

词性标注实验报告

刘天伟 21009306 计算机 1010

一、实验环境

1. 计算机: Intel Pentium Dual-core 2.06GHz, 1.50GB 的内存
2. 操作系统: ubuntu 10.04 Lucid Lynx
3. 程序设计语言: C shell 脚本
4. 编译环境: gcc 4.4.3
5. 调试环境: gdb 7.1-ubuntu

二、附件内容

文件说明

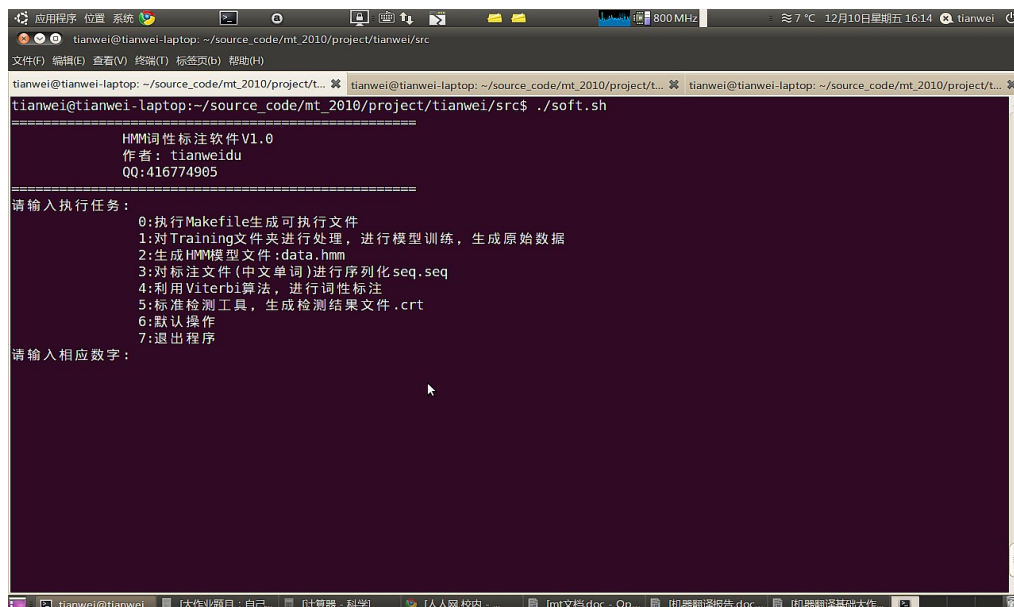
tianwei/	根文件夹
tianwei/src/	源码文件夹
./soft.sh	程序执行总 shell 脚本
./seq.sh	中文词汇测试文件生成序列 shell 脚本
./rightTest.sh	准确律检测 shell 脚本
./hmm.sh	HMM 模型生成 shell 脚本
./showViterbi.sh	Viterbi 训练脚本
./ttrain.sh	分组去除词性标注, 生成概率文件脚本
tianwei/src/	源码文件夹
./eva.c	准确率检测工具源码
./hmm.c	hmm 模型生成源码
./moveMatrix.c	计算概率矩阵源码
./seq.c	检测文件中文单词生成序列
./showViterbi.c viterbi.c viterbi.h	Viterbi 算法
tianwei/src/	
./Makefile	Makefile 文件
tianwei/Training/	初始训练文件
tianwei/ReadMe.txt	程序安装及运行方法

三、实验原理

1. 词性标注，简称标注，即为句子中每个词都标上一个合适的词性，也就是要确定每个词的名词、动词、形容词或其它词性。
2. 词性标注方法：包括基于规则的标注方法（Rule-based tagging）、基于统计的标注方法（Statistical tagging）和基于转换的标注方法（Transformation-based tagging）。实验中采用基于统计的标注方法。
3. 隐马尔科夫模型 HMM 的计算步骤：计算观察序列的概率；计算能够解释观察序列的最大可能的状态序列；根据观察序列寻找最佳参数模型。
4. Viterbi 算法，用于搜索能够生成观察序列的最大概率的状态序列。Viterbi 能够找到最佳解，其思想精髓在于将全局最佳解的计算过程分解为阶段最佳解的计算。

四、实验过程

1. 运行 ./soft.sh, 出现如下界面，选择相应数字，运行程序



```
tianwei@tianwei-laptop: ~/source_code/mt_2010/project/tianwei/src
tianwei@tianwei-laptop: ~/source_code/mt_2010/project/tianwei/src$ ./soft.sh

=====
HMM词性标注软件 V1.0
作者: tianweidu
QQ: 416774905
=====

请输入执行任务:
0: 执行Makefile生成可执行文件
1: 对Training文件夹进行处理, 进行模型训练, 生成原始数据
2: 生成HMM模型文件: data.hmm
3: 对标注文件(中文单词)进行序列化 seq.seq
4: 利用Viterbi算法, 进行词性标注
5: 标准检测工具, 生成检测结果文件.crt
6: 默认操作
7: 退出程序

请输入相应数字:
```

2. 执行 0 和 1 分别进行编译和文件自动分组，生成文件夹 train*, 同时为了提升训练速度，将文件进行合并。训练文件分组由 shell 脚本完成：

```

24 echo "=====完成CRT文件归类===== "
25 cnt=0
26 #对文件进行随意分割
27 for f in $(ls ./Training/*.crt)
28 do
29     #echo $f $cnt "-->" ${set_file[cnt]}
30     mv $f ${set_file[cnt]}
31     cnt=$((expr $cnt + 1))
32     if [ $cnt = 5 ]
33     then
34         #echo cnt
35         cnt=0
36     fi
37 done
38

```

train1	10 项	文件夹
key1	6 项	文件夹
right	1 项	文件夹
ch.crt	302.0 KB	DER/PEM/Netscape-encoded X.509 证书
out.crt	223.6 KB	DER/PEM/Netscape-encoded X.509 证书
res.crt	36.5 KB	DER/PEM/Netscape-encoded X.509 证书
train.crt	2.7 MB	DER/PEM/Netscape-encoded X.509 证书
tst.crt	666.6 KB	DER/PEM/Netscape-encoded X.509 证书
data.hmm	14.7 MB	纯文本文档
seq.seq	366.3 KB	纯文本文档
tags	332 字节	纯文本文档
train2	9 项	文件夹
train3	9 项	文件夹
train4	9 项	文件夹
train5	9 项	文件夹
Training	0 项	文件夹

- 然后，熟悉并清理语料，把训练语料中句子的序号删除（进行交叉测试时需要使用，即交叉测试时，需要把做作为测试语料的训练语料中的词性标记删除）。通过 sed 工具完成。

```

96 echo "=====清除训练集的词性标注===== "
97 #使用sed进行过滤
98 #sed -e 's/[0-9]\{1,\} //' -->过滤开始标号
99 #sed -e 's/[a-zA-Z]\{1,\} /g' -->过滤词性部分
100 #sed -e 's/[[:space:]]\{2,\} /g' -->过滤多余空格
101 #sed -e 's/[[:space:]] //' -->去除开始空格
102 #BAIKE002.crt>BAIKE002 -->文件令存为其他文件
103 for cnt in {1..5}
104 do
105     #echo $(ls ${train_file[cnt-1]}/right/*)
106     for f in $(ls ${train_file[cnt-1]}/right/*)
107     do
108         #echo "$f"
109         sed -e 's/[0-9]\{1,\} //' -e 's/[a-zA-Z]\{1,\} /g' -e 's/[[:space:]]\{2,\} /g' -e 's/[[:space:]] //' $f >> $f+copy
110         rm -f $f
111     done
112     echo "ok:完成 ${train_file[cnt-1]} 处理"
113 done

```

- 接着，执行 2，生成 HMM 训练文件
文件格式如下：

```

66 -----
67 HMM file format:
68 -----
69 M= <中文词汇数量: 观察序列>
70 N= <词性标注数量: 隐藏序列>
71 A:<转移概率矩阵>
72 a11 a12 ... a1N
73 a21 a22 ... a2N
74 . . . .
75 . . . .
76 . . . .
77 aN1 aN2 ... aNN
78 B:<发射概率矩阵>
79 b11 b12 ... b1M
80 b21 b22 ... b2M
81 . . . .
82 . . . .
83 . . . .
84 bN1 bN2 ... bNM
85 pi:<初始概率矩阵>
86 pi1 pi2 ... piN
87

```

5. 接着，对检测文件（去除词性标注的文件）进行序列化，格式如下（数字索引是中文词典中单词位置）

```

1 T= 3 <句子中单词数量>
2 279 1157 1508 <每个句子中中文单词对应词典索引位置>
3 T= 2
4 21545 21546
5 T= 32
6 279 1157 1508 27 3544 2334 5 1406 34 1593 43 1182 68 89 11 181 182 11 6720 320 39 12
  15 74 34 83 2253 139 3200 34 1508 340 26
7 T= 27
8 813 11 75 72 5 279 108 1593 18002 133 89 804 483 399 11 82 5 766 183 86 89 11 119 19
  26 1787 97 26
9 T= 16
10 279 27 8 2334 86 11 3544 2334 27 647 8 115 2334 34 907 26
11 T= 27
12 287 115 2334 5 772 62 38 177 89 11 611 977 3440 43 810 47 333 34 976 11 49 15 21547
  34 2334 1508 26

```

6. 接着，采用 Viterb 算法进行词性标注，生成标注文件。

```

1 /nS /n /n
2 /nP /nP
3 /nS /n /n /vC /nR /n /p /b /uJDE /n /wD /n /vN /f /wP /d /v /wP /d /v /cC /
  d /v /uJDE /v /a /n /n /uJDE /n /n /wE
4 /t /wP /rNP /c /p /nS /b /n /n /n /f /v /a /n /wP /cC /p /n /m /n /f /wP /d
  /dD /v /v /wE
5 /nS /vC /a /n /n /wP /nR /n /vC /m /m /qN /n /uJDE /n /wE
6 /rB /qN /n /p /rB /n /vN /n /f /wP /v /a /n /wD /n /uO /n /uJDE /vN /wP /v
  /uA /v /uJDE /n /n /wE
7 /nS /n /n /d /vC /nS /rB /n /n /uJDE /n /wP /d /v /nR /n /wD /nR /n /wD /nR
  /n /wD /nR /n /wD /nR /n /wD /nR /n /wD /nR /n /wP /cC /nR /wD
  /nR /wD /nR /wD /nR /uO /n /n /wE

```

7. 最后，从自己分好的 5 组训练语料中，选择 4 组作为训练，剩余 1 组作为测试语料，进行交叉实验，并记录实验结果。（共进行 5 次）写入到 res.crt 文件中。

```
4389 =====
4390 单词数量:84466.000000, 匹配单词数量:80079.000000
4391 right: 0.948062
4392 =====
```

五、实验结果

训练数据 (组)	测试数据 (组)	训练时间 (秒)	测试时间 (秒)	精确率
2, 3, 4, 5	1	19.2342	5.7018	94.8062%
1, 3, 4, 5	2	18.5099	6.0020	95.3542%
1, 2, 4, 5	3	20.0001	5.8002	95.0031%
1, 2, 3, 5	4	19.4564	5.6000	95.0371%
1, 2, 3, 4	5	19.3453	5.7038	95.0275%

注：测试时间仅为 `time ./showViterbi` 时间

注：训练时间包括文件移动，分割，取出词性标注，生成矩阵时间之和
`time ./ttrain.sh`