拼音输入法实验报告

孙迅　2019011292

一、算法简述

输入法主体采用动态规划算法实现。在尝试了字二元语法和字四元语法模型后，最终选择了词二元语法辅助的字二元语法模型。

第一部分　预处理

对语料进行注音和分词。

统计语料中所有相邻 k (=1, 2, 3)字组出现的次数，连同注音存于 `char_cnt.txt` 中。

统计语料中出现的所有词语，连同注音存于 `words.txt` 中。

统计语料中所有词语出现的次数，连同注音存于 `word_cnt.txt` 中。

统计语料中所有相邻两词出现的次数，连同注音存于 `phrase_cnt.txt` 中。

第二部分　字二元语法

在字二元语法模型中，某个音对应某个字的概率，取决于已算得的前一个音的概率情况。具体而言，第i个音对应字c的概率

$$p[i][c] = \max(p[i-1][x] * P(c|x))$$

其中，$P(c|x)$的计算公式为

$$P(c|x) = \lambda \frac{\mathrm{cnt}(xc)}{\mathrm{cnt}(x)} + (1-\lambda)\frac{\mathrm{cnt}(c)}{\text{总字数}}$$

从前至后处理所有音节。对每一个音，根据前一个音的概率情况，按上述公式，逐个计算当前所有候选字的概率情况，并记录各自对应的祖先。

处理完成后，从最后一个音的概率最大的候选字出发，沿着记录的祖先回溯，即可得到转化的结果。

第三部分　词二元语法辅助

在字二元语法模型的基础上，引入词二元语法模型辅助，以期得到更准确的结果。

使用词二元语法模型的基本动机是，若某几个连续的候选字可以组合成词，则通过少乘若干次条件概率，使得以该词结尾的概率有所增加，从而创造“捷径”。此外，在条件概率的计算公式中，对于构成“词组”的，适当提高其权重。通过这种方式，转化的结果将更倾向于体现词语之间的连贯性。

具体而言，词c的概率由向前数第len c个音的概率情况决定，条件概率的表达式则改为

$$P(c|x) = k_1 \frac{\text{词组}xc\text{出现的次数}}{\text{词}x\text{出现的次数}} + k_2 \frac{\text{字的组合}xc\text{出现的次数}}{\text{字的组合}x\text{出现的次数}} + k_3 \frac{c\text{出现的次数}}{\text{总数}} (\mathrm{len}\ c)^{\alpha}$$

其中，参数k_1，k_2，k_3根据c和x是字还是词的不同情况确定相应的取值。

二、效果展示

好的例子

机器学习及其应用

你的理解是对的

我去给你买一个橘子

我和一个队员换个位置坐在他的旁边

经济建设和文化建设突出了十八大精神的重要性

取消或停征的涉及个人等事项的行政事业性收费包括

可以看到，该算法不仅对于政务类语句有良好的转化效果，也能对生活化的语句进行准

确的转化。此外，它能根据相邻的词语，对同音异词的情况进行较好的修正。

差的例子

最后他把握送回宿舍

我们做了大概二十六小时的硬卧

他说详情我去喝一杯

给阿姨到已被卡布奇诺

效祖国和定居用（小卒过河顶车用）

创新将可行实际内容（创新讲课形式及内容）

从中可以看出，该算法无法对句子整体的语法结构进行把握，出现了成分错误或残缺的问题。（1、4）

其次，对于间隔稍远的词语之间的匹配，该算法无能为力，出现了搭配不当的问题。（2）

此外，词二元模型的引入增加了对已知的词语搭配的依赖性，容易造成矫枉过正的情况。（3、6）

最后，对于字词较为凝练的词组或习语，该算法往往难以进行准确的转化。（5）

三、性能比较

测试样例来源：一些群聊记录 ＋ 自己想的正常的句子 ＋ 课程征集 ＋ 出师表

与纯字二元模型和字四元模型比较

比较项目	字二元模型	字四元模型	字二元+词二元模型
字准确率	3644/4475=81.43%	3936/4475=87.96%	4018/4475=**89.79%**
句准确率	150/460=32.61%	241/460=52.39%	276/460=**60.00%**
单句平均处理时长	**57ms**	1565ms	363ms

与字二元模型相比，词二元模型在字准确率和句准确率方面均有显著的提高。而与字四元模型相比，词二元模型虽然准确率并无显著的提升，但处理时长降低到了可以接受的水平。

从转化效果上看，字二元模型由于考察范围过于狭窄，时常给出一些不合理甚至荒谬的搭配（如激动车、消化矛）。字四元模型虽然有了更宽的考察范围，但不能很好地从词语的角度进行匹配，因而可能给出一些看似连贯、实则不合语法的句子（如你的理解释对的、测试样力争机），以及一些离谱的词组搭配（如小熊丙肝）。

最终选择的词二元语法辅助的字二元语法模型能在一定程度上解决上述问题，因此是较为理想的解决方案。

调整参数进行比较

在最终的算法模型中，共有 8 个参数。

类别	待匹配的是字			
子情形	词＋字	成词单字＋字	字＋字	字频
权重	12	5	6	2
类别	待匹配的是词			
子情形	词＋词	成词单字＋词	字＋词	词频
权重	700	400	240	50

在这里，不对参数的选取过程进行具体阐述，而是调整参数的取值，简析其影响。

项目	调整至	准确率	典例
不调整	—	89.79%	—
词＋字	0	89.16%	北京市一个美丽的城市
	100	89.52%	背景是一个美丽的城市
成词单字＋字	0	89.69%	拼音输入法有点难些

	50	89.18%	实干才能出城就
字＋字	0	89.87%	油腻好果子吃的
	50	86.84%	你的理解释对的
字频	0.01	89.61%	握好害怕
	20	85.77%	今天回家比较完
词＋词	0	89.09%	及其学习及其应用
	10000	89.61%	维基百可是一个网络百科全书项目
成词单字＋词	0	89.47%	我从未见过犹如此后颜无耻之人
	10000	89.41%	解下来的一年
字＋词	0	89.69%	南京市场江大桥
	10000	88.87%	不在华夏
词频	0.1	87.28%	信力挺难受的
	10000	84.27%	拼音输入法有点男鞋

诚然，单个参数的变化并不足以反映全貌，也不足以对准确率造成太大影响。但通过上面的对比，这些参数对转化结果确实带来了相应的改变。例如，降低“字＋字”的权重，则“有你好果子吃的”的前两个音倾向于被处理为一个完整的词“油腻”；而提高“字＋字”的权重，则“你的理解是对的”中本不相干的 jie 和 shi 会被认为可以组成一个词“解释”。

四、总结反思

收获

通过实现拼音输入法的尝试，我对搜索问题有了更深入的了解，也对其背后的概率统计思想和 Viterbi 算法有了进一步的认识。而通过不同模型和参数选择的比较，我也认识到，在输入法等问题中，向前参考更多的状态，以词而非字为单元进行考察，均能有效提高搜索的准确度。

改进空间

对本次实验进行分析后，我认为可以从以下几个角度入手，进一步提升准确率：

- 将词二元语法进一步拓展至词三元语法，扩大搜索视野。
- 引入汉语语法规则，将词性等因素纳入考虑，提高词语搭配的合理性。
- 针对词语间常见的固定搭配进行优化，如“做……工作”“把……”等。
- 合理处理口语表达中的语气词，如“吗”“吧”“啊”等。