HMM简单拼音输入法实验报告

水工42 张生 2014010330

一、基本思路和实现过程

1. 构建拼音与汉字的对应表

对每个拼音，需要找到相应的汉字。比如：

> a 阿啊呵锕吖腌嗄

2. 基于隐马尔科夫（HMM）的拼音输入法

问题为：

$$\max\prod_{i=1}^{n}P(w_i|w_{i-1})CF(w_i)$$

其中：

$$P(w_i|w_{i-1})=\frac{w_iw_{i-1}\text{同时出现的次数}}{w_{i-1}\text{出现的次数}}$$

为了解决$P(w_i|w_{i-1})$ 可能为0的问题，采用平滑的方法：

$$\lambda P(w_i|w_{i-1})+(1-\lambda)P(w_i)\Rightarrow P(w_i|w_{i-1})$$

实验中，通过选取多组参数确定最佳的λ 的取值。

此外，还有一个问题是，w_{i-1} 也有可能为0. 在我的统计结果中，7272个汉字中以下汉字的出现次数为0：

> 锿 砹 揞 茇 鞴 庳 笾 窆 瘭 醭 骖 楱 镩 榱 鹾 骣 辗 躔 蒇 傺 膪 舡 怛 赕 忉 帱 锦 髑 茳 憨 砘 裰 缍 苊 鲕 镄 篚 唪 鳆 砩 艴 钆 戤 槔 袼 虼 塥 觏 遘 诖 涫 瓯 猓 蜾 鹹 胲 糇 鹱 冱 擐 萑 螅 咴 阍 劐 锪 哜 丌 鐾 裥 鞯 戋 泽 鸺 僬 纟 鲒 骱 赆 獍 刭 弪 僦 鬏 醵 锩 腓 仵 菅 阆 裉 眍 芤 赍 悃 溇 耪 缧 缡 裣 瘵 硇 镥 稆 锷 呒 鞔 矬 猸 镅 蠛 蛘 毪 镎 肭 蝻 穄 旆 堋 仳 擗 螵 缏 氕 肷 锖 靳 劁 愀 吣 赇 彤 膝 鳋 锼 瞍 嗾 谇 唼 彭 胂 矧 铈 锡 耩 慝 掭 韶 酴 腽 芄 阢 饩 荃 蟓 枵 继 砉 痃 茕 獯 曛 屝 阽 蚌 貂 铘 酡 狺 锏 窬 篼 訾 拶 驵 唣 趑 腙 鳜 觜 阼 髓 瘵 嫜 磔 膣 瘛 葑 瘃 窀

显然这些都是非常用字，我基本上一个都不认识。只能直接将概率设为0.

3. 维比特算法

每个拼音对应多个汉字，从而多个拼音对应多层汉字。需要用动态规划的算法寻找最优路径：

1. 对于第一层，不需要求最大路径概率，只需要求该层各个汉字的概率。
2. 对于后面所有层，递推关系式：

$$P(W_{i,j})=\max_k\left(P(W_{i-1,k})\times P(W_{i-1,k}|W_{i,j})\right)$$

其中，$P(W_{i,j})$ 为点$W_{i,j}$ 的最佳路径值，$P(W_{i-1,k}|W_{i,j})$ 为$W_{i-1,j}$ 到$W_{i,k}$ 的发射概率。

4. 统计词频

我们需要知道每一个字出现的次数，每一个词和其他的字同时出现的次数。汉字的个数太多，共七千多个。如果两两组合构建矩阵，大小为7000*7000。当然，这个矩阵是相当稀疏的。可以考虑用字典+字典的数据结构。

首先，已经构造拼音与汉字对应表，由此可以构造pinyin2hanzi词典。

其次，对所有汉字给定编码，构造词典。

再次，对所有汉字统计出现次数。

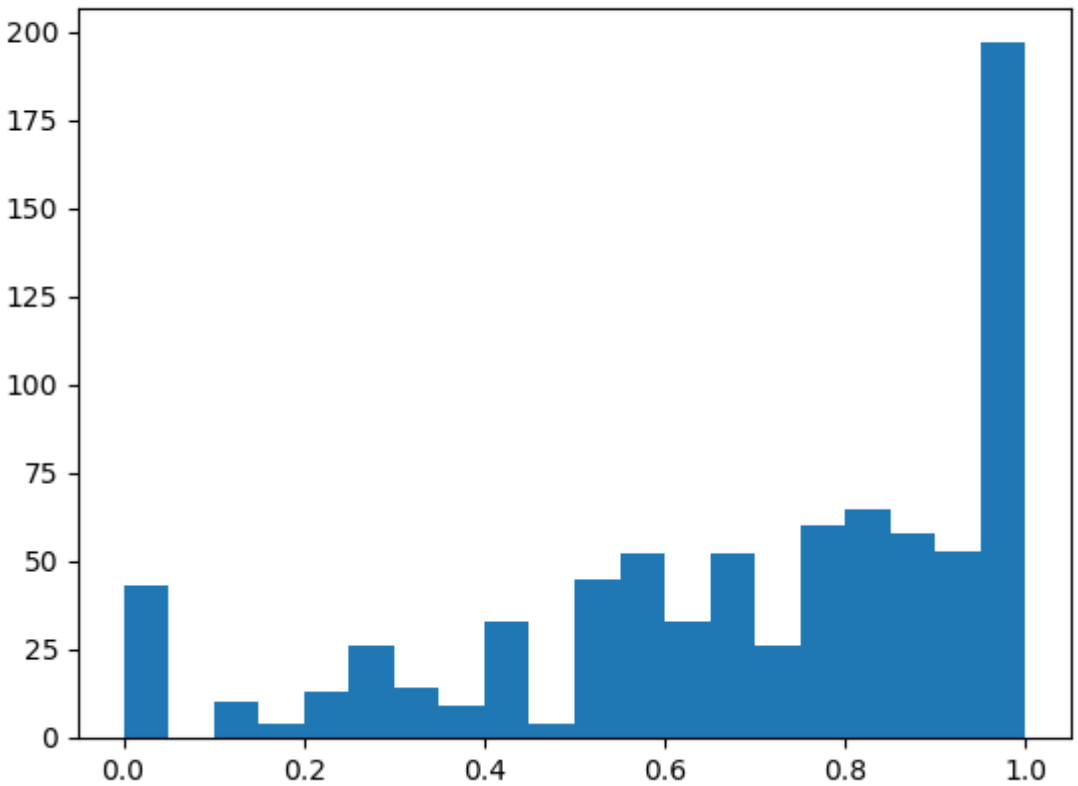
最后，对汉字矩阵统计出现次数。

方法是遍历语料库。每一行的结果需要判断汉字串（可能被各种标点符号切割）。对每个汉字串，遍历即可。

二、实验效果展示

1. 正确率较低的样例

取λ =0.99，对于测试集全集，平均正确率为71.4%。正确率的统计情况如下图：



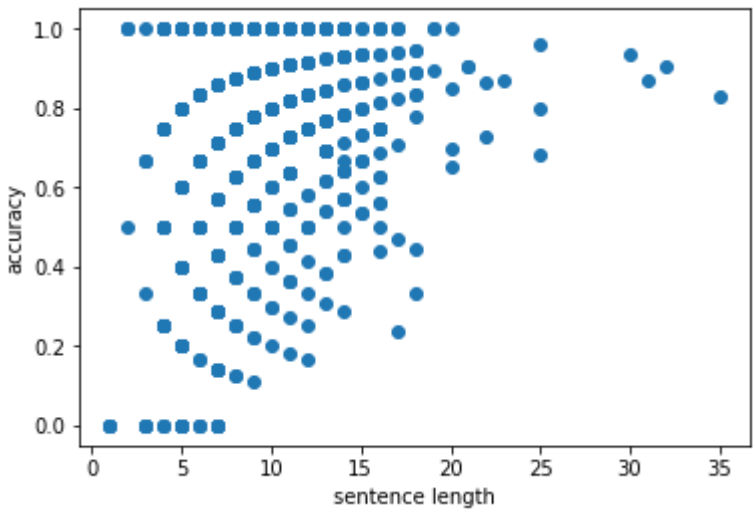
从上图可见大部分的句子正确率较高，但是仍有相当一部分句子正确率很低。下表为正确率在0.2以下的测试样例及输出结果（已经剔除只有一个字的“句子”）。从表可见正确率较低的句子大部分是诗句，可以通过增强这些诗句的权重来改进。另外某些句子的拼音本身就易造成歧义，比如“xiang xiang ji”的输出结果为“想象级”，而正确结果为“香香鸡”。

输入结果	正确结果	正确率
经嗽和大厦也显示认领提	惊搜狐大厦夜现食人灵体	0.181818
他是最阔氏广	她是罪恶之光	0.166667
迹象风不方调篓合收昭山和	既相逢不妨挑灯呵手照山河	0.166667
府网站厮杀场	斧王战死沙场	0.166667
认为深麽舀水较	人为什么要睡觉	0.142857
策影翻沉着孩说	侧影反衬著海水	0.142857
和方殷小切需行	何妨吟啸且徐行	0.142857
务事人肥市食宿	物是人非事事休	0.142857
主播票等都秭归	珠箔飘灯独自归	0.142857
一芟夷山梁静静	一闪一闪亮晶晶	0.142857
立名侨侨华国天便	黎明悄悄划过天边	0.125
烃含四种生情也佛	听寒寺钟声请野佛	0.125
来岜莱巴和向上海拔	来吧来吧互相伤害吧	0.111111
想象级	香香鸡	0
新入名景泰	心如明镜台	0
名警一腓肽	明镜亦非台	0
政部自哐当	争不恣狂荡	0
和叙伦的桑	何须论得丧	0
子失败一圊向	自是白衣卿相	0
岩画像没	烟花巷陌	0
看汛防	堪寻访	0
评省长	平生畅	0
务于茉莉莎	雾雨魔理沙	0
校和财路见建交	小荷才露尖尖角	0
动枫叶芳华签署	东风夜放花千树	0
桂西柳子瘢痕将	贵系六字班很强	0
作业与庶锋舟	昨夜雨疏风骤	0

输入结果	正确结果	正确率
声喊一般选	剩寒意盘旋	0
一孟维码	以梦为马	0
将槲叶玉石碾等	江湖夜雨十年灯	0
五十一道	午时已到	0
策是杨立即	测试样例集	0
构负圭吾乡王	苟富贵毋相忘	0
泰晤拉	太污啦	0
有四阮西嘌醇血	游丝软系飘春榭	0
络虚情展溥修炼	落絮轻沾扑绣帘	0

2. 正确率与句子长短关系

对于每个句子，句子长短和正确率关系如下：



以最长的三个句子为例：

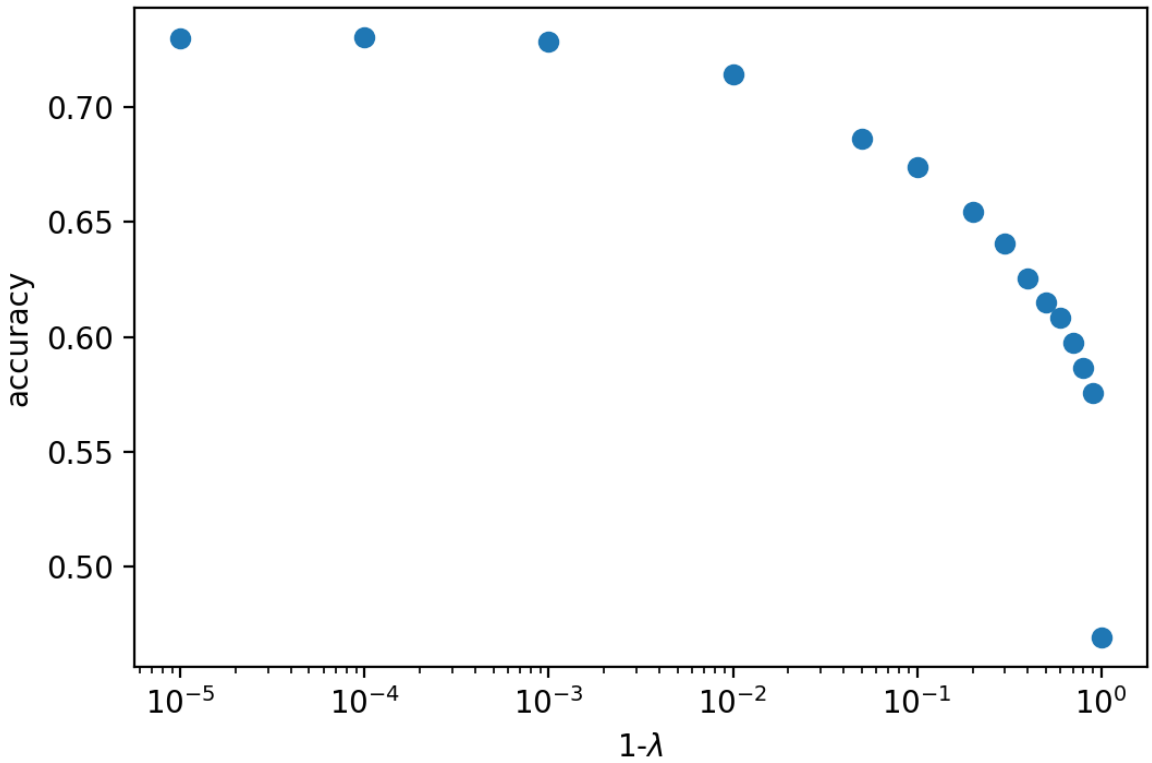
> 输出结果1：安装具有复杂电子控制系统的现代知道无期待提过去相对简单而笨重的火炮系统
>
> 正确结果1：安装具有复杂电子控制系统的现代制导武器代替过去相对简单而笨重的火炮系统
>
> 输出结果2：位于美国家利福尼亚洲旧金山玩的阿尔卡特拉斯岛上的联邦监狱遭到废止
>
> 正确结果2：位于美国加利福尼亚州旧金山湾的阿尔卡特拉斯岛上的联邦监狱遭到废止

可见长句子的正确率也比较高。主要是因为测试集中的长句子词语较多，相互关联性也较强。

三、参数选择及性能分析

我选取了多组λ的取值，并计算其对于测试集全集的正确率。

λ	**accuracy**
0	0.469427
0.1	0.575385
0.2	0.586337
0.3	0.597588
0.4	0.608349
0.5	0.615254
0.6	0.625446
0.7	0.640778
0.8	0.654613
0.9	0.674134
0.95	0.686197
0.99	0.714261
0.999	0.728808
0.9999	0.730258
0.99999	0.730011
1	0.727696



从上表及上图可见，随着λ取值的增加，正确率逐渐增加。在λ超过0.9后，增加得比较平稳。最高值出现在0.9999左右，而之后取0.99999和1正确率都有略微下降。因此，对于本测试集，平滑参数的加入确实能提高正确率，但是提高程度很有限。