Part-of-Speech 词性标注

**任务定义：**

设计一个序列学习模型来推测句子中每个词语的词性

从数据中选择80%作为训练数据，剩下的20%作为测试数据（或者5-hold cross validation）

**输入输出**

利用训练好的模型，可以完成对分好词的文本的词性标注任务

执行分词任务脚本：

**方法**

使用条件随机场工具CRF++对文档进行分词，首先把训练数据转换为CRF++需要的格式：每行一个词，第一列为词语本身，第二列为词语的词性，如图1中给的标注为例

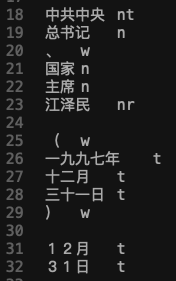


图1 词性标注训练数据格式

CRF++基于条件随机场CRF方法，相对于HMM，CRF的主要优点是它的条件随机性，只需要考虑当前已经出现的观测状态的特性，没有独立性的要求，对于整个序列内部的信息和外部观测信息均可有效利用，避免了MEMM和其他针对线性序列模型的条件马尔可夫模型会出现的标识偏置问题。

**源码运行环境**

|  |
| --- |
| 分词软件基于 CRF++-0.58  **需要先安装，网盘地址：**  链接:<https://pan.baidu.com/s/1-BPjdA1s0Gb2Olwn0MUAdw> 密码:gs8d  **CRF++系统安装**   * Requirements   C++ compiler (gcc 3.0 or higher)   * How to make   % ./configure  % make  % su  % make install  **安装到Python**  % python setup.py install  % 注：最后一步最好用pyhton2运行，python3会有各种各样的问题 |

**标注精确度评价**

经过处理后的测试数据格式与训练数据相同，第一行为词本身，第二列为准确标注的词性。CRF++会把推测出来的词性加在测试数据的第三列。

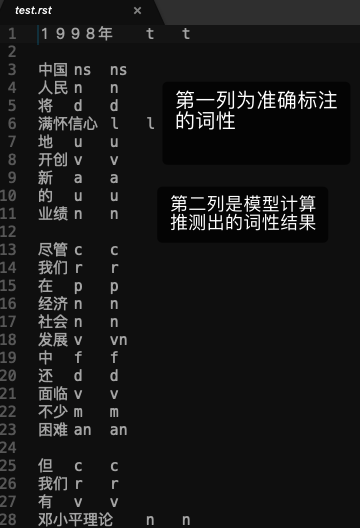


图2

词性标注的检验，比较简单。已经分好词了，我只需要把第二列，第三列不同的情况判断为“错误”就可以统计正确率了

**实验过程结果**

1.把课堂提供的训练数据进行处理，随机取80%作为训练数据，剩下的20%作为测试数据

2 用CRF++开始训练，由于词性标注词性多，特征函数多，训练需要非常长的时间，所以使用了网上下载的模型，但是这个模型的训练集和测试集和课堂提供的有些不同，词性分类更概括，它的训练数据放在data/文件夹下了。

3. 用训练好的模型在测试集上测试，计算准确率。

4 准确率：

|  |
| --- |
| WordCount from result: 98369  WordCount of correct post : 91865  准确率:0.933882 |

**提交代码文件结构：**

PartOfSpeechTagging

----CRF++/

--------data/

------------train\_test\_split.py #将语料库按80%和20%随机分为训练数据和测试数据

------------data\_process.py #处理训练数据和测试数据，转换为CRF++需要的格式

------------test.data 测试集

------------train.data 训练集

--------model/

------------model #训练好的模型

--------train.sh #训练指令

--------test.sh #标注效果测试指令

--------accuracy.py #计算词性标注准确率