## 一．基于词典匹配的方法（字符串匹配、机械分词方法）

定义:按照一定策略将待分析的汉字串与一个“大机器词典”中的词条进行匹配，若在词典中找到某个字符串，则匹配成功。

按照扫描方向的不同：正向匹配和逆向匹配按照长度的不同：最大匹配和最小匹配

## 1.正向最大匹配思想 MM

（1）从左向右取待切分汉语句的m 个字符作为匹配字段，m 为大机器词典中最长词条个数。

（2）查找大机器词典并进行匹配。若匹配成功，则将这个匹配字段作为一个词切分出来。

若匹配不成功，则将这个匹配字段的最后一个字去掉，剩下的字符串作为新的匹配字段，进行再次匹配，重复以上过程，直到切分出所有词为止。

## 2. 逆向最大匹配算法 RMM

该算法是正向最大匹配的逆向思维，匹配不成功，将匹配字段的最前一个字去掉，实验表明，逆向最大匹配算法要优于正向最大匹配算法。

## 双向最大匹配法(Bi-directction Matching method,BM)

双向最大匹配法是将正向最大匹配法得到的分词结果和逆向最大匹配法的到的结果进 行比较，从而决定正确的分词方法。据 SunM.S. 和 Benjamin K.T.（1995）的研究表明，中文中 90.0％左右的句子，正向最大匹配法和逆向最大匹配法完全重合且正确，只有大概 9.0％的句子两种切分方法得到的结果不一样，但其中必有一个是正确的（歧义检测成功），只有不到 1.0％的句子，或者正向最大匹配法和逆向最大匹配法的切分虽重合却是错的，或者正向最大匹配法和逆向最大匹配法切分不同但两个都不对（歧义检测失败）。这正是双向最大匹配法在实用中文信息处理系统中得以广泛使用的原因所在。

## 4. 设立切分标志法

优先在待分析字符串中识别和切分出一些带有明显特征的词，以这些词作为断点（切分标志），可将原字符串分为较小的串再来进机械分词，从而减少匹配的错误率。

**5. 最佳匹配（OM，分正向和逆向）**

对分词词典按词频大小顺序排列，并注明长度，降低时间复杂度。

## 二．基于统计的分词

主要思想：上下文中，相邻的字同时出现的次数越多，就越可能构成一个词。因此字与字相邻出现的概率或频率能较好的反映词的可信度。

主要统计模型为：N 元文法模型（N-gram）、隐马尔科夫模型(Hidden Markov Model, HMM)

## 1. N-gram 模型思想

模型基于这样一种假设，第 n 个词的出现只与前面N-1 个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积 .

我们给定一个词，然后猜测下一个词是什么。当我说“艳照门”这个词时，你想到下一个词是什么呢？我想大家很有可能会想到“陈冠希”，基本上不会有人会想到“陈志杰”吧。

N-gram 模型的主要思想就是这样的。

对于一个句子 T，我们怎么算它出现的概率呢？假设 T 是由词序列W1,W2,W3,…Wn

组成的，那么

*P*(*T* )  *P*(*W*1*W*2*W*3...*Wn* )  *P*(*W*1)*P*(*W*2 |*W*1)*P*(*W*3 |*W*1*W*2 )...*P*(*Wn* |*W*1...*Wn*1)

但是这种方法存在两个致命的缺陷：一个缺陷是参数空间过大，不可能实用化；另外一个缺陷是数据稀疏严重。

为了解决这个问题，我们引入了马尔科夫假设：一个词的出现仅仅依赖于它前面出现的有限的一个或者几个词。

## 2. 隐马尔科夫模型思想

如果一个词的出现仅依赖于它前面出现的一个词，那么我们就称之为 bigram。即

*P*(*T* )  *P*(*W*1*W*2*W*3...*Wn* )  *P*(*W*1 )*P*(*W*2 | *W*1 )*P*(*W*3 | *W*1*W*2 )...*P*(*Wn* |*W*1...*Wn*1)

 *P*(*W*1 )*P*(*W*2 | *W*1 )*P*(*W*3 | *W*2 )...*P*(*Wn* |*Wn*1 )

如果一个词的出现仅依赖于它前面出现的两个词，那么我们就称之为 trigram。

在实践中用的最多的就是 bigram 和trigram 了，而且效果很不错。高于四元的用的很少，因为训练它需要更庞大的语料，而且数据稀疏严重，时间复杂度高，精度却提高的不多。

设 w1,w2,w3,...,wn 是长度为 n 的字符串，规定任意词 wi 只与它的前两个相关，得到三元概率模型

*n*

*P*(*W* )  *P*(*W*1 )*P*(*W*2 | *W*1 )*P*(*Wi* | *Wi*2*Wi*1 )

*i*3

以此类推，N 元模型就是假设当前词的出现概率只同它前面的 N-1 个词有关。

## 3. 基于字标注的中文分词方法

以往的分词方法，无论是基于规则的还是基于统计的，一般都依赖于一个事先编制的词表(词典)。自动分词过程就是通过词表和相关信息来做出词语切分的决策。与此相反，基于字标注的分词方法实际上是构词方法。即把分词过程视为字在字串中的标注问题。由于每个字在构造一个特定的词语时都占据着一个确定的构词位置(即词位)，假如规定每个字最多只有四个构词位置：即 B(词首)，M (词中)，E(词尾)和 S(单独成词)，那么下面句子(甲)的分词结果就可以直接表示成如(乙)所示的逐字标注形式：

(甲)分词结果：／上海／计划／N／本／世纪／末／实现／人均／国内／生产／总值／五千美元／

(乙)字标注形式：上／B 海／E 计／B 划／E N／S 本／s 世／B 纪／E 末／S 实／B 现／

E 人／B 均／E 国／B 内／E 生／B 产／E 总／B 值／E 五／B 千／M 美／M 元／E 。

／S

首先需要说明，这里说到的“字”不只限于汉字。考虑到中文真实文本中不可避免地会包含一定数量的非汉字字符，本文所说的“字”，也包括外文字母、阿拉伯数字和标点符号等字符。所有这些字符都是构词的基本单元。当然，汉字依然是这个单元集合中数量最多的一类字符。

把分词过程视为字的标注问题的一个重要优势在于，它能够平衡地看待词表词和未登录词的识别问题。在这种分词技术中，文本中的词表词和未登录词都是用统一的字标注过程来实现的。在学习架构上，既可以不必专门强调词表词信息，也不用专门设计特定的未登录词(如人名、地名、机构名)识别模块。这使得分词系统的设计大大简化。在字标注过程中，所有的字根据预定义的特征进行词位特性的学习，获得一个概率模型。然后，在待分字串上，根据字与字之间的结合紧密程度，得到一个词位的标注结果。最后，根据词位定义直接获得最终的分词结果。总而言之，在这样一个分词过程中，分词成为字重组的简单过程。然而这一简单处理带来的分词结果却是令人满意的。

## 三．基于规则的分词（基于语义或理解）

通过模拟人对句子的理解，达到识别词的效果，基本思想是语义分析，句法分析，利用句法信息和语义信息对文本进行分词。自动推理，并完成对未登录词的补充是其优点。不成熟.

这种分词方法是通过让计算机模拟人对句子的理解，达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断，即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段。

具体概念:有限状态机、语法约束矩阵、特征词库

## 四．三种分词算法的比较

到目前为止，还无法证明哪一种方法更准确，每种方法都有自己的利弊，有强项也有致命弱点，简单的对比见下表所示：

各种分词方法的优劣对比

|  |  |  |  |
| --- | --- | --- | --- |
| **分词方法** | **基于字符串匹配分词** | **基于理解的分词** | **基于统计的分词** |
| 歧义识别 | 差 | 强 | 强 |
| 新词识别 | 差 | 强 | 强 |
| 需要词典 | 需要 | 不需要 | 不需要 |
| 需要语料库 | 否 | 否 | 是 |
| 需要规则库 | 否 | 是 | 否 |
| 算法复杂性 | 容易 | 很难 | 一般 |
| 技术成熟度 | 成熟 | 不成熟 | 成熟 |
| 实施难度 | 容易 | 很难 | 一般 |
| 分词准确性 | 一般 | 准确 | 较准 |
| 分词速度 | 快 | 慢 | 一般 |

# 1.歧义识别

歧义识别指一个字符串有多种分词方法，计算机难以给出到底哪一种分词算法才是正确的分词系列。如"表面的"可以分为"表面/的"或"表/面的"。计算机无法判断哪一种才是准确的分词系列。

基于字符串的分词算法：仅仅是跟一个电子词典进行比较，故不能进行歧义识别； 基于理解的分词算法：指通过理解字符串的含义，故有很强的歧义识别能力；

基于统计的分词算法：根据字符连续出现次数的多少，得到分词系列，故常常能够给出正确的分词系列选择，但是也有可能判断错误的情况。

**2.新词识别**

新词识别也称作未登录词识别，指正确识别词典中没有出现的词语。姓名、机构名、地址、称谓等千变万化，词典中常常不能完全收录这些词语；另外，网络中出现的流行用语也是一种未登录词的常见来源，如"打酱油"为最近出现在网络中， 并迅速流行，从而成为一个新词。大量的研究证明新词识别是中文分词准确性的一个重要影响因素。

基于字符串的分词算法：无法正确识别未登录词，因为这种算法仅仅与词典中存在的词语进行比较；

基于理解的分词算法：理解字符串的含义，从而有很强的新词识别能力；

基于统计的分词算法：这种算法对第二种未登录词有很强的识别能力，因为出现次数多，才会当作一个新词；对于第二类未登录词，这类词语有一定的规律，如姓名：

"姓"+ 名字，如李胜利；机构：前缀+称谓，如希望集团；故需要结合一定的规则进行识别，仅仅统计方法难以正确识别。

**3.需要词典**

基于字符串的分词算法：基本思路就是与电子词典进行比较，故电子词典是必须的。并且词典越大，分词的正确率越高，因为词典越大，未登录词越少，从而可以大大减少未登录词识别的错误；

基于理解的分词算法：理解字符串的含义，故不需要一个电子词典；

基于统计的分词算法：仅仅根据统计得到最终的结果，故电子词典不是必须的。

**4.需要语料库**

基于字符串的分词算法：分词过程仅仅与一个已经存在的电子词典进行比较，故不需要语料库；

基于理解的分词算法：理解字符串的含义，故不需要电子词典；

基于统计的分词算法：需要语料库进行统计训练，故语料库是必须的；且好的语料库是分词准确性的保证。

**5.需要规则库**

基于字符串的分词算法：分词过程仅仅与一个已经存在的电子词典进行比较，不需要规则库来进行分词；

基于理解的分词算法：规则是计算机进行理解的基础，故准确、完备的规则库是这种分词算法的前提；

基于统计的分词算法：根据语料库统计训练，故规则库不是必须的。

**6.算法复杂性**

基于字符串的分词算法：仅仅进行字符串的比较操作，故算法简单；

基于理解的分词算法：需要充分处理各种规则，故算法非常复杂；事实上到目前为止，还没有成熟的这类算法；

基于统计的分词算法：需要语料库进行训练，虽然算法也比较复杂，但是已经比较常见，故这种分词的复杂性比第一种大，比第二种容易。现在的实用分词系统都采用这种算法。

**7.技术成熟度**

基于字符串的分词算法：是最早出现也是最成熟的算法；

基于理解的分词算法：是最不成熟的一类算法，到目前为止还没有成熟的算法； 基于统计的分词算法：已经有多种成熟的这类算法，基本上能够满足实际的应用。

故技术成熟度：基于匹配的分词算法>基于理解的分词算法>基于统计的分词算法。

**8.实施复杂性**

同上面的道理，实施复杂性：基于理解的分词算法>基于统计的分词算法>基于匹配的分词算法。

**9.分词准确性**

到目前为止还没有一个准确的结论，不过从理论上说，基于理解的分词算法有最高的分词准确性，理论上有 100%的准确性；而基于匹配的分词算法和基于统计的分词算法是一种"浅理解"的分词方法，不涉及真正的含义理解，故可能会出现错误， 难以达到 100%的准确性。

**10.分词速度**

基于匹配的分词算法：算法简单，操作容易，故分词速度快，所以这种算法常常作为另外两种算法的预处理，进行字符串的粗分；

基于理解的分词算法：这种算法常常需要操作一个巨大的规则库，故速度最慢； 基于统计的分词算法：这种分词算法仅仅是与一个统计结果进行比较，故速度一般。

故一般的分词速度从快到慢依次为：基于匹配的分词算法>基于统计的分词算法>基于理解的分词算法。

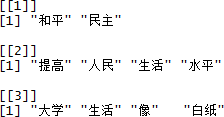
**五．分词算法的几个案例**

对下列三句话分词："和平民主"、"提高人民生活水平"、"大学生活像白纸"。

1. 用最大正向匹配方法得到的结果截图：



1. 用基于统计的方法（隐马尔科夫模型）截图：



将"人民生活水平"和"大学生活"添加到词库中，则得到很好的分词效果。

