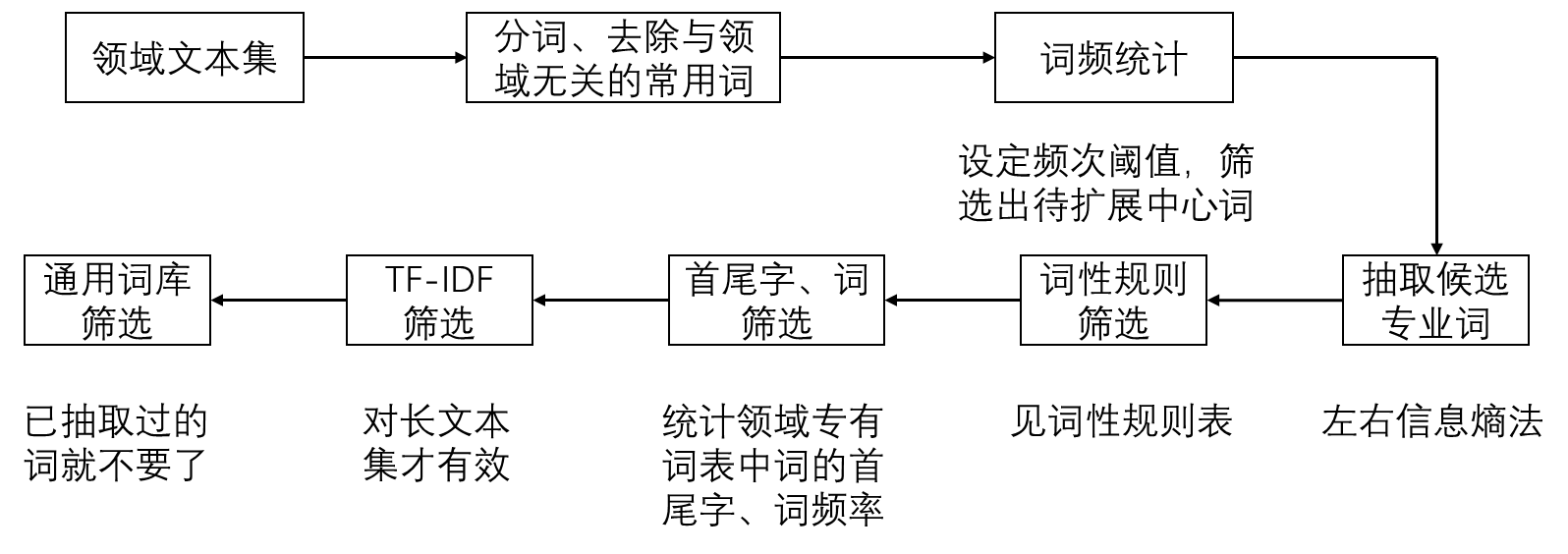
信息抽取技术应用

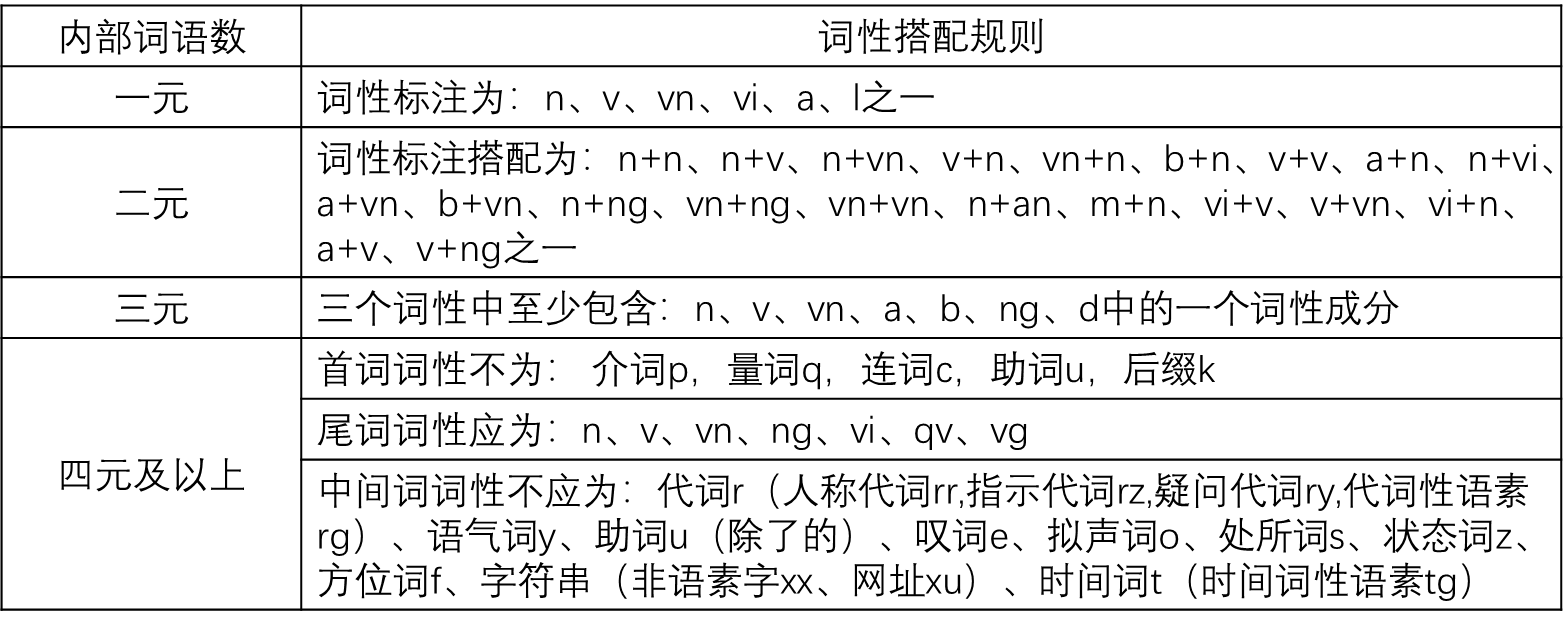
# 新词识别

## 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

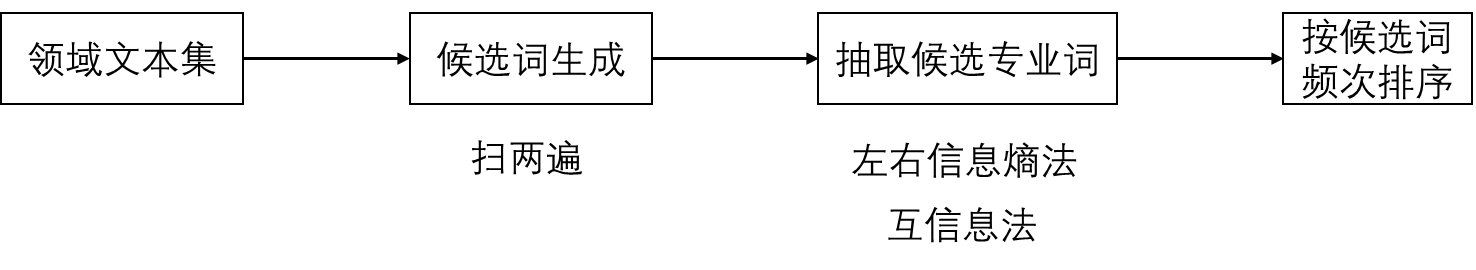


### 边界信息概率库

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，表示候选新词出现的次数。

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，表示候选新词出现的次数。

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。

示例：已知有下边的两句话

|  |
| --- |
| 两只蝴蝶飞呀飞  这些蝴蝶飞走了 |

计算和

解：

计算候选新词左侧字符的分布

|  |  |  |
| --- | --- | --- |
|  | 只|蝴蝶 | 些|蝴蝶 |
| p | 1/2 | 1/2 |

，，

计算候选新词右侧字符的分布

|  |  |
| --- | --- |
|  | 飞|蝴蝶 |
| p | 1 |

，

显然，右熵已达最小值，说明候选新词“蝴蝶”需要继续向右扩展，变为“蝴蝶飞”。HanLP中，将左右熵的最小值作为候选新词的熵，当熵值小于设定阈值后，HanLP就会丢弃掉这个候选新词，也就相当于向两侧扩展。

### 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。

## 素问术语抽取

句子开头词的左熵的计算

句子开头词的右熵的计算

# 搭配

## 搭配的定义

一个指定词的搭配就是对这个词的**惯用位置**的描述。搭配包括：**名词短语、动词短语、固定搭配**。搭配的构成方式有：复合构成、非复合构成。如果可以从搭配的各组成部分意思推出整体的意思，称这个搭配是**复合构成**的。如果不能从搭配的各组成部分意思推出整体的意思，称这个搭配是**非复合构成**的。常见的非复合构成的搭配有：成语或俚语。自然语言生成、计算词典编撰学、句法分析、语料库语言学等研究和应用中都需要用到搭配抽取技术。

## 搭配抽取方法

### 频次法

基本方法是，统计2gram频次，认为高频的2gram最有可能成为搭配，输出搭配时再用词性序列规则和停用词表过滤。这一过滤过程，类似于专业词汇抽取时用到的过滤方法。

### 位置均值、偏差法

在一个3-4个词的窗口内，统计核心词与候选搭配词之间距离的均值和偏差（标准差），如果偏差很低（小于1.0），认为两个词之间具有搭配关系。这种词汇组合之间的联系比固定短语之间的联系更加松散，并且在它们之间插入的符号及其相对位置都是可变的。

### 假设检验法

前边两种搭配抽取方法存在这样的缺陷：如果一个频繁出现的2gram的两个构成词也是频繁出现的，例如new companies（new和companies都频繁出现），那么我们可以预期这两个词同现很多次仅仅是一个偶然现象，它们甚至不能形成一个搭配。实际上，前边新词识别计算词语内互信息时也存在这样的问题，不过那里我们只取词语内所有可能子串对的最小互信息值作为词语互信息，在一定程度上避免这种问题。

## NLTK和HanLP v1.7.6抽取类目核心短语

# 实体识别

## 识别方法

## 商品属性抽取