# 歧义字段的最大概率切分算法

刘挺　王开铸　姜兴海
（哈尔滨工业大学计算机系　150001）
邵艳秋
（黑龙江交通高等专科学校　150050）

**摘要**：本文提出了一种汉语歧义字段的切分算法，该算法根据词频估算每一种切分形式出现的概率，并以概率最大的切分形式作为分词结果。

**关键词**：计算语言学　汉语自动分词　歧义消解

# The Maximum Probability Segmentation Algorithm of Ambiguous Character Strings

Liu Ting　WangKaizhu　Jiang Xinghai
(Dept. of Computer, Harbin Institute of Technology, 150001)
Shao Yanqiu
(Heilongjiang College of Communications, 150050)

**Abstract**: This paper presents one kind of segmentation algorithm of Chinese ambiguous character strings. The algorithm evaluates the probability of every kind of segmentation form, and takes the most prossible form as the result.

**Keywords**: Computational Linguistics, Chinese Automatic Segmentation , Unambiguation

## 一、引言

歧义字段的切分是汉语自动分词问题的关键，所谓歧义字段是指存在多种分词形式的汉字串，而歧义字段的机械式切分方法往往是将一种评价函数作用于各种分词形式之上，并以一种搜索算法求得函数值最大或最小的分词形式作为最终的分词结果。通常评价函数可以取为：词长、词频、词数等，也有的取整合程度系数、关联系数等[1]，搜索算法有爬山式算法、Dijkstra算法等。

机械式分词算法有很多种，其中1991年出现的最少分词词频选择算法(FWF)十分引人注目[2]。该算法先利用Dijkstra算法求得若干个词数最少的分词形式，再从这些分词形式中选择词频和最大的一种作为分词结果。例如：字符串“所有的”，根据最少分词的原则可分为“所有\的”和“所\有的”，前一种形式的词频和为：(1408+27c)/N，后一种形式的词频和为：(70+9c)/N，其中c为某一正整常数，N为语料字数，此处N=3万字。显然前一种切分形式被选中，因为它出现的可能性大。

和最大匹配法相比，FWF明显地提高了消解歧义的能力，但是由于该算法的第一步采用最少匹配的原则，因而对于组合型字段AB，一概分成AB，而不会考虑分成A\B的可能性，对于象“结合成分子时”这样的歧义字段也只有“结合\成分\子时”这一种分法。此外，将分词过程分为最少匹配和词频选择两个过程增加了时空复杂度，而用词频和作为评价函数则缺乏充足的理论根据。

笔者经过分析，提出了一种针对汉语歧义字段的最大概率切分算法(MP)。该算法追求分词结果中各词词频的乘积最大，而词长和词数不再是分词的依据，搜索算法采用人工智能中的$A^*$算法，实验显示：MP算法除了能够解决FWF算法所能够解决的歧义切分外，还能够轻松地解决象“结合成分子时”这样的切分难题，同时，对于组合型字段AB，当A和B的词频的乘积大于AB的词频时，将切分为A\B。

## 二、算法描述

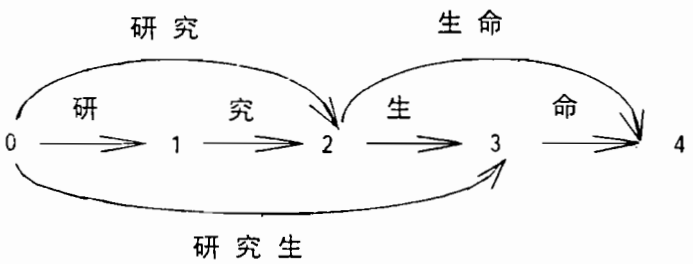


图1　歧义字段切分的状态空间图

本算法将歧义字段的切分问题视为状态空间的搜索问题，状态空间的示例如图1，初始结点为0，目标结点为n，n=字段中的汉字总数。字段中的一个词对应于图中的一条弧，每一条弧的费用设定为-logF，其中F为与该弧相对应的词的词频。最大概率分词问题归结为从状态空间图中搜索出一条从初始结点到达目标结点的最小费用路径的问题。

假设在最小费用路径上有m个词，则该路径的费用$\sum_{i=1}^{m}-\log F_i$是各条路径中最小的，根据对数的性质有$-\log\prod_{i=1}^{m}F_i$最小，又因为-log是反比例函数，所以$\prod_{i=1}^{m}F_i$是最大的，即最小费用路径就是最大概率路径。

MP采用$A^*$算法完成对最小费用路径的搜索，有关$A^*$的详细描述请参阅有关书目，如文献[3]等。MP算法设定任意结点的评价函数值f(v)=g(v)+h(v)，其中g(v)为从结点0到结点v的最小费用，h(v)根据n-v的值从预先填写的h值表得出。h值表是一个向量，h[i]表示当n-v=i时h(v)的值，显然有h[0]=0，当i>0时，假设面临一个i长汉字串的分词问题，该汉字串中的任意j长子串的费用均为词典中最高频j字词的词频的负对数，利用等费用搜索算法[3]可以获得该汉字串从串首至串尾的最小费用值，将此值赋给h[i]即可。显然，对于任何结点v，有$h(v)\le h^*(v)$，从而保证了算法的可纳性。

## 三、实验结果举例

本文实验中使用的词频词典主要来源于文献[4]，词典中含词46343个，总词次13027024。

(1) 25个典型歧义字段的切分结果

乒乓球拍 卖 完 了
结合 成 分子 时
研究生 --般 年龄 较 大
研究 生命 起源
这个 研究所 不大
这 项 研究所 涉及 的 问题 很 复杂 *
实现 在 情报 工作 方面 的 自动化
不同 情况 下 有 不同 解释
用 方块图 形式 来 描述
他 看见 一 只 白 天鹅
让 位移 小于 两 毫米
独立自主 和 平等 互利 原则
这 支 歌 太 平淡 无味 了
产品 需求 和 规格 说明
其实 也 是 看中 和 中国 大陆 做 生意 的 机会
战事 已经 有 了 结局
发展中国家 的 经济 状况 很 好

发展　中国　家庭　副业
使用　户外　天线　要　注意　避雷
使　用户　满意　的　做法
昨天　下午　他　不　在
他　将来　上海　*
将来　的　上海　会　有　严重　污染
他　从　马上　下来　*
老师　叫　你　马上　去

加黑的地方为句中重点解决的歧义字段，标记*号的句子切分有误。

(2)　两个有趣的切分实验

在以下两组实验中，尽管交集链的链长不断增加，但每次的切分结果都是令人满意的。

结合
结合　成
结合　成分
结合　成　分子
结合　成　分子　时

研究
研究生
研究　生命
研究生　命题
研究　生命　题目
研究生　命题　目的
研究　生命　题目　的确
研究生　命题　目的　确实

以下列出上例中各词的频度供参考（词频=词的频度/词典中的总词次）：

| | | | | | |
|---|---|---|---|---|---|
| 结 3471 | 结合 3721 | 合 3946 | 合成 832 | 成 20608 | 成分 1605 |
| 分 17880 | 分子 4457 | 子 6453 | 子时 13 | 时 43223 | |
| 研 357 | 研究 9580 | 研究生 213 | 究　402 | 生 8722 | 生命 1096 |
| 命 1376 | 命题 635 | 题 2126 | 题目 313 | 目 1411 | 目的 3195 |
| 的 849386 | 的确 357 | 确 1167 | 确实 661 | 实 2937 | |

表1　词的频度列表

# 四、讨论

[1] 本文提出的算法是针对歧义字段的，状态空间图中的初始结点和目标结点均对应着字段的端点。在实际的分词系统中，需要先将句子的分词问题归约为若干字段的分词问题，即划分字段。这一步较为简单，有兴趣的读者请参阅文献[5]。

[2] 本算法是在假设词与词之间相互独立的前提下给出的，即采用的是Unigram模型，若采用Bigram或Trigram模型，切分精度将进一步提高，但开销太大。

[3] 通常将歧义字段分为交集型歧义字段和组合型歧义字段两种，其中组合型歧义字段通常定义如下：AB既可分为AB，也可分为A\B，则AB为组合型歧义字段。由于约有2000多常用汉字可作为单字词使用，根据上述定义，大多数二字或二字以上词均可视为组合型歧义字段。这样组合型歧义字段的数量远远超过交集型歧义字段的数量，这与通常的看法正好相反。在文献[6]关于组合型字段的定义中补充了一点，即A\B这种切分方法在汉语中是有意义的，然而某中切分是否有意义是很难断定的，比如：在“第7位置1”中“位置”应切分为“位\ 置”，脱离了这种环境谁又能断定“位\ 置”这种切分形式是否有意义呢。

根据以上的思考笔者认为：所谓歧义字段就是从段首到段尾存在两条或两条以上的通路的汉字串。根据这种形式上的定义，歧义字段在句子中是大量存在的，而人们意识到的歧义字段往往具有这样的性质：在该字段中，除了最小费用路径以外，还存在一条从初始结点到目标结点的路径，该路径的费用和最小费用路径的费用接近。这两条路径的费用越接近，切分难度越高。例如：“将来”的费用为4.08，“将\来”的费用为2.84+2.49=5.33，比值为4.08/5.33=0.77；“马上”的费用为4.17，“马\上”的费用为3.55+2.22=5.77，比值为4.17/5.77=0.72。而“位置”的费用为3.55，“位\置”的费用为3.33+3.89=7.22，比值为3.55/7.22=0.49。因此“将来”和“马上”通常被收集为组合型歧义字段，而“位置”通常视为无歧义。“结合\成\分子\时”的费用为3.54+2.80+3.47+2.48=12.29，“结合\成分\子时”的费用为3.54+3.91+6.00=13.45，比值为12.29/13.45=0.91，这一比值显示该字段的切分难度很高。（这里词的费用为词频的负对数，对数的底数取10）人们感到组合型歧义很少，而交集型歧义较多，这是因为绝大多数组合型字段的切分难度很低，而交集型字段的切分难度偏高。

[4] 本算法给出了孤立看待一个歧义字段时该字段的最佳切分方法。也可以说，是根据歧义字段的固有性质进行切分的，没有考虑句内及篇中上下文的影响。歧义字段进入上下文后，可能出现反常的切分形式，例如，“马上”的固有切分形式是“马上”，而不是“马\上”，但在“他\从\马\上\下来”一句中，应取其反常的切分形式“马\上”。从宏观上讲，这种反常的情况很少出现，因而本算法的精度很高。

[5] 本算法中每个词的费用取其频率的负对数，而某一事件的发生频率的负对数恰好等于该事件的自信息量。费用最小的解路径对应着信息量最小的切分形式，这一结论的心理学

依据是：人们面对一条消息时往往倾向于将其理解为自己最熟悉的一种模式，而熟悉的模式所负载的信息量要小于不熟悉的模式所负载的信息量。

[6] 当每个词的费用均取1，并采用Dijkstra算法时，MP算法褪化为最少分词算法。和基于规则的歧义切分算法相比，MP算法省去了收集规则、组织规则的麻烦，但由于缺乏上下文知识，无力解决象“他从马上下来”这样的反常切分现象。

## 五、结语

本文提出了一种只依靠词频来解决汉语歧义字段切分的算法，该算法采用了人工智能中的状态空间搜索技术，能够迅速有效的解决绝大多数歧义切分问题。

近来，笔者通过计算生词的费用和词在上下文中的频率，提高了本算法识别生词和利用上下文信息的能力，由于篇幅所限这方面的内容待另文发表。

## 参考文献


[1] 马宴，基于评价的汉语自动分词系统的研究与实现，清华大学硕士论文，1991年6月

[2] 王晓龙等，自然语言理解中的音字流自动分词，中文信息学报，1991年，第5卷，第3期

[3] 傅京孙等，人工智能及其应用，清华大学出版社，1987年9月

[4] 刘源等，现代汉语常用词词频词典，宇航出版社，1990年

[5] 王晓龙，最少分词问题及其解法，科学通报，1989年第13期

[6] 刘源等，信息处理用现代汉语分词规范及自动分词方法，清华大学出版社、广西科学技术出版社，1994年6月

[7] 何克抗等，书面汉语自动分词专家系统设计原理，中文信息学报，1991年，第2期

[8] 候敏等，汉语自动分词中的歧义问题，计算语言学进展与应用，清华大学出版社，1995年10月