, MIT=1, ITH=1
    - FMan=0, FCan=0

3.5 实验分析

- 1) 实验数据集
- 2) 评价方法
- 3) 评价结果

1) 实验数据集

	Origin	# entries	Romanization System
Latin-Scripted Personal name corpus for **ENOR** (90%+10%)	Eng	88,799	English
	Man	115,879	*Pinyin*
	Can	115,739	*Jyutping*
	Jap	123,239	Hepburn

	Origin	# entries
Personal name corpus written in Chinese characters for **CNOR** (90%+10%)	Eng	37,644
	Chi	29,795
	Jap	33,897

2) 评价方法

- Evaluation Metrics

$$P = \frac{\#\ correctly\ recognized\ entries\ of\ the\ given\ origin}{\#\ entries\ recognized\ as\ the\ given\ origin\ by\ the\ system}$$

$$R = \frac{\#\ correctly\ recognized\ entries\ of\ the\ given\ origin}{\#\ entries\ of\ the\ given\ origin}$$

$$F = \frac{2PR}{P+R}$$

$$Acc = \frac{\#\ all\ correctly\ recognized\ entries}{\#\ all\ entries}$$

3) 评价结果(1)

Features	*Acc*
FUni	85.29
+FBi	96.72
+FTri	97.97
+FPUni	98.16
+FPBi	98.80
+FPTri	98.30
+FFre	98.35
+ FMan + FCan	98.44

- Observations:
  - All individual features are useful
  - Bi-gram feature are the most useful
  - MaxEnt method can integrate the strengths of previous rule-based and statistical methods and easily integrate any diverse other features.

Table 1: Contribution of each feature for **ENOR**

3) 评价结果(2)

Features	Eng	Jap	Man	Can
FMan	-0.357	0.069	0.072	-0.709
FCan	-0.424	-0.062	-0.775	0.066

Table 2: Features (FMan and FCan) weights in ENOR task

- The two features are indicative
- Some Japanese names can be successfully parsed by the Chinese Mandarin *Pinyin* system due to their similar syllable structure. “Ta-na-ka Mi-ho”

3) 评价结果(3)

Features	*Acc*
FUni	96.97
+FBi	96.28
+FLen	97.14
+FPUni	97.77
+FPBi	97.56
FUni +FLen + FPUni	98.10

Table 3: Contribution of each feature for **CNOR**

- Observations:
  - Unigram features are the most informative
  - Bigram features degrade *ACC*. This is largely due to the data sparseness problem: in the test, 1080 out of 2980 names of Chinese origin haven't any bigrams learnt from training data, while 2888 out of 2980 names haven't any learnt trigrams.
  - Name Length feature is also useful

3) 评价结果(4)

Origin	**Trigram SUM**	**Trigram PP**	**MaxEnt**
	F	***F***	***F***
Eng	82.11	95.28	**96.45**
Man	90.65	98.66	**99.60**
Can	91.91	97.89	**99.65**
Jap	90.96	97.24	**97.63**
Overall *Acc* (%)	*89.57*	*97.39*	***98.44***

Table 4: Benchmarking different methods in **ENOR** task

3) 评价结果(5)

Origin	**Trigram SUM**	**Trigram PP**	**MaxEnt**
	F	***F***	***F***
Eng	97.28	97.60	**98.56**
Chi	91.59	91.06	**97.22**
Jap	95.28	94.07	**98.34**
Overall *Acc* (%)	*95.00*	*94.53*	***98.10***

Table 5: Benchmarking different methods in **CNOR** task

3) 评价结果(6)

Methods	# of parameters for ENOR	# of parameters for CNOR	
	Trigram	Unigram	Bi-gram
MaxEnt	124,692	13,496	182,116
PP	16,851	4,045	86,490
SUM	16,851	4,045	86,490

Table 6: Numbers of parameters used in different methods

- MaxEnt can incorporate diverse feature effectively
- PP outperforms SUM although having the same # of parameters

3) 评价结果(7)

	SUM	**PP**	**MaxEnt**
Unigram Features	90.55	97.09	**98.10**
Bigram Features	95.00	94.53	**97.56**

Table 7: Overall accuracy using unigram and bi-gram features in CNOR task

- Bi-gram shows lower *Acc* than unigram except SUM
- Re-examine Table 5 (CNOR), # of Chinese characters used in CNOR of different origins
  - Chi:2319 Jan:1413 Eng:377

3.6 结论

- MaxEnt model together with diverse features can effectively solve the name origin recognition problem
- Data sparseness is the most challenging issue to the task
- Future Work:
  - to study the issue of name origin recognition in context of sentence and use contextual words as additional features
  - to integrate our name origin recognition method with a machine transliteration engine to further improve transliteration performance

4. 小节

- ME模型可用于名实体识别
- ME模型是典型的多知识源融合方法
- 特征模板通常考虑位置、词形、词性、当前词类别等特征
- 原子特征、复合特征
- 局部特征与全局特征相结合
- 基于启发式规则的优化处理
- 识别结果的深入分析
- 不同识别模式的对比研究
- 制约因素：时空复杂性

本章结束

谢　谢！