人工智能大作业

拼音输入法项目

作业要求

编程实现一个简单的汉语拼音输入法。
实现从拼音（**全拼**）到汉字（**字符串**）内容的转换。

课内讲解

- 整句概率 $P(S)=\prod_{i=1}^{n}P(w_i \mid w_1 \ldots w_{i-1})$

二元语法时：$P(S)=\prod_{i=1}^{n}P(w_i \mid w_{i-1})$

求 $\max\left(\prod_{i=1}^{n}P(w_i \mid w_{i-1})\right)$ 所对应的句子

等价于：

求 $\min\left(-\sum_{i=1}^{n}\log\left(P(w_i \mid w_{i-1})\right)\right)$ 所对应的句子

任务理解&整体思路（以二元模型为例）

1. 任务理解：

   读入一行拼音，输入对应的汉语

   wo shang xue qu le -> 我上学去了

   1. 需要知道拼音可以对应哪些汉字：shang 上商尚伤…
   2. 需要知道两个字在一起的概率 P(学|上)、P(血|上)、P(雪|上)…
   3. 需要得到概率最大的句子 $P(S)=\prod_{i=1}^{n}P(w_i \mid w_{i-1})$

2. 思路：

   1. 拼音对应汉字：建立拼音->汉字map
   2. 两个字连接的概率：词频表与概率模型
   3. 概率最大的句子：动态规划（viterbi算法）

第一步：拼音汉字表的处理

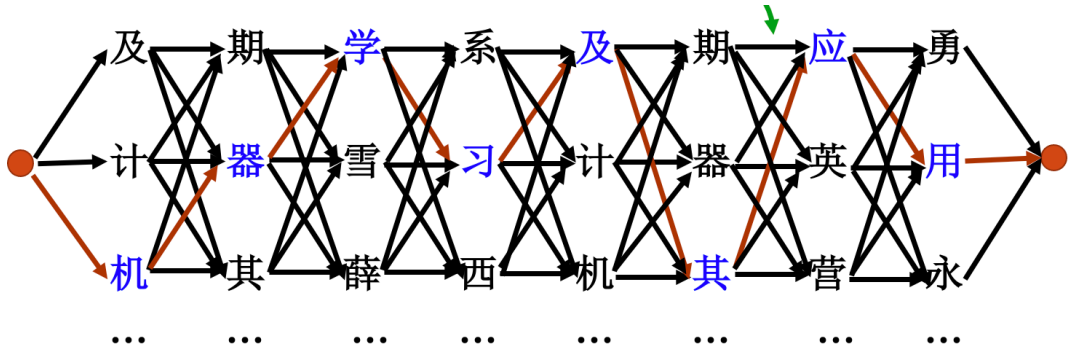
- 读入拼音汉字表及一二级汉字表
- 建立 {拼音:[汉字列表]}的dict（汉字列表可只保留一二级汉字）
- 存储该dict

第二步：使用语料库建立词频表

- 按文件读入语料库
- 处理语料：去除汉字以外的符号（数字、标点等）
  - 循环逐个符号判断？正则表达式抽取？自己决定
- 计算词频：
  - 建立词频dict：{w1w2: count}
  - 每碰到两个字，加入词频dict中/词频dict的count+1
  - 不需要对拼音汉字表中每两个字都建立词频表
- 存储词频表
  - 不需要存储所有的二元词组

第三步：输入法实现

$$P(S) = \prod_{i=1}^{n} P(w_i \mid w_{i-1})$$



- Viterbi算法：
  - 在每一个汉字处，记录：
    - 1. 达到该点的最大概率/Top k大的概率： F(期) = max(F(ji) * P(期|ji))
    - 2. Top k概率对应的前一个汉字
  - 达到结尾时，回溯最大概率对应的路径，得到输出

- 条件概率的计算：
  - P(期|及) ≈ Count(及期) / Count(及)

其他问题&优化与讨论

问题：

1. 编码：gbk编码
2. 第一个字怎么办？
3. 需要存储多少词的词频？怎么选择？
4. 词频为0怎么办？
5. ……

优化与讨论：

1. 时空效率与准确度的trade off
2. 多元字模型、词模型的尝试
3. 语料库扩充
4. 评价指标的分析
5. 多音字处理
6. ……

Thank you!
Q & A

王贝宁

wbn23@mails.Tsinghua.edu.cn