

1. 最大匹配法 (Maximum Matching, MM) 词典切片, 乱排切片.

{	正句	From
	逆句	But
	22句	But

for: $s = 66 \dots 69$

求切: $w_L = G_L - C_L$ m 为 零木子数.
 零比同的

简单易懂。无而因法才觉得
正确率 85%，收又为饼的能力差。

$m = 7, 6, 5, \dots$ 按↑去比时.

算法描述:

c1) 令 $i=0$, 将指针 PL 指向输入字符串的末尾

4. 计算 P_2 列中未读的个数 n , $n=1$, 转 (4).

例 2 $m = 1$ 值 ~~值~~

3) 从 P_i 选取 m 个因子作为候选, 判断:

a). 此是词表中单词, 在此后添加一个切分标志, 如 (c).

a). 以是词典中的词且以的长度大于1. 将以从右端去掉一个字, 转
b). 以不是词典中的词且以的长度大于1. 将以从右端去掉一个字, 转

a) 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, 700, 701, 702, 703, 704, 705, 706, 707, 708, 709, 710, 711, 712, 713, 714, 715, 716, 717, 718, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 733, 734, 735, 736, 737, 738, 739, 740, 741, 742, 743, 744, 745, 746, 747, 748, 749, 750, 751, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 798, 799, 800, 801, 802, 803, 804, 805, 806, 807, 808, 809, 810, 811, 812, 813, 814, 815, 816, 817, 818, 819, 820, 821, 822, 823, 824, 825, 826, 827, 828, 829, 830, 831, 832, 833, 834, 835, 836, 837, 838, 839, 840, 841, 842,

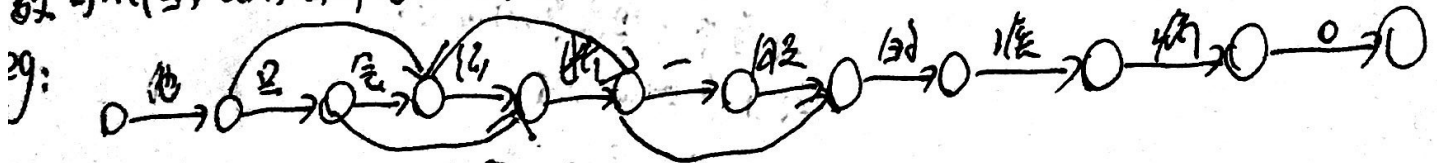
4. 根据 w_2 的长度修改 P_2 位置, $i = |x|$, 转 4).

41. 蛭类百果

2. 最少分词法(最脆弱法) (词数最少的和为15果)

设特别字符串 $S = a_1 a_2 \dots a_n$, 其中 a_i 为单字符, n 为串的长度, 建立一个节点

设有向图 G ，各节点编号依次为 $v_0, v_1, v_2, \dots, v_n$



输出使选中元件同改型少的

在描述: 1. 相邻点 v_{i-1}, v_i 之间建立有边 $\langle v_{i-1}, v_i \rangle$, 边对应的权重为

$$C_n(C^{\infty}, 2, \dots, n)$$

4. $w = C_1 C_2 \dots C_j$

$$v_{i-1}, v_j \text{ 相互有向边 } \langle v_{i-1}, v_j \rangle$$

-31. repeat ②. 直至无新出现.

41. 蜀山志

4. 仇: 倡言忠孝不孝
符合自然规律

缺: 歧义
长度相同的字符串增加



第7章 词法分析器与词性标注

词：NLP的基本单位
 自动词法分析：分析形态等
 词性/词类

词析器：形态分析
 分析器：分词
 标注器：分词+形态还原

7.2 英语的形态分析

7.2.1 单词识别 → 自动机随着样本数增大，逐渐不再重要。

7.2.2 形态还原 → 词典，按规则还原

7.3 汉语自动分词

问题 { 单字词与词素 } → 汉语分词规范
 词与短语
 词素型歧义 } 歧义切片 链长：共享字
 组合型歧义 } 12.1 (根本)
 人为标注与事实 } 卡诺示词的识别
 新出现的词汇 }

原则 { 合并原则：语义上不问相加而得
 无法合并得到（不符合组合规律）

辅助原则 { 切片原则：有明显片断词
 (线性) 合并 { 附音符号
 频率高的字串
 双音节+单音节的偏正式
 双音节的偏正式

切片 { 内部复合，合并冗长

7.4 分词与词性标注结果评价指标

正确率 (P) : $P = \frac{n}{N} \times 100\%$

召回率 (R) : $R = \frac{n}{n'} \times 100\%$

R_{out} R_{in}
 集外词 集内词

F-测度值: $F-measure = \frac{(R_{in} + 1) \times P \times R}{R_{in} \times P + R} \times 100\%$: 平均

n: 正确结果

n': 系统输出

R_{in} : 标注合集的召回



第八章 语法理论

8.1 功能句-文法 (FUG)

提出起因: Chomsky 语法结构语法生成能力太强, 产生许多不合语法/有歧义的句子

改进: 单一形式的结构模式 | 句-运算

复杂特征集的定义: 设 α 为一个功能描述 FD, 当且仅当 α 可以表示为:

$$\left(\begin{array}{l} t_1 = v_1 \\ t_2 = v_2 \\ \vdots \\ t_n = v_n \end{array} \right)^{n \geq 1} \quad \begin{array}{l} t_i = \text{特征名} \\ v_i = \text{特征值} \end{array}$$

$\alpha(t) = v_i \quad (i=1, \dots, n)$
可以用复杂特征集描述句法

- 词法
 - $\text{cat} = n$
 - $\text{sent} = \text{equp ment}$
 - $\text{lex} = \text{什 么 机}$
- 短语
 - $S \rightarrow NP + Verb$
 - $\text{cat} = S$
- 句子
 - $\text{cat} = S$
 - $\text{subject} =$

特点: 允许利用多个特征描述同一个语言单位。

嵌套、层次、运算方便。

句-运算:

若 α, β 均为复杂特征集, 则 α, β 是相容的, 当且仅当:

1) 如果 $\alpha(t) = a, \beta(t) = b$, 且 a, b 都是原子的, 那么 α, β 是相容的。
当且仅当 $a = b$

2) $\alpha(t), \beta(t)$ 相容。



递归定义: 1) a, b 均为原子, 如果 $a=b$, 则 $a \sqcup b = a$, 否则 $a \sqcup b = \emptyset$

2) a, b 均为复合项, 则

1) 若 $\alpha(t) = \nu$, 但 $\beta(t)$ 的值未定义, 则 $t = \nu$ 属于 $\alpha \sqcup \beta$

2) 反之亦然.

3) $\alpha(t) = \nu_1, \beta(t) = \nu_2, \nu_1$ 与 ν_2 相左 (不相兼容).

则 $t = (\nu_1 \sqcup \nu_2)$ 属于 $\alpha \sqcup \beta$, 否则为 \emptyset .

8.2 句法功能语法 (LFG)

成分结构层次 (上下文无关文法)
功能结构层次 (句法树)
句法树
语义树
子树结构
上述三种关系.

$s \rightarrow NP \quad VP$
 $(PSUBJ) = \downarrow \quad T = \downarrow$
终结

$\uparrow \downarrow$: 自指支配元素.

由 $c \rightarrow t$ 代换

注: 一个函数只能有一个值.

8.4 树连接语法 (TAG)

$G = \langle V_n, V_t, S, T_n, T_a \rangle$

替换
附加
整体 \rightarrow 左半
else: 右半

8.5 广义的句法功能语法

GPSG.

句法树

特征结构的树

树升结构

语义解释

语义解释树.

8.3 GB理论 (管轄约束理论)

规则系统
句法规则
语类和轻根规则
语素规则
语义规则

句法树

大理论

句法理论

句法理论

句法理论

句法理论

句法理论

句法理论

句法理论

\rightarrow 可移位

