拼音输入法实验报告

1. 算法实现：

   我采用了二阶隐马尔可夫算法来实现拼音输入法。

   预处理：

   首先读入拼音汉字表，初始化发射矩阵和汉字列表。接着我根据汉字列表初始化开始向量和转移矩阵。

   接着开始读入文件，对于每一条数据，先将语段去除括号内的内容（通常不是完整连续的语句），再按标点符号分割成独立的语句，最后去除语句中所有非汉字的字符。

   训练：

   对于每一条语句，每个汉字都计入开始向量中。使用 pypinyin 将语句转化成拼音，将对应的拼音-汉字对计入发射矩阵中，最后将连续的汉字计入转移矩阵中。

   使用前向-后向算法（BW 算法）对已有的模型更新权重，进行再训练。参考《模式识别》第三章的内容实现。发现训练效率极低，且成果不显著，遂放弃。

   平滑：

   对于开始向量，我采用加一平滑的方法，在初始化时对每一个元素加一。然后我对每个语句的第一个字额外计入一次，提高它出现在语句首项的概率。

   对于发射矩阵和转移矩阵，我采用 KN 平滑的算法[1]，得到矩阵中每一项平滑后的概率，并取十的对数储存。若某个字一次都没出现过，则取均匀分布。

   解码：

   使用维特比算法[2]解码，从观测序列（拼音）得到隐状态序列。计算 T=t-1 时各个汉字出现的最大概率乘以转移概率，再乘以发射概率，得到 T=t 使各个汉字的概率。最后再逆推得到当前概率最大的汉字，输出结果。

2. 实验效果

   a. 基于二阶隐马尔可夫矩阵的拼音输入法在长句上表现不错，如随机选取的一篇新闻报道：http://news.sina.com.cn/o/2018-04-21/doc-ifznefkh2981650.shtml

      原语段：规划建设雄安新区，是以习近平同志为核心的党中央深入推进京津冀协同发展作出的一项重大决策部署，是习近平总书记亲自谋划、亲自部署、亲自推动的重大历史性工程，是千年大计、国家大事。

      输出其中一段的成果：

      规划建设凶案新区

      是以习近平同志为核心的党中央深入推进京津冀协同发展作出的一项重大决策部署

      实习近平总书记亲自谋划

      亲自部署

      亲子推动的重大历史性工程

      是千年打击

[1] https://blog.csdn.net/baimafujinji/article/details/51297802

[2] https://en.wikipedia.org/wiki/Viterbi_algorithm

国家大事

b. 一些不常出现在新闻的短句表现较差：
ni shi shui
泥石水
ji suan qi
计算起

3. 性能分析：
这个模型没有使用具体的参数（除了平滑中通用的 0.75）。
隐马尔可夫模型的训练用了将近两个小时，效率较差。对一个语句的训练的复杂度是 O(t)，但是由于语料库太大，实际复杂度也很高。我猜测是由于平滑的复杂度是 $O(n^2)$。
解码过程中，复杂度是 $O(tp^2)$（p 是单个拼音对应的汉字数量），因此计算的速度很快。

4. 总结收获
实际上，我考虑过使用三阶隐马尔可夫模型。但是训练的过程中内存要求较高，而且对汉语的分析发现，大多数词是二字词组，三阶模型的结果提高较少。其次，实际上与其使用三元模型，不如收集更多的新闻语料。根据 NFL 理论[3]，模型对于新闻语句的输出提升不如训练集的增加。但是由于时间关系，遂放弃扒新闻。
除此之外，为了解决多音字问题，还可以使用伪三阶 HMM，对发射矩阵中的拼音对汉字的概率，令其受前一隐状态的影响。但是实际上这类问题不常出现，因此用 $O(n^2p)$的复杂度有点得不偿失。
这次项目让我学会了应用算法解决实际问题。虽然只是个 demo，但是在特殊领域（如术语输出）有切实的应用可能。但是在算法中还有很多可以改进的地方，没有做到尽可能完美，略有可惜。

[3] https://en.wikipedia.org/wiki/No_free_lunch_theorem