Introduction to AI
Pinyin Input Method Report

姓名: 任一

学号:2018011423

ry18@mails.tsinghua.edu.cn

2020 年 4 月 5 日

实验环境	
操作系统:	Windows10 家庭版 18362.72
Python 版本:	Python 3.7.3 64-bit

1 实验概述

本次实验我使用了字的二元模型，结合 Viterbi 算法，得到了拼音到汉字的转换程序。在同学们共同制作的测试集上逐字准确率为 79.3%, 逐句准确率为 31.2%.

此外，我还根据今年的时事热点，爬取了今年 1-3 月人民日报的内容，加入语料库训练，从而能够对当前时事热词表现出较高的识别精度。[1]

2 实验思路

本实验主要任务是,给定一个拼音的输入序列 $l_1l_2l_3...l_n$,求得汉字序列 $w_1w_2w_3...w_n$,使得 $P(w_1w_2w_3...w_n)$ 最大, 其中 w_i 为拼音 l_i 对应的一个汉字。由概率的链式法则,$P(w_1w_2w_3...w_n) = P(w_1)\prod_{i=2}^{n} P(w_i|w_{i-1}w_{i-2}...w_1)$. 再由二元模型用到的一阶马尔科夫假设，可得

$$P(w_1w_2w_3...w_n) = P(w_1)\prod_{i=2}^{n} P(w_i|w_{i-1}w_{i-2}...w_1) \approx P(w_1)\prod_{i=2}^{n} P(w_i|w_{i-1}) \tag{1}$$

其中 $P(w_i) = \frac{\#\{w_i\}}{len(corpus)}$, $P(w_i|w_{i-1}) = \frac{\#\{w_{i-1}w_i\}}{\#\{w_{i-1}\}}$.

因此，二元模型只需要计算语料库中的单字和二元词得出现次数即可得到上述公式所需的概率。

但在实现中,还需要进行平滑处理。平滑处理的原因在于,若在计算某汉字序列的概率 $P(w_1w_2w_3...w_n)$ 时，其中 w_i 在语料库中出现的次数为 0，在公式(1)的连乘当中，就会导致这个序列的概率为 0。这个现象是不合理的，因为出现某一项为 0 时，其他项的概率" 贡献" 会被清零，整个序列的概率仅决定于 w_i 没有出现，这种情况是不合理的。因此通过资料的查找，我使用了 Add-k smoothing [2] 来解决这个问题。即使用下面的公式来近似 $P(w_i|w_{i-1}) = \frac{\#\{w_{i-1}w_i\}}{\#\{w_{i-1}\}}$

$$P^*_{\text{Add-k}}\ (w_i|w_{i-1}) = \frac{\#\left(w_{i-1}w_i\right) + k}{\#\left(w_{i-1}\right) + kV}$$

其中 V 为可能出现的字的数量，k 为平滑系数。

此外，为了防止公式(1)中连乘出现数值下溢，我采用了对数概率，即对公式(1) 两边取对数，计算对数概率的和。由于对数函数的单调性，这样操作正确性可以得到保证。

给定输入拼音序列 $l_1l_2l_3...l_n$，如何选取到 $\arg\max_{w_1w_2...w_n} P(w_1w_2w_3...w_n)$ 也是有一定技巧的。记 N 为平均每个拼音对应的汉字数量，T 代表输入拼音序列 (待查询汉字序列) 的长度，如果我们采用枚举法，选取最佳汉字序列的复杂度为 $O(N^T)$，这样高的复杂度在拼音输入法这样对速度要求较高的场景下是不适用的。因此我采用 Viterbi 算法，利用截止到每个拼音的最佳汉字序列仅与前一个拼音有关的性质，可以将复杂度降低到 $O(N^2T)$，从而能够高效地得到结果。

[1]爬取人民日报的程序来源于`https://github.com/ppmm/get-people-daily`，由于爬取的结果为 PDF 格式，我还参考了`https://blog.csdn.net/xc_zhou/article/details/81009809` 中关于 PDF 转文字的相关实现，最终将 2020 年 1 月至 2020 年 3 月的人民日报内容转换为 txt 文件并运用于模型训练。该语料信息存储于`https://cloud.tsinghua.edu.cn/f/3ae6f6d6f7c847abb827/`

[2]此处参考 Speech and Language Processing Chapter3 N-gram Language Models. Daniel Jurafsky & James H. Martin. Copyright c 2019. All rights reserved. Draft of October 2, 2019.

3 分析实例

3.1 好的实例及其分析

1. zu guo wan sui qing nian jia you
   祖国万岁青年加油
2. zhong guo gong chan dang ren de chu xin he shi ming jiu shi wei zhong hua ren min mou xing fu wei zhong hua min zu mou fu xing
   中国共产党人的初心和使命就是为中华人民谋幸福为中华民族谋复兴
3. xin xing guan zhuang bing du
   新型冠状病毒
4. yi qing jiu shi ming ling fang kong jiu shi ze ren
   疫情就是命令防控就是责任
5. wei shen me ao da li ya de sen lin da huo shao le si ge yue dou mei neng bei pu mie
   为什么澳大利亚的森林大火烧了四个月都没能被扑灭
6. jin tian shui guo gao te jia
   今天水果搞特价

由上面较好的例子可以看出，我设计的拼音输入法总体性能优异，在长句中也能表现较好，如例 2 例 5。此外，由于加入了近 3 个月的人民日报语料，本输入法也能较好地识别新的一些说法，如例 3 例 4. 另外虽然训练语料基本都是新闻内容，例 6 也展示出了本输入法一定的泛化能力，能够在一些生活场景中有较好表现。

3.2 不好的实例

1. wo hao xi wang neng gou hui dao xue xiao
   错误：**握**好希望能够回到学校
   正确：我好希望能够回到学校
2. jin tian hui jia bi jiao wan
   错误：今天回家比较**完**
   正确：今天回家比较晚
3. ji qi xue xi de lu bang xing
   错误：机器学习的**绿邦刑**
   正确：机器学习的鲁棒性
4. tuan yuan de chun jie
   错误：**团员**的春节
   正确：团圆的春节
5. qi qu de shan lu
   错误：**起去**的山路
   正确：崎岖的山路

上述实例体现了本输入法的一些局限性。我将从语法、词库、语境三方面进行分析。

从语法上讲，词的二元模型甚至多元模型都没有明确考虑语言在语法层面的性质。如例 1，从语法角度来看以“我”开头明显是更合适的选择，例 2 中以“比较完”结束句子也是从语法上明显不合理的。

从词库上讲，本输入法模型较为依赖词库的丰富性。如例 3 所示，机器学习的鲁棒性在学术语境下是非常自然的，然而由于词库主要是新闻，学术方面的词汇就不会学习到。

从语境来讲，词的二元模型对语言的前后关联考虑不足。偏正短语 [3] 尤其体现了这一点，可以看到例 4 例 5 的偏正短语，都是较为自然的说法。但是由于词的二元模型仅考虑了两个字的相互关联，“的”前面的修饰语和后面的中心语之间就不能产生很好的关联，以至于修饰语和中心语不搭配的情况很有可能出现。

4 参数调整与分析

在本部分，我主要对两种平滑方式中的参数调整，并分析结果。

4.1 Add-k Smoothing

在第 2 部分实验思路中，我已经详细介绍了 Add-k Smoothing. 公式如下：

$$P^{*}_{\text{Add-k}}\left(w_i|w_{i-1}\right)=\frac{\#\left(w_{i-1}w_i\right)+k}{\#\left(w_{i-1}\right)+kV}$$

其中 V 为可能出现的字的数量，k 为平滑系数。

本部分将调节 k 的大小，并说明 k 的大小对实验结果的影响，实验结果图表如下：

表 1: Add-k Smoothing

k	Sentence Accuracy	Character Accuracy
10^{-50}	0.3064	0.7961
10^{-40}	0.3064	0.7961
10^{-30}	0.3064	0.7961
10^{-20}	0.3064	0.7961
10^{-10}	0.3064	0.7961
10^{0}	0.302	0.793
10^{1}	0.3036	0.786
10^{2}	0.2758	0.769
10^{3}	0.2256	0.7497

从图 1 和表 1 中可以看到，当 k 值小于 1 时，模型表现都非常稳定，当 k 大于 1 时，模型性能会出现下降。这可能是由于，当 k 大于 1 时，平滑项 ($\frac{k}{kV}$) 的权重会进一步增大，甚至在一定程度上掩盖原有的概率项 ($\frac{\#(w_{i-1}w_i)}{\#(w_{i-1})}$), 这样就很有可能造成一些偏差。考虑极端情况 $\lim_{k\to+\infty}\frac{\#(w_{i-1}w_i)+k}{\#(w_{i-1})+kV}=\frac{1}{V}$, 这样就完全消除了原有概率项的影响，明显是不合理的，

[3] 偏正短语又叫偏正词组，是由修饰语和中心语组成。例如“好看的姑娘”、“前进的方向”、“强劲的性能”都是偏正短语。

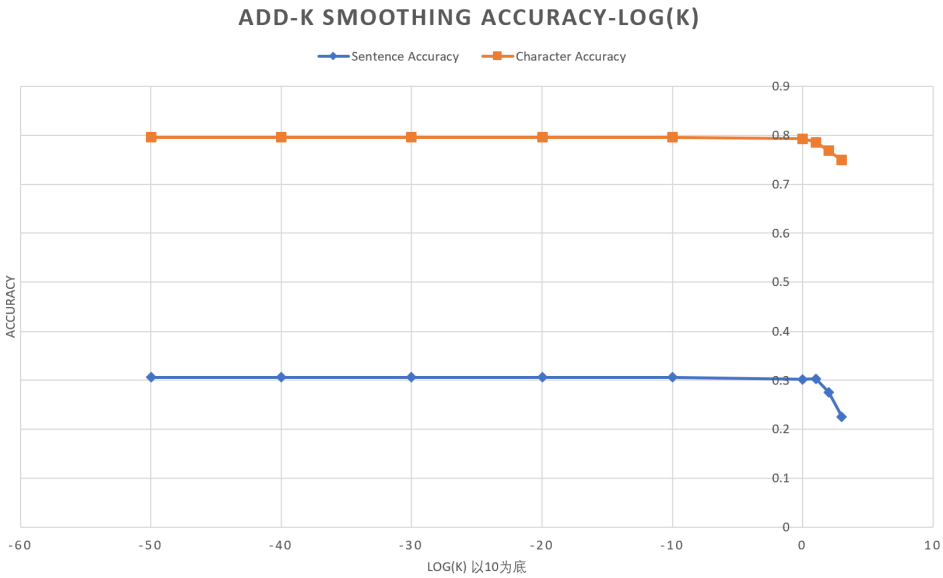

图 1: Add-k Accuracy - log(k)

4.2 Small Probability Smoothing

除了之前介绍的 Add-k Smoothing, 另外一种简便的平滑做法是，对于 $P(w_i|w_{i-1}) = \frac{P(w_{i-1}w_i)}{P(w_{i-1})}$, 若 $P(w_i) = 0$ 或 $P(w_{i-1}w_i) = 0$，则直接令 $P(w_i|w_{i-1}) = \alpha$, 其中 α 为一个较小的常数。对于这种方法，我调整参数 α 的值，得到实验结果图表如下：

表 2: Small Probability Smoothing

α	Sentence Accuracy	Character Accuracy
10^{-50}	0.3064	0.7961
10^{-40}	0.3064	0.7961
10^{-30}	0.3064	0.7961
10^{-20}	0.3064	0.7961
10^{-10}	0.3064	0.7961
10^{0}	0	0.0903
10^{1}	0	0.0903
10^{2}	0	0.0766
10^{3}	0	0.059

从图 2 和表 2 中可以看出，当 α 的值小于 1 时，模型表现都很稳定。但当 $\alpha >= 1$ 时，模型性能出现断崖式下跌。这令我思考参数 α 的含义。我认为 α 体现了模型对于未在语料库中出现的字或词的一种“惩罚”程度。对于某个拼音输入序列，一种包含生僻字的汉字序列的可能就会被 α“惩罚”，且 α 越小，“惩罚”程度就越大。当 $\alpha >= 1$ 时，由于正常字词的概率值小于 1，这种“惩罚”甚至变为了“奖励”，即在一定程度上鼓励了生僻字在最后结果中的出现，这样的情况明显不符合常识，也因此解释了为什么 $\alpha >= 1$ 时模型性能断崖式的下跌。

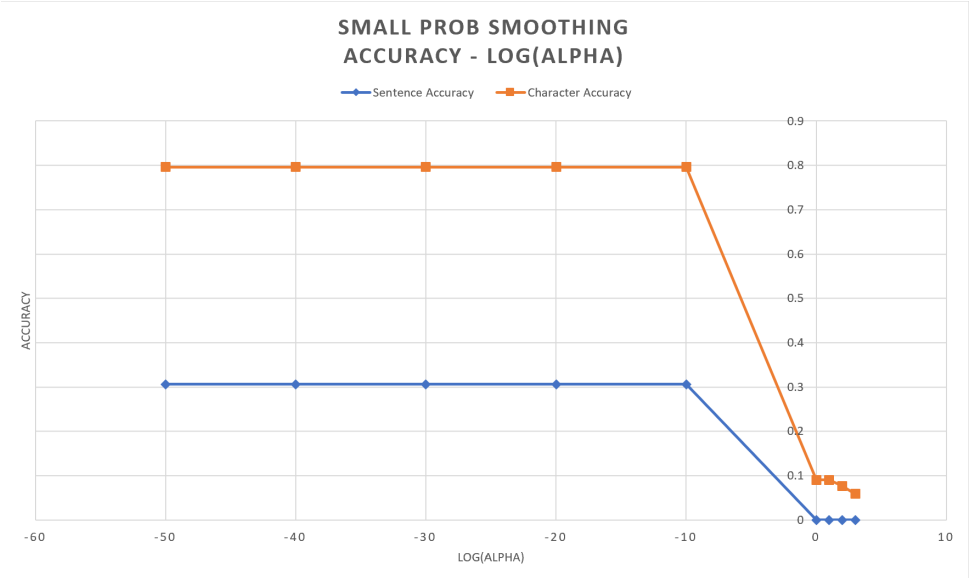

图 2: Small Prob Accuracy - log(α)

5 模型对比

为了提升原始语料库 (仅含新浪新闻) 对于 2020 年新词的识别率，我自行爬取了 2020 年 1-3 月的人民日报语料信息形成了第 2 个语料库，并且对新浪新闻和人民日报语料信息做 jieba 分词，将分词后的语料以二元词的形式加入词频词典形成了第 3 个语料库。下面我将对比这 3 个语料库的表现。

表 3: Different Model Accuracy

	Sentence Accuracy	Character Accuracy
Sina	0.312	0.7930
Sina+People	0.312	0.7930
Sina+People+Jieba	0.312	0.7936

从图 3 和表 3 可以看出，加入人民日报语料和 Jieba 分词后的词汇对模型整体提升并不大。对这个结果我的理解如下：助教提供的新浪新闻语料库大小为 948MB，而我爬取的近 3 个月人民日报语料仅有 14.3MB. 因此加入后提升不会非常大。此外由于同学们收集的语料主题和内容较为分散，没有很多时事内容，因此也没有很大提升。

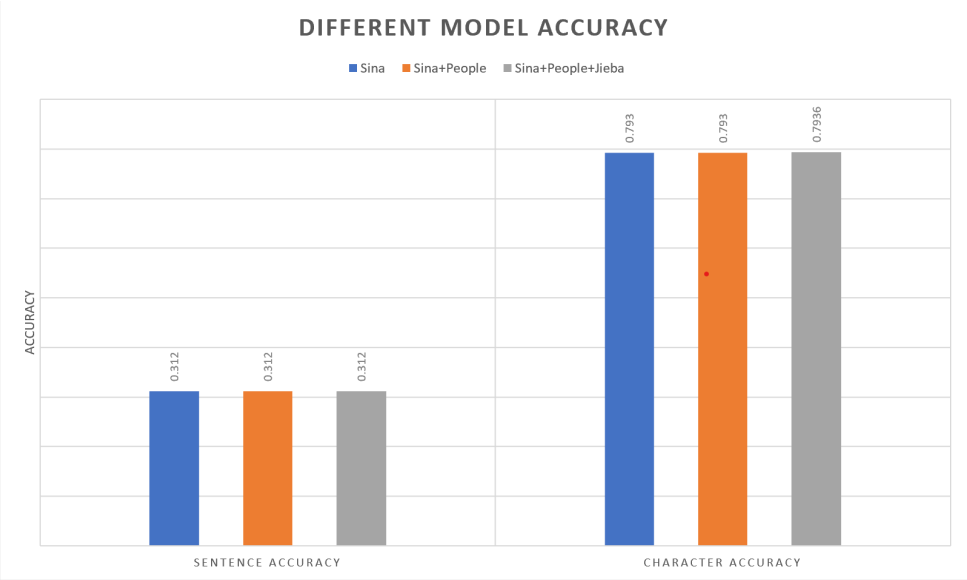

图 3: Different Model Accuracy

下面我会针对一些时事热点词汇，对比加入人民日报语料前后的模型识别情况。

1. xin xing guan zhuang bing du
   加入前：新型**灌装**病毒
   加入后：新型**冠状**病毒
2. yi qing jiu shi ming ling fang kong jiu shi ze ren
   加入前：疫情就是命令**防空**就是责任
   加入后：疫情就是命令**防控**就是责任
3. xin guan fei yan yi qing
   加入前：**信管**肺炎疫情
   加入后：**新冠**肺炎疫情
4. fu gong fu chan
   加入前：**府共服产**
   加入后：**复工复产**
5. ying xiong de cheng shi ying xiong de ren min
   加入前：英雄的城市英雄的人民
   加入后：英雄的城市英雄的人民
6. dao qiong si gong ye zhi shu da fu xia die
   加入前：道琼斯工业指数大幅下跌
   加入后：道琼斯工业指数大幅下跌
7. mei guo gu shi chu fa rong duan ji zhi
   加入前：美国故事处发熔断机制
   加入后：美国故事处发熔断机制

从上面的测试样例的例 1-4 中可以看出，新模型对于疫情相关的信息表现出了非常好的适应性。虽然新补充的语料较少，但也能够在某方面极大提升准确性，这也为拼音输入法的增量更新与改进

提供了可能性。而例 5、6 这样语法结构较为简单，说话方式符合正常习惯的句子是否改进表现都不错。对于例 7 这样的样例，我分析可能是由于人民日报对于国外股市报道较少，最近集中于疫情的报道，故没有产生明显的改进。

6 可能的改进方案

针对 3.2 不好的实例中体现出的问题，我想到的可能的改进方案如下：

从语法上讲，拼音输入法应该在一定程度上考虑语法的合理性。比如可以对字词做一定的语法分析，例如“你”、“我”、“我们”等词是主语，“前往”、“上学”等词是谓语。并且增大主谓宾、主谓等特定语法结构出现的可能性，以增强拼音输入法对合理语法的接受度。

从词库上讲，针对用户使用场景和偏好的不同，应该增强特定领域的词库。很多成熟的拼音输入法也会在用户安装时，建议用户选择自己偏好和常用的词语方向。此外，针对时事热点，也应该定期更新词库。面对用户的输入习惯，拼音输入法也应该有自我学习的能力，讲用户常输入的一些词加入词库中。

从语境上讲，拼音输入法应加强词语之间的关联度。可以考虑采用三元、四元等模型来丰富二元模型的表达能力，也可以考虑在某些特定词汇间建立更强的联系，例如“喜庆”这一词就和“春节”这一词的联系应该较为紧密，以增强拼音输入法在特定语境下的表达能力。

7 总结

在本次实验中，我利用 Viterbi 算法和一阶马尔科夫假设，基于大量的语料信息完成了拼音输入法的设计。此外还定向加入了近期时事新闻以提高当前热词的识别率。在实现的过程中锻炼了自己的工程能力，在分析结果的时候锻炼了自己的思维能力。我感觉到，很多任务的乐趣和意义不仅仅来源于调参优化性能，还在于认真分析结果，对现有实验结果提出解释和猜想，并通过进一步的实验来验证或者反驳自己的猜想，在这样理论和实践的反复验证和反驳的过程中培养科学思维。感谢老师和助教给予我们的悉心指教和帮助！