第34卷 第5期
1997年5月

计算机研究与发展
COMPUTER RESEARCH & DEVELOPMENT

Vol. 34, No. 5
May 1997

利用汉字二元语法关系解决汉语自动分词中的交集型歧义

孙茂松 黄昌宁 邹嘉彦* 陆 方 沈达阳

(清华大学计算机科学与技术系 北京 100084)

*(香港城市大学语言资讯科学研究中心 香港九龙达之路)

摘 要 本文提出了一种利用句内相邻字之间的互信息及t-测试差这两个统计量解决汉语自动分词中交集型歧义切分字段的方法. 汉字二元语法关系(bigram)为相关计算的基础,直接从生语料库中自动习得. 初步的实验结果显示,可以正确处理90.3%的交集字段.

关键词 汉语自动分词,汉字二元语法,互信息,t-测试差

中图法分类号 H08; TP301.6

USING CHARACTER BIGRAM FOR AMBIGUITY RESOLUTION IN CHINESE WORD SEGMENTATION

SUN Mao-Song HUANG Chang-Ning Benjamin K. Tsou*

LU Fang SHEN Da-Yang

(*Department of Computer Science & Technology, Tsinghua University, Beijng* 100084)

*(*Research Centre of Language Information Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong*)

Abstract This paper presents a method of using two kinds of statistical measures, mutual information and difference of t-test of adjacent characters in sentences, to deal with ambiguities in Chinese word segmentation. A Chinese character bigram matrix, derived automatically from the raw corpus, serves as a basis for the related calculations. Preliminary experiments show that 90.3% correct rate is achieved for overlapped ambiguities.

Key words Chinese word segmentation, bigram, mutual information, difference of t-test

Class number H08; TP301.6

收稿日期:1996-01-15;修回日期:1996-11-19. 本项研究得到国家自然科学基金资助. 孙茂松,1962年生,1988年于清华大学获硕士学位,副教授,主要研究领域为计算语言学、中文信息处理. 黄昌宁,1937年生,1961年毕业于清华大学,教授,博士生导师,中国中文信息学会计算语言学专委会主任,主要研究领域为人工智能、计算语言学和机器翻译. 邹嘉彦,香港城市大学语言资讯科学研究中心主任,教授,主要研究领域为语言学和计算语言学.

0 引　言

汉语自动分词问题是制约中文信息处理发展的瓶颈之一. 歧义切分字段又是影响分词系统切分精度的重要因素. 所谓歧义切分字段，系指文句中某个片断可能存在不止一种的切分形式，通常包括交集型与包孕型两个基本类型[1]. 本文仅讨论交集型歧义切分字段. 这是因为:①交集型占全部歧义切分字段的85%以上[2];②本文的方法专门针对交集型的特点而设计.

右边给出几个交集型歧义切分字段的例子.

从 头 年　老 人 们　偏 心 底

在 下 议 院　全 党 代 表　剩 余 劳 动 力

实验表明，利用词频信息构造适当的算法来处理切分歧义，十分有效[3,4]. 此类研究的关键是在词频信息的获取上:

(1) 需要相当规模、预经人工分词的语料作为训练样本. 设 C 为某个汉语语料库，C 的词容量(C 所含词的个数之和)为 N. 任给词 w，令 δ_w 为 w 的样本频率 f_w 与概率 p_w 的相对误差(f_w 为 p_w 关于 C 的估计)，则 δ_w，N，f_w 三者之间的关系可用公式表示[5]:

$$\delta_w=\frac{Z_\rho}{\sqrt{N*f_w}}\quad(Z_\rho\text{为常数，通常取 2})\quad\text{或}\ N=\frac{Z_\rho^2}{f_w*\delta_w^2}.$$

汉语常用词不低于50,000个，若视之为等概率独立事件，则 $f_w=1/50,000$. 如果期望任一词频的相对误差小于10%，即 $\delta_w=0.1$，那么理论上训练样本 C 的容量应该为:

$$N=\frac{2^2}{(1/50000)*0.1^2}=20,000,000\ (\text{词}).$$

(2) 词频对领域有一定的敏感性. 即使是从精心挑选的"平衡语料库"中计算而来，将之应用于不同领域也会产生偏移，导致切分精度下降. 要得到针对某个专门领域、比较翔实可靠的词频信息，又要诉诸人工预分的训练语料，这种代价是无法承担的.

(3) 无论是人工分词还是机器分词人工校正，均在不同程度上依赖于人的语感，而语感是因人而异甚至因时而迁的. 并且大批量数据处理过程中，长时间枯燥单调的工作也会使发生错误的机率渐次增加. 所以，分词质量远不能达到人们期望或者想象的水准.

我们提出了一种新的处理策略：借鉴计算词汇学(computational lexicology)与计算辞典学(computational lexicography)的思想[6]，直接从未经加工的生语料库(raw corpus)出发，通过字的统计信息模拟词频，并据之设计交集型歧义切分字段的分析算法. 由于字的统计信息的获取过程是全自动的，因而避免了人工预分引起的各种弊端，数据的准确性、一致性、方法的简明性、移植性均可保证.

1 基本统计判据

1.1 互信息(mutual information)

定义 1. 对有序汉字串 xy，汉字 x，y 之间的互信息定义为:

$$I(x;y) = \log_2 \frac{p(x,y)}{p(x)p(y)} . \tag{1}$$

其中 $p(x,y)$是 x,y 的邻接同现概率，$p(x)$，$p(y)$分别代表 x,y 的独立概率.

若在字容量为 N 的汉语语料库中，x,y 邻接同现的次数为 $r(x,y)$，x,y 独立出现的次数分别为 $r(x)$，$r(y)$，则式(1)各量可用式(2)估计：

$$p(x,y) = \frac{r(x,y)}{N}, \quad p(x) = \frac{r(x)}{N}, \quad p(y) = \frac{r(y)}{N}. \tag{2}$$

互信息反映了汉字对间结合关系的紧密程度：

(1)$I(x;y) \gg 0$，则 $p(x,y) \gg p(x)p(y)$，此时 x,y 之间具有可信的结合关系，并且 $I(x;y)$值越大，结合程度越强；

(2)$I(x;y) \approx 0$，则 $p(x,y) \approx p(x)p(y)$，此时 x,y 之间的结合关系不明确；

(3) $I(x;y) \ll 0$，则 $p(x,y) \ll p(x)p(y)$. 此时 x,y 之间基本没有结合关系，并且 $I(x;y)$值越小，结合程度越弱.

请观察 1.3 节中的例 1. 其中 $I(x;y)$标识的行为两个相邻汉字之间的互信息值(系根据一个含 2000 万汉字的新闻语料库计算得到).

(1)$I(x;y) \gg 0$ “网球”(7.8)，“巴黎”(7.6)，“今天”(7.0)，“战幕”(4.2)，“西郊”(3.7)

(2)$I(x;y) \approx 0$ “天在”(3.4)，“球公”(3.2)，“公开”(3.2)，“开赛”(2.6)，“赛今”(2.5)，“法国”(2.3)，“在巴”(2.1)，“拉开”(1.9)，“开战”(1.4)，“国网”(1.0)，“郊拉”(−1.3)

(3)$I(x;y) \ll 0$ “黎西”(−3.5)

值得注意的是 $I(x;y) \approx 0$ 这个区域：①某些字对结合成立，应连，如“公开”，“法国”；某些字对的结合不成立，应断，如“天在”，“赛今”；②结合成立的字对间的互信息值可能小于结合不成立的字对间的互信息值，如“天在”(3.4)，应断. “公开”(3.2)，却应连. 甚至“拉开”(1.9)，“开战”(1.4)，也可认为结合成立. 互信息是两个字之间结合力的绝对度量. 但在 $I(x;y) \approx 0$ 附近相当宽的区域内，结合能否实现仅依靠互信息有时难以裁定，必须参照一定的上下文，通过上下文字对之间的比较进一步寻找判据.

1.2 t-测试(t-test)

定义 2. 对有序汉字串 xyz，汉字 y 相对于 x 及 z 的 t-测试定义为：

$$t_{x,z}(y) = \frac{p(z|y) - p(y|x)}{\sqrt{\sigma^2(p(z|y)) + \sigma^2(p(y|x))}} . \tag{3}$$

其中 $p(y|x)$，$p(z|y)$分别是 y 关于 x，z 关于 y 的条件概率，$\sigma^2(p(y|x))$，$\sigma^2(p(z|y))$则代表各自的方差. 式(3)各量可用式(4)、式(5)估计：

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{r(x,y)}{r(x)}, \; p(z|y) = \frac{p(y,z)}{p(y)} = \frac{r(y,z)}{r(y)} . \tag{4}$$

$$\sigma^2(p(y|x)) = \frac{r(x,y)}{r^2(x)}, \quad \sigma^2(p(z|y)) = \frac{r(y,z)}{r^2(y)} . \tag{5}$$

下面对式(5)进行推导：

任给邻接有序字对 $x1x2$，设

(1)事件 E 表示 $x1x2$ 在字容量为 N 的语料库中是否出现. E 只能有两种结果：A(出现)及$\sim A$(不出现)；

(2)R_{x1x2}表示在该语料库中(相当于 N 重贝努利试验)事件 A 发生的次数($R_{x1x2}=0,1,\ldots,N$).

则可以认为随机变量 R_{x1x2}近似服从参数为 N 和 $p(A)$的二项分布，记为 $R_{x1x2}\sim B(N, p(A))$，而

$$p(A)=\frac{r(x1,x2)}{N},\text{所以}\quad R_{x1x2}\sim B(N*\frac{r(x1,x2)}{N})$$

由二项分布的性质，有：

$$\sigma^2(R_{x1x2})=N*p(A)*(1-p(A))=N*\frac{r(x1,x2)}{N}*(1-\frac{r(x1,x2)}{N})$$

$$\approx r(x1,x2)\qquad (r(x1,x2)\ll N).$$

另外，显然有 $\sigma^2(R_{x1x2})=\sigma^2(r(x1,x2))$.

于是 $\sigma^2(p(x2|x1))=\sigma^2(\frac{p(x1,x2)}{p(x1)})=\sigma^2(\frac{r(x1,x2)}{r(x1)})$

$$=\frac{1}{r^2(x1)}*\sigma^2(r(x1,x2))\qquad (r(x1)\text{为常数})$$

$$=\frac{1}{r^2(x1)}*\sigma^2(R_{x1x2})\approx\frac{r(x1,x2)}{r^2(x1)}.$$

从 t-测试的定义，可知：①$t_{x,z}(y)>0$ 时，字 y 有与后继字 z 相连的趋势. 值越大，相连趋势越强；②$t_{x,z}(y)=0$ 时，不反映任何趋势；③$t_{x,z}(y)<0$ 时，字 y 有与前趋字 x 相连的趋势. 值越小，相连趋势越强.

1.3 节例 1 中 $t_{x,z}(y)$标识的行显示了相应的 t-测试值.

t-测试是三个汉字之间结合力的相对度量. 如片断"球公开"：I(球;公)，I(公;开)均为 3.2，前者应断，后者却应连，互信息无法区别. 但 $t_{网,公}$(球)$=-17.6<0$,说明"球"倾向于与"网"相连，$t_{球,开}$(公)$=3.2>0$,说明"公"倾向于与"开"相连. 这意味着"球公"之间倾向于断，而"公开"之间倾向于连. 利用 t-测试的不便之处是，t-测试是挂靠在汉字上的，不似互信息挂靠于两个汉字之间的位置,故有必要将两者的挂靠对象统一起来.

1.3 t-测试差

定义 3. 对有序汉字串 $vxyw$，汉字 x,y 之间的 t-测试差为：

$$\Delta t(x;y)=t_{v,y}(x)-t_{x,w}(y).$$

可以分几种情形讨论：

$t_{v,y}(x)>0$，$t_{x,w}(y)<0$. 此时 x,y 之间相互吸引，必有 $\Delta t(x;y)>0$，x,y 之间倾向于连，且趋势比单独使用 $t_{v,y}(x)$或 $t_{x,w}(y)$更显得突出.

$t_{v,y}(x)<0$，$t_{x,w}(y)>0$. 此时 x,y 相互排斥，必有 $\Delta t(x;y)<0$，x,y 之间倾向于断.

$t_{v,y}(x)>0$，$t_{x,w}(y)>0$. 此时 y 吸引 x,同时 w 吸引 y，产生"竞争"：

$\Delta t(x;y)>0$ 倾向于连；$\Delta t(x;y)<0$ 倾向于断.

$t_{v,y}(x)<0$，$t_{x,w}(y)<0$. 此时 x 吸引 y，同时 v 吸引 x，产生“竞争”；

$\Delta t(x:y)>0$ 倾向于连；$\Delta t(x:y)<0$ 倾向于断.

例 1.

	法	国	网	球	公	开	赛	今
$I(x:y)$	2.3	1.0	7.8	3.2	3.2	2.6	2.5	7.0
$t_{x,x}(y)$	43.7	−43.8	20.0	−17.6	3.2	−15.8	5.3	96.4
$\Delta t(x:y)$	87.5	−63.8	37.6	−20.8	19.0	−21.1	−91.1	165.6

	天	在	巴	黎	西	郊	拉	开
$I(x:y)$	3.4	2.1	7.6	−3.5	3.7	−1.3	1.9	1.4
$t_{x,x}(y)$	−69.2	−53.4	24.3	−32.2	0.4	−2.7	7.3	−13.5
$\Delta t(x:y)$	−15.8	−77.7	56.5	−32.6	3.1	−10.0	20.8	−16.2

	战	幕
$I(x:y)$	4.2	
$t_{x,x}(y)$	2.7	−10.4
$\Delta t(x:y)$	13.1	

其中，Δt(网:球)属情形(1)，Δt(球:公)属情形(2)，Δt(赛:今)则属情形(3). 请继续注意片断“球公开”：$|\Delta t(球:公)-\Delta t(公:开)|=39.8$，较之 $|t_{网,公}(球)-t_{球,开}(公)|=20.8$，差异进一步拉开.

1.4 互信息与 t-测试差的关系

对任意邻接有序字对 xy，$I(x:y)$反映了 x,y 之间的静态结合能力，$\Delta t(x:y)$则动态考虑了 $vxyw$ 四个字的耦合影响.

例 2.

	重	点	工	程	威	墅	堰	电
$I(x:y)$	5.9	1.7	6.3	3.1	11.9	12.3	0.5	3.1
$\Delta t(x:y)$	87.3	−78.8	103.3	−63.9	0.3	2.1	−1.5	11.7

	厂	第	二	期	扩	建	工	程
$I(x:y)$	−0.3	5.9	2.3	2.4	4.7	0.1	6.3	2.6
$\Delta t(x:y)$	−63.2	101.2	−41.0	−21.8	26.1	−68.9	97.5	−41.4

	已	完	成
$I(x:y)$	2.1	4.7	
$\Delta t(x:y)$	−44.1	97.3	

“工程”在例 2 中出现两次，I(工:程)恒定(6.3)，Δt(工:程)依上下文不同有所变化(一处为 103.3，另一处为 97.5).

综上所述，互信息和 t-测试差各有特点，且具一定程度的互补性. 把它们结合起来，可以形成更趋合理的统计判据. 由于计算 $I(x:y)$，$\Delta t(x:y)$均离不开 $r(x,y)$，故本质上这是一个关于汉字的一阶 Markov 模型，亦即关于汉字的二元语法(bigram)模型.

2 利用互信息和 t-测试差处理交集型歧义切分

2.1 设计原理

一般地，交集型歧义切分字段 JS：$a_1\cdots a_ib_1\cdots b_mc_1\cdots c_n(i>0, m>0, n>0)$ 存在两种切分方案：

$SEG1$：$\underbrace{a_1 \cdots a_i b_1 \cdots b_m}_{w11}$ ↑$pt1$ $\underbrace{c_1 \cdots c_n}_{w12}$　　$SEG2$：$\underbrace{a_1 \cdots a_i}_{w21}$ ↑$pt2$ $\underbrace{b_1 \cdots b_m c_1 \cdots c_n}_{w22}$

其中 $w11, w12, w21, w22$ 均为词，$pt1, pt2$ 分别对应 $b_m c_1$ 与 $a_i b_1$ 之间的位置. 对词频法，易作出判断：

如果 $p(w11) * p(w12) > p(w21) * p(w22)$，则肯定 $SEG1$，否则肯定 $SEG2$.

我们不妨转换一下角度：歧义字段 JS 有两个可能断点 $pt1$ 和 $pt2$，即位置 $b_m c_1$ 和 $a_i b_1$（两断点必有一个出现，但不能同时出现），前者对应 $SEG1$，后者对应 $SEG2$. 最终哪一个断点成立，可以认为是字串 $a_1 \cdots a_i b_1 \cdots b_m c_1 \cdots c_n$（而不再是词 $w11$，$w12$，$w21$，$w22$）共同作用的结果. 注意力自然凝聚到 $I(b_m;c_1)$，$\Delta t(b_m;c_1)$ 及 $I(a_i;b_1)$，$\Delta t(a_i;b_1)$ 上. 由于绝大多数情况下，有 $m \leqslant 2$，$i+n \leqslant 4$，故这四个参量的视域基本覆盖了整个字串 JS. 换言之，可以认为交集型字段 JS 的排歧是上述四个参量的函数：

$$Disambi(JS) = F(I(b_m;c_1), \Delta t(b_m;c_1), I(a_i;b_1), \Delta t(a_i;b_1))$$

函数 F 应适应参量的变化：

(1) I 及 Δt 的判断一致

($I(pt1) > I(pt2)$ 且 $\Delta t(pt1) > \Delta t(pt2)$) 或

($I(pt1) < I(pt2)$ 且 $\Delta t(pt1) < \Delta t(pt2)$).

例 3.（加洛斯）　网　球　场　（拉开战幕）

$I(x;y)$　　7.8　3.5

$\Delta t(x;y)$　　52.8　−9.0　　⇒　　网球　场

例 4.　（今年）　共　生　产　（汽车八万辆）

$I(x;y)$　　−0.9　6.6

$\Delta t(x;y)$　　−142.4　232.8　　⇒　　共　生产

(2) I 及 Δt 的判断不一致

①$|I(pt1)-I(pt2)| \geqslant \alpha$. 此时由 I 值判断.

例 5.　（新机制使它）　进　发　出　（旺盛活力）

$I(x;y)$　　7.0　2.3

$\Delta t(x;y)$　　2.0　2.3　　⇒　进发　出

②$|I(pt1)-I(pt2)| < \alpha$ 但 $|\Delta t(pt1)-\Delta t(pt2)| \geqslant \beta$. 此时主要由 Δt 值判断.

例 6.　（对社会对自己都）　有　益　处

$I(x;y)$　　2.8　2.1

$\Delta t(x;y)$　　−2.9　2.4　　⇒　有　益处

虽然 I(益;处)$<I$(有;益)，但仍取"益处"连.

例 7.　（第一次）　全　党　代　表　（大会）

$I(x;y)$　　1.6　1.9

$\Delta t(x;y)$　　−13.1　−114.0　　⇒　全党　代表

虽然 Δt(全;党)，Δt(党;代)均<0（倾向于断），但程度上的差异使得比较仍有意义.

③$|I(pt1)-I(pt2)| < \alpha$ 且 $|\Delta t(pt1)-\Delta t(pt2)| < \beta$.

例 8. 此时退回来，仍由 I 值判断.

（14 次登） 上 海 拔 （7534 米的…）

$I(x;y)$ 6.1 7.2

$\Delta t(x;y)$ 24.1 24.0 ⇒ 上 海拔

2.2 算法实现

算法的基本流程为：

(1) 利用词典进行正向及反向最大匹配分词

(2) 如果对字段 JS，正、反向给出两种切分方案 $SEG1$ 和 $SEG2$，则：

①如果两种方案切分段数不同，则选择切分段数少的方案作为结果

②如果两种方案切分断数相同，则计算可能断点 $pt1$，$pt2$ 处的 I 值及 Δt 值

- 先利用 I 值判断. 如果 $I(pt2) - I(pt1) \geqslant \alpha$ 则肯定 $SEG1$；如果 $I(pt1) - I(pt2) \geqslant \alpha$ 则肯定 $SEG2$；
- 如果 $|I(pt1) - I(pt2)| < \alpha$，再利用 Δt 值判断. 如果 $\Delta t(pt2) - \Delta t(pt1) \geqslant \beta$ 则肯定 $SEG1$；如果 $\Delta t(pt1) - \Delta t(pt2) \geqslant \beta$ 则肯定 $SEG2$；
- 如果 $|I(pt1) - I(pt2)| < \alpha$ 且 $|\Delta t(pt1) - \Delta t(pt2)| < \beta$，退回来利用 I 值作最后判断. 如果 $I(pt2) > I(pt1)$ 则肯定 $SEG1$，否则肯定 $SEG2$.

常数 α，β 由实验测定.

除词典(常用词 6 万余条)外，支持本算法的另一个重要资源是任意两个相邻汉字的同现概率矩阵(bigram 矩阵). 二级汉字共 6775 个，故 bigram 矩阵单元数为 6775 * 6775 ≈45.6M. 若每单元以二字节计，需 91.2M 磁盘空间. 这种开销显然难以忍受.

bigram 矩阵的训练样本为包含 2000 万(20M)汉字的新闻语料库. 即便如此，矩阵中每个单元容纳的平均字对数只有 20M/45.6M=0.44 个，尚不足 1. 可见 bigram 矩阵中必含大量的零单元，即存在所谓数据稀疏问题.

bigram 矩阵中非零单元比例依训练样本规模的变化规律显示：统计伊始，非零单元增长很快(0—600 万字区间，平均增长速度约 0.28%/百万字)，以后渐次放慢(600—1400 万字区间，平均增长速度约 0.13%/百万字)，而在统计末期，更趋平缓 (1400—2000 万字区间，平均增长速度约 0.08%/百万字)，训练结束时非零单元仅占矩阵单元总数的 3.3%. 利用此特点，我们设计了矩阵的压缩存储机制，压缩后需空间<2MB，仅为完整 bigram 矩阵(91.2M)的 2%.

此外，bigram 矩阵中的零单元会导致计算互信息时对数运算失败. 故需进行数据平滑(smoothing)处理：令 r 是统计实际所得的次数，r^* 是经平滑调整后的次数，则

$$r^* = \frac{(r+1)*N}{N+S}.$$

其中 N 为语料库字数，S 为汉字集个数(S=6775). 平滑后满足：

$$\sum_1^N r^*/N = 1, \quad \text{即} \quad \sum p = 1.$$

3.3 实验结果及其分析

我们从新闻语料库中随机抽取了 12.5 万字的文本进行测试. 实验结果如下：

交集型歧义切分字段数：	504
正确切分字段数：	455
错误切分字段数：	49

对交集型歧义切分字段的切分正确率：　90.3%

对这504个交集型歧义字段，若单纯用正向最大匹配法，切分正确率为35.7%；单纯用反向最大匹配法，切分正确率为64.3%．与之相比较，本算法将切分正确率分别提高了54.6%和26.0%．

通过实验结果的分析发现，错误往往发生于某些不常见的用法．如：

例9.（打着金星）　化　　工　　厂　（招牌）

$I(x;y)$	3.3	4.8		
$\Delta t(x;y)$	10.7	47.1	⇒	化　工厂

例10.（调整广东省潮汕地区的）　行　　政　　区　　划

$I(x;y)$	4.9	1.7		
$\Delta t(x;y)$	50.1	−35.9	⇒	行政区　划

例9、例10中，“工厂”、“行政区”的使用频率远远大于“化工”、“区划”的使用频率．即使应用词频法（实际上是词的一元语法模型unigram），此类错误也不能避免．这是由词的unigram模型或字的bigram模型的固有缺陷所决定的．一个可能的改进是诉诸词的bigram模型，不过必须得到足够规模的、经正确分词的语料的支持，代价无疑会十分高昂．

3　结束语

本文讨论了基于汉字互信息及t-测试差（汉字的bigram模型）解决汉语自动分词中交集型歧义的一般方法．方法的最大特点是知识获取（统计）过程的完全自动化．实验显示，此方法具有与采用词频信息大致相同的处理能力，可以满足较高程度的分词需要．

我们进一步考察了同一bigram矩阵对于不同领域语料的适应性．bigram矩阵由新闻语料库训练，分词系统据此对经济类语料进行切分．交集型歧义切分字段的切分正确率约为81%．这说明：一方面，领域不同的文本由于内容、风格、体裁的不同，对bigram矩阵的表现会产生一定的影响；但另一方面，也应该看到，利用领域A的语料训练系统参数，据之切分领域B的文本，效果仍明显好于单纯的最大匹配法．汉语在遣词造句时毕竟有一些最基本的用法、习惯或规律（字词一级如常用字、常用词），相对稳定，并不因领域或文体而发生太大的变化．于是导致了bigram矩阵一条可贵的性质：对领域的敏感度有限．

参　考　文　献

1　Lai B Y, Lun S, Sun C F, Sun M S. A conditional maximal matching segmentation algorithm using mainly tags for resolution of ambiguities for Chinese texts. In: Proc of ROCLING-IV, Kenting, Taiwan, 1991

2　梁南元．书面汉语自动分词系统——CDWS．中文信息学报，1987(1)

3　王晓龙、王开铸，李仲荣等．最少分词问题及其解法．科学通报，1989，13

4　Zhang J S, Chen Z D, Chen S D. A method of word identification for Chinese by constraint satisfaction and statistical optimization techniques. In: Proc of ROCLING-IV, Kenting, Taiwan, 1991

5　冯志伟．数理语言学．上海：知识出版社，1985

6　Church K W, Hanks P, Hindle D. Using statistics in lexical analysis. In: Zernik ed. Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale, N J: Erlbaum, 1991.115—164