分词算法设计中的几个基本原则：

1、颗粒度越大越好：用于进行语义分析的文本分词，要求分词结果的颗粒度越大，即单词的字数越多，所能表示的含义越确切，如：“公安局长”可以分为“公安 局长”、“公安局 长”、“公安局长”都算对，但是要用于语义分析，则“公安局长”的分词结果最好（当然前提是所使用的词典中有这个词）

2、切分结果中非词典词越少越好，单字字典词数越少越好，这里的“非词典词”就是不包含在词典中的单字，而“单字字典词”指的是可以独立运用的单字，如“的”、“了”、“和”、“你”、“我”、“他”。例如：“技术和服务”，可以分为“技术 和服 务”以及“技术 和 服务”，但“务”字无法独立成词（即词典中没有），但“和”字可以单独成词（词典中要包含），因此“技术 和服 务”有1个非词典词，而“技术 和 服务”有0个非词典词，因此选用后者。

3、总体词数越少越好，在相同字数的情况下，总词数越少，说明语义单元越少，那么相对的单个语义单元的权重会越大，因此准确性会越高。

最大匹配法：最大匹配是指以词典为依据，取词典中最长单词为第一个次取字数量的扫描串，在词典中进行扫描（为提升扫描效率，还可以跟据字数多少设计多个字典，然后根据字数分别从不同字典中进行扫描）。例如：词典中最长词为“中华人民共和国”共7个汉字，则最大匹配起始字数为7个汉字。然后逐字递减，在对应的词典中进行查找。