最大匹配法

匹配：分词过程中用文本中的候选词去跟词表中的词匹配，匹配成功就认为候选词是词

最大匹配：尽可能用最长的词来匹配句子中的汉字串

切出来的词尽可能长，词尽可能少

“时间就是生命”

步骤	s1	s2	w
1	时间就是生命	null	时间就是生命
2	时间就是生命	null	时间就是生
3	时间就是生命	null	时间就是
4	时间就是生命	null	时间就
5	时间就是生命	null	时间
6	就是生命	时间/	
7	就是生命	时间/	就是生命
8	就是生命	时间/	就是生
9	就是生命	时间/	就是
10	就是生命	时间/	就
11	是生命	时间/就/	
12	是生命	时间/就/	是生命
13	是生命	时间/就/	是生
14	是生命	时间/就/	是
15	生命	时间/就/是/	
16	生命	时间/就/是/	生命
17	null	时间/就/是/生命/	

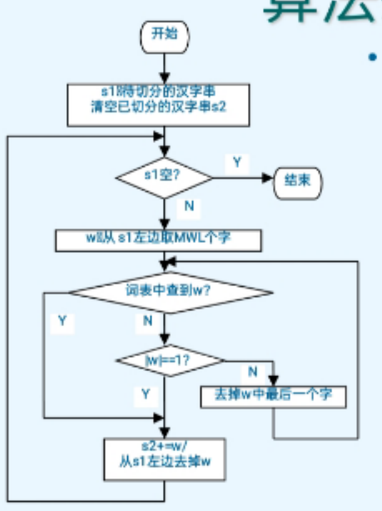
s1：待分词字符串

s2：当前切分出的词（结果）

w：当前匹配到的候选词

算法描述

- 两重循环
  - 外循环—切词
    - 从输入串s1中复制候选词w
    - w尽可能长，是为了进行“最大匹配”
    - 每次切掉s1左边的若干个字
  - 内循环—匹配
    - 用w去匹配词表中的词
    - 最好的情况是一次匹配成功
    - 最坏的情况是每次匹配均不成功，直到剩下一个字



优点：简单，仅需很少的语言资源

缺点：消解切分歧义的能力差、切分正确率低（95%）

逆向扫描

每次从汉字串右边取一个候选词，候选词不止一个汉字而且在词表中查不到，则将它最前面的汉字去掉

“使用户满意”

步骤	s1	s2	w
1	使用户满意	null	用户满意
2	使用户满意	null	户满意
3	使用户满意	null	满意
4	使用户	/满意	
5	使用户	/满意	使用户
6	使用户	/满意	用户
7	使	/用户/满意	
8	使	/用户/满意	使
9	null	/使/用户/满意	

注：最大匹配法是一种"正向扫描"，在这道题中，正向扫描会这么划分：使用/户/满意

源句子	正向扫描	逆向扫描
使用户满意	使用 / 户 / 满意×	使 / 用户 / 满意√
需求和规格说明	需求 / 和 / 规格 / 说明√	需 / 求和 / 规格 / 说明×

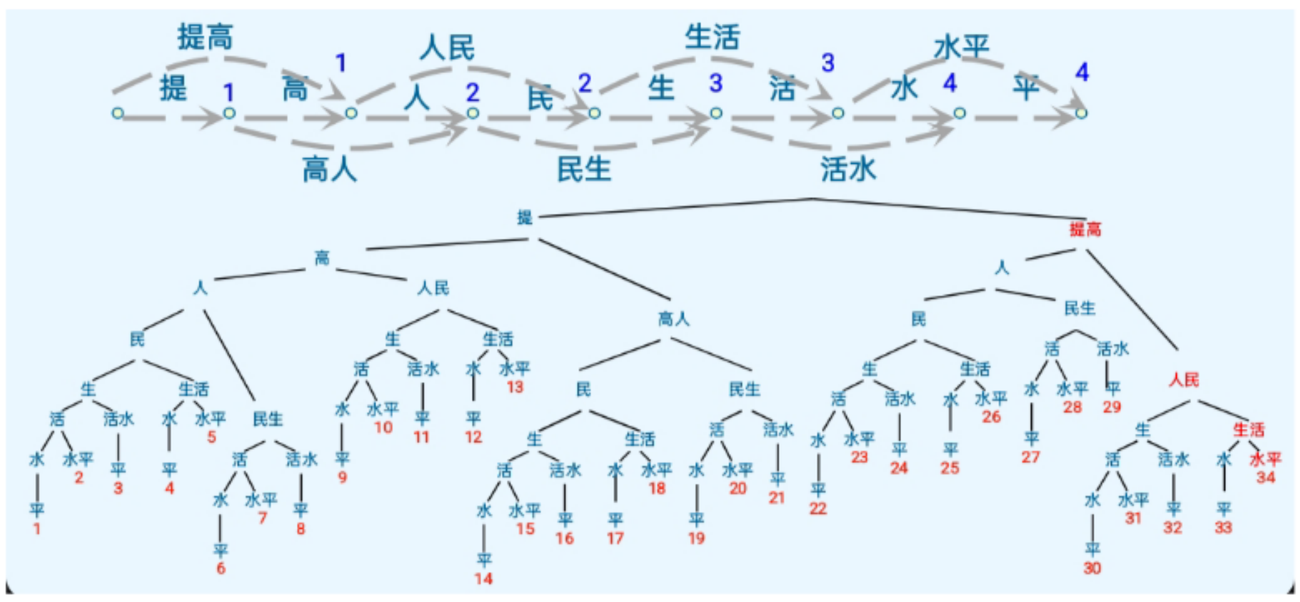
歧义切分：若对一个汉字串进行正、逆两次扫描而切分结果不同，便认为有切分歧义。但是依旧存在问题：双同最大匹配法存在切分歧义检测盲区

如：结合成分才时

切分为：结合/成分/才时

最少分词法

等价于在有向图中搜索最短路径，即使分词结果中含词数最少



存在问题：含词数最少的切分方式可能不止一种

"他说的确实在理"

他/说/的确/实在/理

他/说/的确/实/在理

他/说/的/确实/在理

这时需要其它方式辅助选择

最大概率法

以概率最大的那个词串W作为汉字串Z的分词结果，即已知输出Z，反推概率最大的分词为法W

$$\hat{W} = \arg\max_W P(W|Z) = \arg\max_W \frac{P(W)P(Z|W)}{P(Z)}$$

由于W本就由Z切分得到，所以 $P(Z|W)=1$

而P(Z)无论对哪个W，数值都不会变，不影响 $\arg\max_W$ 结果

所以也约掉

$$\hat{W} = \arg\max_W P(W)$$

现在我们来考虑如何计算P(W)，这里有两种假设，分别对应两种模型

(一) 一阶马尔可夫假设（二元语法模型）

$$P(W) \approx \prod_{i=1}^{m} P(w_i|w_{i-1})$$

w_i：第i个词　　w_0：虚设的句首词，如<BOS>

(二) 独立性假设（一元语法）

$$P(W) = \prod_{i=1}^{m} P(w_i)$$

最大概率法步骤：①根据词表找出Z中所有词

②找出所有可能的切分路径(W)

③找出 $\arg\max_W P(W)$ 作为Z的切分

注：词表是不可能完备的，为了保证至少有一种切分

①把每个汉字都作为候选词，

②若切分里有未登录词就给它一个很小的概率

问题：当Z过长，每种切分结果的P(W)都会趋近于0

碍于机器精度会直接显示为0，导致无法比较

方法：用对数将 $\prod P(W_i)$ 化为 $\sum P(W_i)$

定义 $Fee(W) = -\log P(W)$　　Fee称为费用

便有 $Fee(W) = \sum_{i=1}^{n} -\log P(W_i)$

P越高 Fee越小，于是找"P(W)最大的词串"

变为了找"Fee(W)最小的词串"

最大概率法性能分析
- 优：可以发现所有切分歧义
- 缺：依赖词典统计概率与决策算法
  需要大量标注语料

词性标注法

- 例："他从小学高尔夫球。"

他 | 从小 | 学 | 高尔夫球 | 。
代词 副词 动词 名词 句号

他 | 从 | 小学 | 高尔夫球 | 。
代词 介词 名词 名词 句号

P(代词 + 副词 + 动词 + 名词+句号) >> P(代词 + 介词 + 名词 + 名词+句号)

将自动分词和基于Markov链的词性自动标注技术结合起来，利用从人工标注语料库中提取出的词性二元统计规律来消解切分歧义

将分词和词类标注结合起来，利用丰富的词类信息对分词决策提供帮助，并且在标注过程中又反过来对分词结果进行检验、调整，从而极大地提高切分的准确率

基于互现信息的分词方法

词是稳定的字的组合，因此，相邻的字同时出现的次数越多，就越可能构成一个词

$$M(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

x、y：相邻的两个字

P(x)：字x单独出现的概率（边缘概率）

P(x,y)：字x与y相邻出现的概率（联合概率）

$\frac{P(x,y)}{P(x)P(y)}$
- >1　x,y实际共现比随机共现稠密，成词
- $=1$　x,y完全独立，不成词
- <1　x,y实际共现比随机共现更稀疏，不可能成词

优点：无需词典

缺点：共现频度高但并不是词的字组并不罕见，如"之一""这一""我的"

对常用词的识别精度差

时空开销大

基于HMM的分词方法

将分词看作字的状态估计问题，每个字有4种词位状态

①词首(B) ②词中(M) ③词尾(E) ④单独成词(S)

如：小明/毕业于/西工大

B E B M E B M E

观察序列O：小明毕业于西工大

状态序列Q：BEBMEBME

初始状态概率同量元：词头的第一个字属于{B,M,E,S}四种状态的概率

状态转移矩阵A：前一个字状态为B，后一个字属于{B,M,E,S}四种状态的概率

观测概率矩阵B：某一位置处于某个状态下，出现某个汉字的概率

于是问题为已知观察序列O，求概率最大的状态序列Q，即预测问题

Viterbi算法复习：

①初始化：$\delta_1(i) = \pi_i b_i(O_1)$　　$\psi_1(i) = 0$

②递归计算：$\delta_t(i) = \max_j [\delta_{t-1}(j) a_{ji}] \times b_i(O_t)$

$\psi_t(i) = \arg\max_j [\delta_{t-1}(j) a_{ji}]$

③路径回溯：$\hat{q}_T = \arg\max_i \delta_T(i)$

$$\hat{q}_{t-1} = \psi_t(\hat{q}_t)$$

- 维特比(Viterbi)算法求解步骤

1.初始化及递归计算

算法得到两个矩阵，δ 和 ψ:

		小	明	硕	士	毕	业	于	西	北	工	业	大	学
		0	1	2	3	4	5	6	7	8	9	10	11	12
B	0													
E	1													
M	2													
S	3													

概率数组δ[4][13] 4是状态数(0:B,1:E,2:M,3:S)，13是输入句子的字数

比如 δ[0][2] 代表 已知观察字序列的条件下，‘硕’字推断为状态B的最大概率。

路径数组ψ[4][13] 4是状态数(0:B,1:E,2:M,3:S)，13是输入句子的字数

比如 ψ[0][2] 代表 δ[0][2]取到最大时，前一个字的状态，比如 ψ[0][2] = 1, 则代表 δ[0][2]取到最大时，前一个字(也就是明)的状态是E。

2.边界条件和路径回溯

边界条件：最后一个字的状态只可能是 E 或者 S，不可能是 M 或者 B

3.回溯，得到序列

EMMMMBEMBEBEB

4.倒序

BEBEBMEBMMMME

5.切词

BE/BE/BME/BMMMME

小明/硕士/毕业于/西北工业大学

还有基于字分类的分词方法、基于实例的分词方法等

分词规范

- 空格与标点符号是计算机中单位的分隔标记
- 二字或三字词，以及结合紧密，使用稳定的二字、三字词组一律为分词单位
- 四字成语、四字词以及结合紧密、使用稳定的四字词组一律为分词单位
- 五字和五字以上的俗语等，只要切分后不影响本义，应予切分

- 举例
  - 时间名词或词组的分词规则
    - 一年的十二个月份以及每周的七天，一律为分词单位。
      - 五月 元月 3月 星期日 礼拜三
    - “年、日、时、分、秒”分别为分词单位。
      - 1988/年 15/日 11/时/42/分/8/秒
    - “前、后、上、下、大前、大后”等直接与时间名词或量词组合时，它们为一个分词单位。
      - 前天 后年 上星期 下月 大前天 大后年
    - “初”加十以内的数字一律为分词单位。
      - 初一 初二

在实际应用中，不同应用规范略有不同

歧义切分

一、交集型歧义切分

Z = ABC，AB和BC都是词，便形成交集型歧义切分：

AB/C or A/BC

如："使用/户"与"使/用户"

二、组合型歧义切分

它含至少两个汉字的字符串，它们组合起来是词，拆分开来也是词

如："马上"与"马/上"　　"个人"与"个/人"

三、混合型歧义切分

在交集型字段里包含组合型字段

未登录词

①专有名词　②实体名词

③衍生词：叠词：打牌/打打牌　高兴/高高兴兴

一个/一个个

派生词：非党员　成功者　看见/看得见

相信/相不相信

离合词：打架/打了一架

④新词：蓝紫、内卷、新冠

分词主要评测指标

精确率：$P=\frac{系统输出正确词个数}{系统输出词的个数}\times 100\%$

召回率：$R=\frac{系统输出正确词个数}{标准答案中词个数}\times 100\%$

精确率与召回率属于此消彼长的关系，P个就要少切词，切长词，容易漏切导致R↓，R个就要多切词，切短词，容易错切导致P↓

F-测度：$F=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}\ \beta=1\ =\frac{2PR}{P+R}$

β：权重系数

β默认为1，表示P、R同等重要

β>1 R更重要 β<1 P更重要

词性标注

由于汉语不像西方语言那样不同词性的词形态不同，且汉语常用词兼类现象严重，这导致词性歧义问题在汉语中十分突出

法一、词性标注集

法二、基于规则的词性标注法

按照兼类词搭配关系和上下文语境建立词性消歧规则

① 使用一部词典给每个单词指派一个潜在词性表

② 使用歧义消解规则表筛选原来的潜在词性表，使每个单词得到一个单独的词性标记

通用规则：只使用词类或短语信息

个性规则：直接使用词汇信息，针对特定词写的规则

法三、基于统计的词性标注法

① 基于一阶马尔可夫模型的词性标注

即当前词的词性只依赖上一个词

$$P(T)=\prod_{i=1}^{l} P(t_i \mid t_{i-1}) \qquad T=t_1 \cdots t_l \text{ 词性标注串}$$

② 基于HMM的词性标注：

观测序列W：词串　　隐状态序列T：词性标注串

$$T=\arg\max_T P(T \mid W)=\arg\max_T \frac{P(T)P(W \mid T)}{P(W)}$$

$$=\arg\max_T P(T)P(W \mid T)$$

$$P(T)=P(t_1 \cdots t_n)$$

$$=\prod_{i=1}^{n} P(t_i \mid t_{i-1})$$

$$=\pi_{t_1} a_{t_1 t_2} a_{t_2 t_3} \cdots a_{t_{n-1} t_n}$$

$$P(W \mid T)=P(w_1 w_2 \cdots w_n \mid T)$$

$$\overset{\text{独立性假设}}{=\!=} P(w_1 \mid T) \cdots P(w_n \mid T)$$

$$=\prod_{i=1}^{n} P(w_i \mid T)$$

$$=\prod_{i=1}^{n} P(w_i \mid t_i)$$

$$=\prod_{i=1}^{n} b_{t_i}(w_i)$$

③ 分词与词性标注一体化模型

$$W,T=\arg\max_{W,T} P(WT \mid Z)$$

$$=\arg\max_{W,T} \frac{P(WTZ)\ P(Z \mid WT)}{P(Z)}$$

$$=\arg\max_{W,T} P(WTZ)$$

$$= \operatorname{argmax}_{W,T} P(WT)$$

$$= \operatorname{argmax}_{W,T} P(T)\,P(W|T)$$