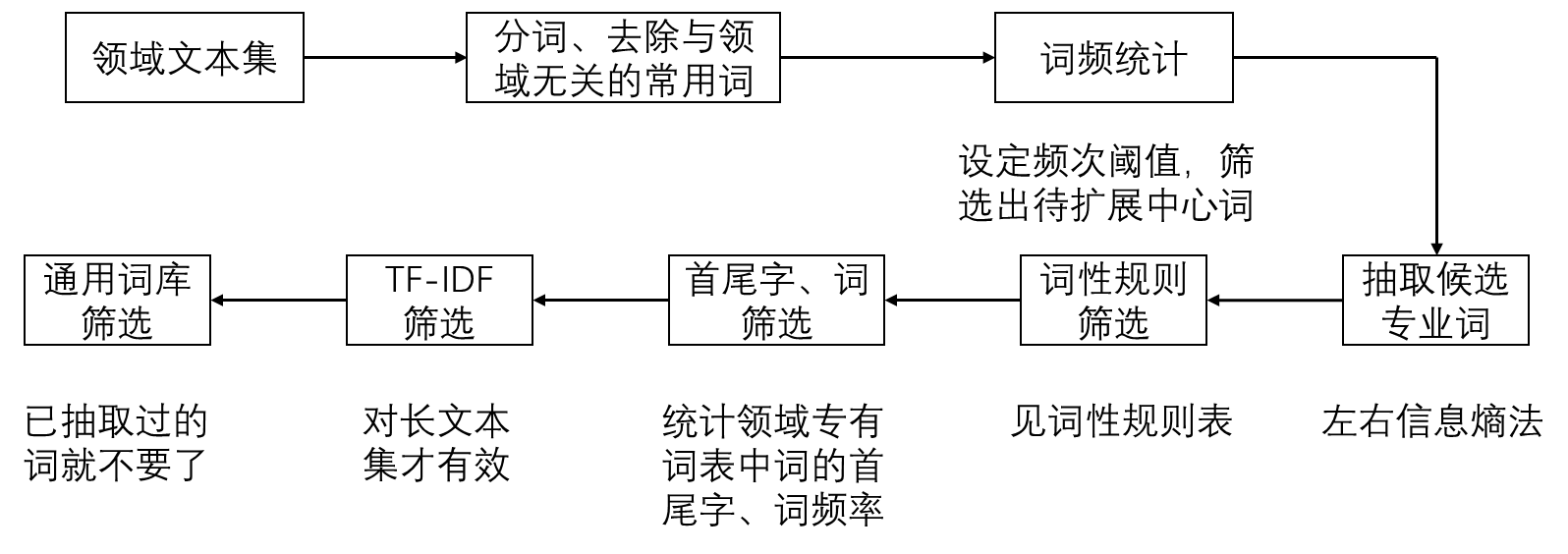
信息抽取技术应用

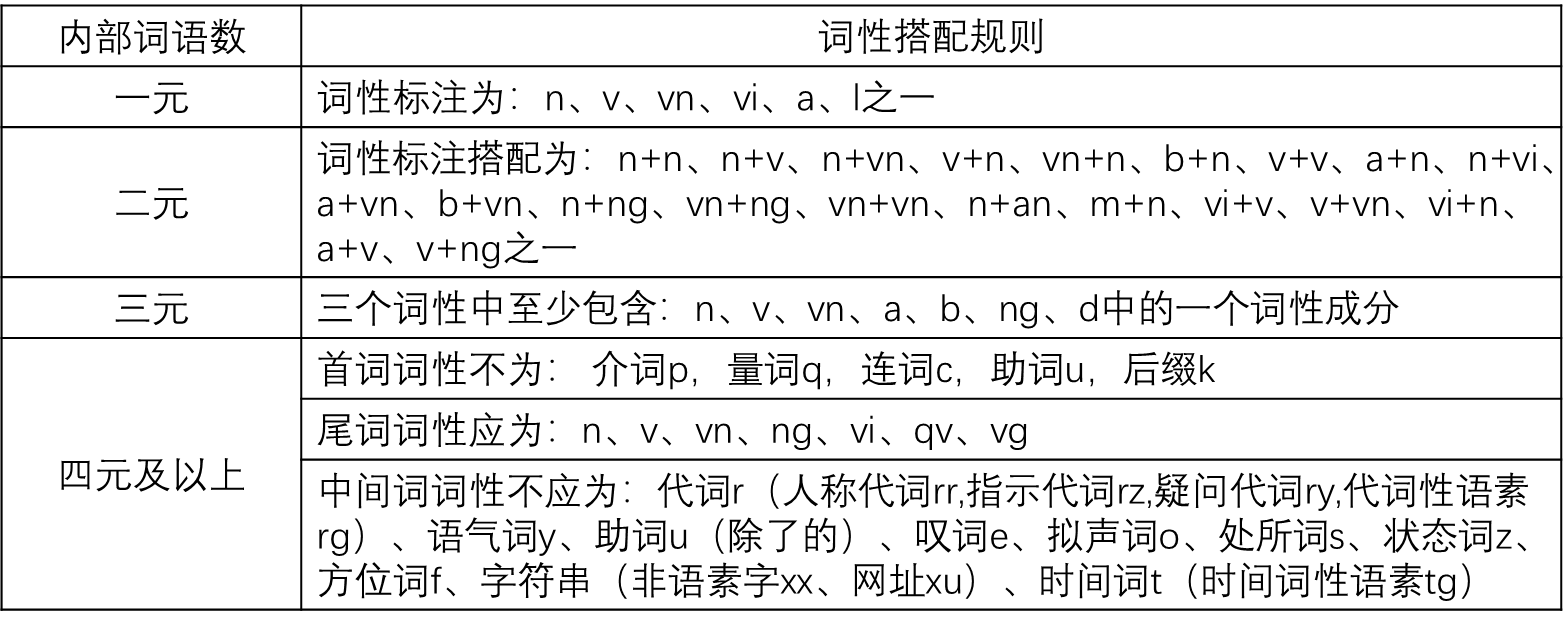
# 新词识别

## 面向特定领域的专业词汇抽取

### 抽取算法流程



### 词语搭配规则

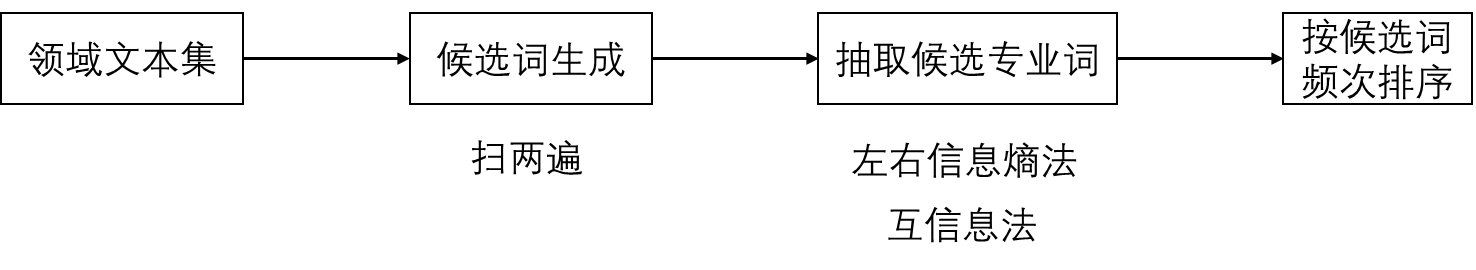


### 边界信息概率库

如果想用边界信息概率库来过滤掉不能成词的候选词，则需要用专业词表来提供首尾字词概率信息。在《统计语言建模与中文文本自动校对技术》中，作者使用了《人民日报》语料中长度大于4的词，2000篇计算机领域文献的关键词，搜狗计算机领域专业词库作为专业词表。

## HanLP-v1.7.6新词抽取

### 抽取算法流程



1. 遍历输入文档集的每一行或输入的一行文本串，遍历时使用两个指针i和j，i从头扫到尾，j从i后开始扫到尾，每次i、j的移动抽取出一个候选新词，更新候选词词典中候选词词频以及候选词左右的字种及字频。对于含有n个字符的一行文本，**遍历时间复杂度为o(n^2)**。从候选词生成的过程以及后续的信息熵和互信息的计算，我们可以看到这种新词抽取**要求输入的文本不能太短**。从最后的按候选词频率排序还可以看到，**如果候选词的词频太低，也是抽取不出来的**。我们能否按照一定规则把高频的非专业词的候选词过滤出来，是这种基于统计方法抽取新词的关键。
2. 计算左右信息熵时，如果熵值小于算法输入参数min\_entropy，则说明候选词的左侧或右侧的字符分布不均匀，更有可能只有一种或两种词语，此时就可以向左或向右扩展出新的候选词。在HanLP中，将候选词的左、右熵的最小值作为候选词的熵值，对于熵值小于**min\_entropy**的候选词，直接过滤掉。
3. 互信息用于度量两个子串间的紧密程度，互信息值越大说明两个子串联系越紧密。对于一个候选词，HanLP将拆分出所有可能的左右子串，将所有成对子串互信息的最小值作为候选词的互信息值，对于互信息值小于**min\_aggregation**的候选词，直接过滤掉。
4. HanLP提供的其他参数
5. size 抽取出的topN候选词数量。
6. newWordsOnly 如果该参数值为true，则过滤掉HanLP词库中的词。
7. max\_word\_len 抽取的候选专业词最多字符数。
8. min\_freq 过滤掉词频低于该值的候选词。
9. 由于**没有分词步骤**，所以也就谈不上词性规则过滤，高频的常见词会被抽取出来。对于小说类文本来说，称呼的频率高，很容易被抽取出来。因此，可以考虑首先对文本用2-gram+Viterbi分词得到词表，用该词表过滤抽取出的新词。

### 候选新词左右熵

候选新词的左熵定义为：

其中，，表示候选新词左侧出现过的字或词，表示候选新词左侧出现过的字种或词种集合，表示字符串出现的次数，表示字符串出现的次数。

候选新词的右熵定义为：

其中，，表示候选新词右侧出现过的字或词，表示候选新词右侧出现过的字种或词种集合。表示字符串出现的次数，表示字符串出现的次数。

注意：无论是左熵还是右熵，都不是条件熵的定义，所以不可以说计算候选词左、右条件熵。

### 候选新词内部互信息

两个子串与间的互信息定义为：

其中，表示词语总数，表示与共同出现的次数，表示出现的次数，表示出现的次数。

注意：张仰森在《统计语言建模与中文文本自动校对技术》中指出，当两者出现次数较少时，反而互信息的值越大，因此，他引入了三次互信息的公式。