拼音输入法作业报告

张晨2017011307

1 算法基本原理

1.1 二元语法

令$f[i][cur]$表示第i个拼音处的汉字是cur的概率，则

$$f[i][cur] = max(f[i-1][pre] * p(cur|pre))$$

其中，$p(cur|pre)$表示在汉字pre之后出现汉字cur的概率，可近似为

$$p(cur|pre) = \frac{cnt(pre, cur)}{cnt(pre, PY(cur))}$$

其中，$cnt(pre, cur)$表示语料库中pre和cur一起出现的次数，$cnt(pre, PY(cur))$表示语料库中pre与读音与cur相同的字同时出现的次数。

1.2 三元语法

令$f[i][pre][cur]$表示第$i-1$，i个拼音处的汉字分别是pre,cur的概率，则

$$f[i][pre][cur] = max(f[i-1][k][pre][cur] * p(cur|k, pre))$$

其中，$p(cur|k, pre)$表示在k,pre一起出现的情况下出现汉字cur的概率，可近似为

$$p(cur|k, pre) = \frac{cnt(k, pre, cur)}{cnt(k, pre, PY(cur))}$$

其中，$cnt(k, pre, cur)$表示语料库中k, pre, cur一起出现的次数，$cnt(k, pre, PY(cur))$表示语料库中k、pre、读音与cur相同的字同时出现的次数。

2 算法实现

2.1 二元语法

2.1.1 算法2-1

语料库中$cnt(pre, cur)$、$cnt(pre, pinyin(cur))$可能为0，故将dp方程调整为

$$f[i][cur] = max(f[i-1][pre] * (\alpha * \frac{cnt(cur)}{cnt(PY(cur))} + \frac{cnt(pre, cur)}{cnt(pre, PY(cur))}))$$

边界条件

$$f[1][cur] = cnt(PY(cur))$$

其中，$cnt(cur)$表示语料库中字cur的出现次数，$cnt(PY(cur))$表示语料库中与字cur读音相同的字的总出现次数，当$cnt(pre, pinyin(cur))$为0方程第二项取0，α是一个超参数，其值取0.01时在测试集上达到最高字正确率60.7%。

2.1.2 算法2-2

$p(cur|pre) = \frac{cnt(pre,cur)}{cnt(pre,PY(cur))}$并不是一个很好的选择字$cur$的指标，如对于“xian lai”，$p$(来|现) = 99.7%，$p$(来|先) = 99.9%，没有明显的区别，但“先来”实际出现的概率远大于“现来”（语料库中“先来”有1408次，“现来”有389次），这导致了算法2-1出现以下错例：

- ming ming shi wo xian lai de
  （输出）明明是我**现来**的
  （参考）明明是我先来的

因此，将dp方程变成

$$f[i][cur] = max(f[i-1][pre] * (\alpha * \frac{cnt(cur)}{cnt(PY(cur))} + cnt(pre, cur)))$$

该算法在$\alpha = 0.1$时达到字准确率71.6%

2.1.3 算法2-3

由于动态规划的边界条件是字频，它对以较为稀有的字开头的句子不太友好，这导致算法2-2出现以下错例

- gu du de ren ta jiu zai hai shang
  （输出）**古都**的人他就在海上
  （参考）**孤独**的人他就在海上

因此，将dp方程的边界变成

$$f[1][cur] = 1(cur\text{的读音是第一个拼音})$$

在$\alpha = 0.1$时，其字准确率达到72.1%.

2.2 三元语法

2.2.1 算法3-1

考虑到$cnt(k, pre, cur)$, $cnt(k, pre, PY(cur))$可能为0的问题，借鉴二元语法算法中的经验，将dp方程写为

$$f[i][pre][cur] = max\Big\{\Big[\beta * \big(\alpha * \frac{cnt(cur)}{cnt(PY(cur))} + cnt(pre, cur)\big) + cnt(k, pre, cur)\Big] * f[i-1][k][pre]\Big\}$$

	guo jia dui	jia dui zhan	dui zhan sheng
对	国家对(2630)	家对战(47)	对战胜(16)
队	国家队(3533)	家队战(12)	队战胜(24)”

表 1: 部分词的出现次数

由于该dp方程状态数过多，对每个位置i，仅保留$f[i][pre][cur]$最大的100个状态。

将超参数设置为$\alpha = 0.1, \beta = 10^{-5}$，得到字准确率83.9%

2.2.2 算法3-2

在算法$3-1$的输出结果中，发现如下错例：

- zhong guo guo jia dui zhan sheng han guo guo jia dui
  （输出）中国国家**对**战胜韩国国家队
  （参考）中国国家**队**战胜韩国国家队

如表1所示，“国家对”，“家对战”，“对战胜”在语料库中的出现次数都较多，然而尽管“国家队”“队战胜”出现次数略多，但“家队战”的出现次数明显小于“家对战”，这导致了模型倾向于在这个位置选择“对”字，因此，需要设法增加“词组”的权重。

我的做法是将jieba分词dict.txt.small文件中的词的出现次数都乘以一个权重，调参后得二元词权重*8，三元词权重*16后，$\alpha = 0.1, \beta = 10^{-5}$时字准确率84.2%。

2.2.3 算法3-3

一般而言，如果几个相连的字在语料库中有极高的出现次数，那么即便它们和在它们之前、之后相邻的字一同出现的次数不多，它们也应当被优先考虑放到这个位置。因此我认为，选择某个词的概率与在语料库中的出现次数之间的关系不是线性的，而是一个下凸函数。我用幂函数$g(x) = x^k$(x表示语料库中的词频，k是需要调整的超参数)代替方程中的词频项，在三元词$k_3 = 1.5$，二元词$k_2 = 1$时获得最优准确率84.7%

2.3 多音字的处理

2.3.1 使用谷歌翻译

在实现以上算法之前，我曾尝试寻找处理为多音字注音的软件，但没有发现合适的。因而，对于存在多音字的词，我只能把各个可能读音下的词频都视为语料库中的出现次数，这势必会导致模型出现错误，如将”shou lian”(收敛)识别为“熟练”、”jiao xiu”(娇羞)识别为“脚臭”、”ju zi”(句子)识别为“车子”等，故我尝试对多音字进行粗略处理。

谷歌翻译会对输入的中文进行注音，故我将出现次数较多的二元词打印出来，用谷歌翻译进行注音。如果一个三元词的前两个字或后两个字出现在了我

的注音集合里，那么这两个字的发音便以注音为准。然而谷歌翻译仅有段落注音的接口，没有词注音的接口，故我只能将一大段词用空格分离组成段落再进行注音，这时对某个词的注音可能会受到相邻词的干扰，因此有一些错误。

2.3.2 使用pypinyin

之后，有同学向我推荐了pypinyin，一个可以方便进行词注音的库。我使用此库对所有的二元、三元多音词进行了注音，再对参数进行微调，在$\alpha = 0.1$，$\beta = 10^{-6}$，jieba三元词词频*16，jieba二元词权重*8，将三元词权值视为其词频的1.5次方时，得到了85.4%的准确率。

3 结果分析

3.1 二元、三元方法对比

3.1.1 输出结果

以下是一些使用二元语法输出错误，但使用三元语法输出正确的句子。上面一句为二元的输出，下面一句为三元的输出。

- ni wei shen me lai wan le
  你为什么来**玩乐**
  你为什么来**晚了**

- xiao peng you men dou xi huan qu jiao you
  小朋友们都喜欢去交由
  小朋友们都喜欢去郊游

- ben zhan ji zhe pao de kuai
  本站记者跑的快
  本站记者跑得快

3.1.2 分析

在校正拼音后，二元语法的字准确率为76.1%，三元语法的最终字准确率为85.4%。可见三元语法使准确率有了大幅提升，它成功的将“跑的快”校正为“跑得快”，对“来晚了”，“去郊游”等三元词的识别率也明显增高。

3.2 缺陷分析

1. 语料库选择问题。由于语料库来自于新浪新闻，模型倾向于输出一些新闻中常出现的词，如将“句子”输出为“巨资”。

2. 测试集可靠性不好。测试集是前几届学生合编的，其中包含一些专业词汇及诗句等。由于没有使用相应的语料库，模型对这些句子的处理能力极为有限，如会将“函数”输出为“寒暑”。

3. 缺乏对语义的理解，如无法正确输出“他”“她”“它”。
   • ta shi wo de mu qin → **他**是我的母亲
   • ta shi wo de gou gou → **他**是我的狗狗

4. 较为短视，对长词组的处理效果不佳，也会产生每个小段都相对合理，但连起来十分奇怪的句子，如第二个句子，“香港市最繁华的”与“最繁华的大都市”都很合理，但连起来却不对。
   • mei li jian he zhong guo (美利坚合众国) → 美丽建和中国
   • xiang gang shi zui fan hua de da du shi zhi yi → 香港市（应当为“是”）最繁华的大都市之一

5. 没有音调带来的问题
   • yi qi zai huang dao shang deng dai xing chen
   一起在**黄道**上等待星辰
   一起在**荒岛**上等待星辰

3.3 进一步尝试的方向

1. 选取更大、更全面的语料库。
2. 考虑四元、五元甚至更长的语法。
3. 导入专有名词的词库。
4. 使用基于词的算法。
5. 在分析相邻词的基础上，分析整句中各个词之间的关系。

4 demo

使用pyqt5实现了一个输入法小程序，启动后可之间通过键盘进行输入，另有如下操作：

1. 通过按键1-9在候选列表内进行选择
2. 若候选列表多于一页，则可以使用加号、减号进行翻页
3. 使用shift键将输入框里的内容直接加入文本
4. 使用back space键删除输入框、文本框内的最后一个字符

该小程序是基于作业的功能进行开发的，因而只能支持选出最好的一个选项，而不支持选出最好的若干个选项。但在输入仅为一个字的拼音的时候，它会显示该拼音对应的所有字。该程序demo见`https://cloud.tsinghua.edu.cn/f/c1e4584712474988a0b2`

图 1: 输入法demo截图

5 程序运行说明

5.1 源程序

本机的环境为python3.7.0，理论上也支持python3的其他版本。运行命令为python3 pinyin.py input.txt output.txt

程序需要约1分钟的时间加载词频模型，望助教耐心等待。

5.2 demo程序

在根目录demo文件夹下，运行命令为python main.py，需要使用PyQt5，程序启动可能需要花费较长时间。

附录

A 参数调整过程记录

A.1 算法2-1

见表2

α	0.001	**0.01**	0.1	1	10	100
字准确率	0.606	**0.607**	0.606	0.606	0.601	0.511

表 2: 算法2-1准确率与参数α的关系

A.2 算法2-2

见表3

α	0.001	0.01	**0.1**	1	10	100
字准确率	0.716	0.716	**0.716**	0.715	0.715	0.709

表 3: 算法2-2准确率与参数α的关系

A.3 算法2-3

见表4

α	0.001	0.01	**0.1**	1	10	100
字准确率	0.721	0.721	**0.721**	0.720	0.720	0.720

表 4: 算法2-3准确率与参数α的关系

A.4 算法3-1

参数α表示一元词与二元词之间的权重关系，可直接沿用算法2-3中的$\alpha = 1$，参数β衡量了算法2-3得到的二元结果与新加入的三元词之间的权重关系，是该算法调参的主要调整对象，其对准确率的影响如表5所示。

A.5 算法3-2

三元词的词频对模型的影响较大，故先调整三元词的词频，结果如表6所示，在权重为16时达到最优准确率84%。在此基础上，进一步调整二元词所乘权重，所得结果见表7。

A.6 算法3-3

首先调整三元词的幂次k_3，结果见表8 之后调整三元词的幂次k_2，发现准确率几乎不变，因此取$k_2 = 1$。

β	1e-8	1e-7	1e-6	**1e-5**	1e-4	0.001
字准确率	0.836	0.836	0.838	**0.839**	0.836	0.826

表 5: 算法3-1准确率与参数β的关系($\alpha = 0.1$)

权重	1	2	4	8	**16**	32
字准确率	0.839	0.840	0.840	0.841	**0.842**	0.842

表 6: 算法3-2准确率与三元词所乘权重的关系

B 最终模型的部分输出

B.1 正确输出

- shen du xue xi ji shu tui dong le ren gong zhi neng de fa zhan
  深度学习技术推动了人工智能的发展
- wo he wo de xiao huo ban men dou jing dai le
  我和我的小伙伴们都惊呆了
- xin ru ming jing tai
  心如明镜台
- re xue nan er dang zi qiang
  热血男儿当自强
- wei ji fen shi xue bu hao de
  微积分是学不好的
- qi yin huo fu bi qu zhi
  岂因祸福避趋之
- la tiao shi yi zhong you ming de xiao chi
  辣条是一种有名的小吃
- gou jian she hui zhu yi he xie she hui
  构建社会主义和谐社会
- quan qiu hua wen ti xu yao quan qiu ren xie shou jie jue
  全球化问题需要全球人携手解决
- chi pu tao bu tu pu tao pi
  吃葡萄不吐葡萄皮

B.2 错误输出

- suo yi zhe liang ge han shu ji hu chu chu xiang deng
  所以这两个**汉**数几乎处处想等
  所以这两个**函**数几乎处处相等

权重	1	2	4	**8**	16	32
字准确率	0.842	0.842	0.846	**0.846**	0.845	0.842

表 7: 算法3-2准确率与二元词所乘权重的关系（三元词已乘权重16）

k_3	1	1.1	1.3	1.5	**1.7**	2
字准确率	0.846	0.846	0.847	0.847	**0.849**	0.848

表 8: 算法3-3准确率与三元词幂次的关系

- jiu xiang lei shui xiao shi zai yu zhong
  就像泪水**小时**在雨中
  就像泪水**消失**在雨中

- gou li guo jia sheng si yi
  苟利国家生死一
  苟利国家生死**以**

- mei ren ti gong zhi shao shi ge ju zi
  每人提供至少**是个巨资**
  每人提供至少**十个句子**

- dan shi long cheng fei jiang zai
  但**是**龙城飞将在
  但**使**龙城飞将在