POStag

MG1733099，周天烁，tianshuo.zhou@smail.nju.edu.cn

2017 年 12 月 5 日

OverView

part-of-speech tagging (POS tagging or PoS tagging or POST)即词性标注是自然语言处理（NLP）中的重要任务之一。具体就是给定由单词序列构成的句子，设计程序自动将单词的词性识别。这里利用隐马尔可夫模型，把词性作为隐状态空间，单词作为观测状态，模型参数λ=[A,B,π]利用最大似然估计求出，然后利用实验一的维特比算法计算给定句子的最可能的词性序列序列。具体实现如下。

Part1.标注预料

使用 python自然语言处理工具包nltk的brown语料。该语料的大约有5万个英文句子，其中的每个单词在句子中的词性都已由人工标注，在质量较高，此处想选择的标记符号为最简化的版本，如图1所示。

Part2.参数估计

各转移矩阵的值通过最大似然估计（Maximum Likelihood Estimation）即统计频率得出。具体实现则可以通过直接调用nltk库函数求出。

Part3.词性预测

估计出参数后，分别构造单词和词性的索引word2index和tag2index，再建立相应的转移矩阵[A,B,π]，通过索引把转移矩阵的每个值填充。

Part4.实验示例

该程序具有良好的用户界面，用户根据命令行提示输入句子，程序返回标注的词性，然后继续提示用户输入句子，输入大写字母E退出。程序运行示例如图2所示。

Universal Part-of-Speech Tagset

Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

图 1: Universal Part-of-Speech Tagset

```
In [7]: runfile('E:/课件/AML/assign2_code_v2/hmmpostag.py', wdir='E:/
课件/AML/assign2_code_v2')
Reloaded modules: myHMM

please input the sentece(E to exit):I love it
['START', 'PRON', 'VERB', 'PRON', 'END']

please input the sentece(E to exit):aa bb cc
error:some word not in the dictionary

please input the sentece(E to exit):I want to go
['START', 'PRON', 'VERB', 'PRT', 'VERB', 'END']

please input the sentece(E to exit):E
***program end***
```

图 2: 实验示例

Part5.运行环境

- python3
- import nltk；import myHMM
- from nltk.corpus import brown（需要下载brown语料库）