**词性标注**

词性用来描述一个词在上下文中的作用。例如描述一个概念的词叫做名词，在下文引用这个名词的词叫做代词。有的词性经常会出现一些新的词，例如名词，这样的词性叫做开放式词性。另外一些词性中的词比较固定，例如代词，这样的词性叫做封闭式词性。因为存在一个词对应多个词性的现象，所以给词准确地标注词性并不是很容易。比如："改革"在"中国开始对计划经济体制进行改革"这句话中是一个动词，在"医药卫生改革中的经济问题"中是一个名词。把这个问题抽象出来就是已知单词序列  ，给每个单词标注上词性 。

不同的语言有不同的词性标注集。比如英文有反身代词，例如myself，而中文则没有反身代词。为了方便指明词的词性，可以给每个词性编码。例如《PFR人民日报标注语料库》中把"形容词"编码成a；名词编码成n；动词编码成v等。

词性标注有小标注集和大标注集。例如小标注集把代词都归为一类，大标注集可以把代词进一步分成三类。

人称代词：你 我 他 它 你们 我们 他们

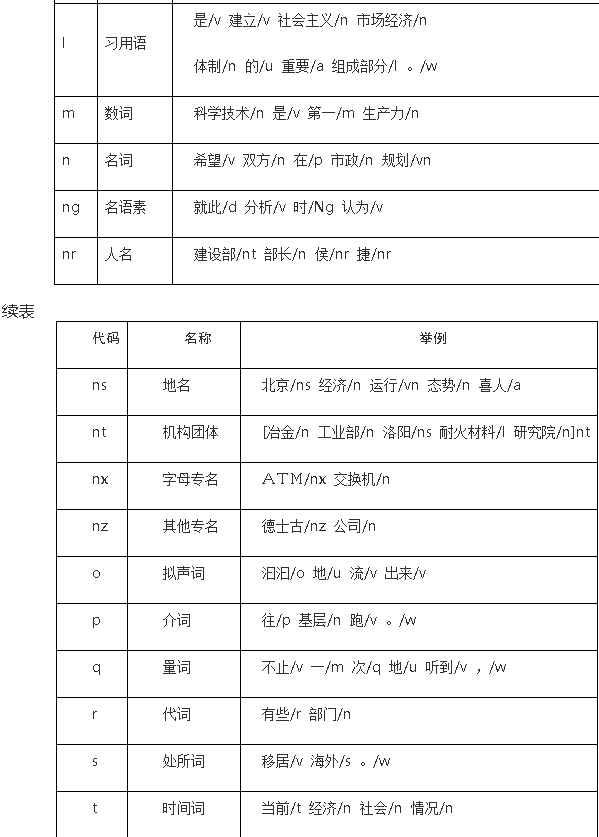
疑问代词：哪里  什么  怎么

指示代词：这里 那里  这些  那些

采用小标注集比较容易实现，但是太小的标注集可能会导致类型区分度不够。例如在黑白两色世界中，可以通过颜色的深浅来分辨出物体，但是通过七彩颜色可以分辨出更多的物体。

参考《PFR人民日报标注语料库》的词性编码表，如表4-4所示：







以[把][这][篇][报道][编辑][一][下]为例，有5×1×1×2×2×2×3＝120种词性标注可能性，如图4-10所示，哪种可能性最大？

解决标注歧义问题最简单的一个方法是从单词所有可能的词性中选出这个词最常用的词性作为这个词的词性，也就是一个概率最大的词性，比如"改革"大部分时候作为一个名词出现，那么可以机械地把这个词总是标注成名词，但是这样标注的准确率会比较低，因为只考虑了频率特征。

考虑词所在的上下文可以提高标注准确率。例如在动词后接名词的概率很大。"推进/改革"中的"推进"是动词，所以后面的"改革"很有可能是名词。这样的特征叫做上下文特征。

|  |
| --- |
| <http://images.51cto.com/files/uploadimg/20110615/155347652.jpg> |
| （点击查看大图）图4-10  词性标注 |

隐马尔可夫模型（Hidden Markov Model，HMM）和基于转换的学习方法是两种常用的词性标注方法。这两种方法都整合了频率和上下文两方面的特征来取得好的标注结果。具体来说，隐马尔可夫模型同时考虑到了词的生成概率和词性之间的转移概率。