**《自然语言处理》实验报告**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **年级、专业、班级** | | **2018级计算机科学与技术卓越 01班** | | | **姓名** | **屈湘钧** |
| **实验题目** | **HMM模型训练实践** | | | | | |
| **实验时间** | **2020/12/7** | | **实验地点** | **DS3401** | | |
| **实验成绩** |  | | **实验性质** | **□验证性 □设计性 □综合性** | | |
| 教师评价：  **□**算法/实验过程正确； **□**源程序/实验内容提交 **□**程序结构/实验步骤合理；  **□**实验结果正确； **□**语法、语义正确； **□**报告规范；  其他：  评价教师签名： | | | | | | |
| 一、实验目的  理解、掌握隐马尔可夫模型，N元语法等自然语言处理的基本思想、算法，并将其应用于从汉语拼音到汉字的自动转换过程。 | | | | | | |
| 二、实验项目内容  （1）对训练语料及相关资源进行预处理；  （2）通过学习算法，训练HMM模型；  （3）使用合理的数据平滑方式，解决N元文法中的数据稀疏问题；  （4）利用HMM模型和维比特算法，实现从任意拼音到汉字的自动转换，要求程序能对转换中的歧义进行正确识别和处理。  （5）利用给定测试集，评价上述程序的转换准确率。 | | | | | | |
| 1. 实验过程或算法（源程序） 2. HMM模型   对于一个随机事件，有一个可以观测到的值序列：  该事件的每一个观察到的值都对应一个生成他的状态，则其背后存在一个状态序列：  假设1：（马尔科夫假设）每一个状态的值都与其前n个状态的值相关  假设2：（不动性假设）状态与具体的时间无关  假设3：（输出独立性假设）输出只与当前状态有关  则一个HMM模型是一个五元组  其中            解码问题：对于给定的模型和观察值序列，求出最大可能性的状态序列。   1. 拼音转汉字算法设计   拼音转汉字即对应HMM模型的解码问题。由已知的语料库训练出来汉字到汉字的转移概率和汉字到拼音的发射概率，然后用户输入拼音序列为已知的观察值序列，求大嘴可能性的汉字状态序列。  用一个简单的例子来表示这个识别的过程及原理。  若用户想在计算机得到汉字“我爱中国”，则需要往键盘敲入“wo ai zhong guo”这四个英文字符串。从HMM模型出发，“wo ai zhong guo”是观测值序列，如下观测流程图。  图1 我爱中国汉字拼音输入HMM观测过程  图中蓝色圆圈为隐藏的状态，即汉字，橙色圆圈代表可以观测到观测值，拼音。联系HMM模型，汉字“我”到汉字“爱”的过程是一个转移过程，如果用二元语法模型，则汉字“爱”在“我”的出现情况下有一个转移概率，如后汉字分析同理。同时，汉字我到拼音“wo”有一个发射过程，也有一个概率。则上图可变化为基于概率的识别流程图，如下。  图2 我爱中国汉字拼音输入基于概率的识别过程  但是，拼音“wo”的对应汉字处理“我”还有“卧”、“窝”等，拼音“ai”对应汉字处理“爱”，还有“哎”，“唉”等，那么“wo ai”的组成情况就还会出现“卧爱”、“窝爱”、“我哎”等。如次就出现了如下图3的基于隐马尔科夫模型HMM的拼音转汉字模型图。  如次，基于HMM模型的解码问题，可以求解出状态转移链中概率最大的一条路径，此条路径即所求的汉子序列。    图3 “wo ai zhong guo”拼音输入基于HMM的识别过程   1. 维特比算法   维特比(Viterbi)算法用于解码，在给定模型μ和观察序列O的条件下，使条件概率P(Q|O，μ)最大的状态序列，即  wpsoffice  维特比算法运用动态规划的搜索算法求解这种最优状态序列。为了实现这种搜索，首先定义一个维特比变量。  维特比变量是在时间t时，HMM沿着某一条路径到达状态，并输出观察序列的最大概率:  wpsoffice  有如下递归关系:  wpsoffice  这种递归关系使我们能够运用动态规划搜索技术。为了记录在时间t时，HMM通过哪一条概率最大的路径到达状态。  其伪代码如下：  -----------------------------------------------------------  **维特比算法(Viterbi algorithm）**  **初始化：**  **wpsoffice**  **归纳计算**  **wpsoffice**  **终结**  **wpsoffice**  **wpsoffice**  -----------------------------------------------------------   1. 模型实现与构建   其构建的流程如下图。  D0BB069BA514332899E7745B80574369  图4 项目模型构件图  1.语言模型训练  本处使用老师给的toutiao\_cat\_data.txt文件中的数据。   1. 语料清洗   由于文件中的文字段都是带有非法字符和大段文字的。所以这里本人使用了正则表达式来匹配中文字符，对于非中文字符都直接忽略，并切断句。  如此，即可得到如图5所示的规则汉字串。    图5 清洗非法字符后的汉字串  但是，HMM模型需要拼音到汉字的发射概率，现在我们还缺少汉字的标准注音。所以，本人调用了pypinyin第三方库来对所有的汉字注音。获得拼音列表。     1. 语言模型训练   基于如上的大量语料库文本，由一段文字可以得到文字库和拼音库，然后统计文字的频次、文字到拼音的频次、文字到文字的频次。如此，根据N-Gram 语言模型原理    得出一元语言模型和二元语言模型：      同时训练时采用加一平滑技术得到如下公式：      依照如上公式带入语料库文本信息则可以训练得到HMM模型的概率矩阵。同时，为方便HMM模型程序调用整理好的语料数据，将此些加工后的语料文件保存为python易使用的npy文件。  此外，为了方便使用如上的语言模型计算，这里保存了四个语料文件，分别为每一个拼音对应的所有已知汉字集合文件py2hanzi.npy、为汉字编码后的汉字编码字典文件my\_hanzi\_dict.npy、汉字编码到汉字编码二字词映射频次对应的二维矩阵文件my\_moving\_array.npy、单个汉字出现次数的列表文件my\_hanzi\_num.npy和汉字对应的各个拼音文件my\_emission\_dic。  其生成的代码如下：  统计汉字的出现频次，用于计算一元语言模型。  汉字的编码列表  汉字编码到汉字编码的次数映射，用于统计二元语法模型。  汉字到拼音的频数，其格式为{'了':{'le':5, 'liao':10},'屈':{'qu':5}}。  至此，所有的语料文件都已训练好并保存为npy格式的文件，方便调用。  2.HMM模型的构建  依照图4的模型，首先构建一个HMM的模型模块，输入为一串拼音序列，并判断“l”、“n”遇上元音“ü”的情况，替换“v”为“ü”，然后加载语料数据，构建HMM模型如图3，每一个汉字状态圆圈代表一个节点，节点存储此汉字和汉字对应拼音的语料数据。然后通过维特比算法获得最优的汉字序列，输出汉字序列。其伪代码如下：  **---------------------------------------------------------------**  **输入：拼音序列**  **过程：**  **1： 加载语料数据文件，获得汉字到汉字、拼音的概率表和汉字的概率表**  **2： 输入拼音**  **3： 若“l”、“n”遇上元音“ü”的替换“v”为“ü”**  **4： 初始化每个拼音的汉字状态节点，汉字状态节点初始化三类概率**  **5： 维特比算法求解**  **6： 输出汉字序列**  **---------------------------------------------------------------**  下面我们用代码实现如上的伪代码。  首先是加载语料数据文件。  然后是设计了一个Graph邮箱图类来存储HMM模型的结构。其初始化为将输入的拼音字符串分解，为每个拼音构建其汉字节点，然后将每个节点初始化。  其中每个汉字都是一个节点，节点也是一个类，保存了其汉字、次数、维特比变量和前一个节点等熟悉。  在初始化好HMM模型结构后，我们就可以开始用维特比算法计算每一个节点的维特比变量，来对整个模型的节点进行遍历计算。其中初始节点的维特比变量用一元语法模型计算，后续节点用二元语法模型状态，都是用的加一平滑技术。然后发射概率也是加一平滑的。  在训练完成后，HMM模型就可以找到最优的路径了，这时候利用每一个节点保存的最优上一节点可以反向遍历得到最佳的路径，也就是最优的字符串序列，期待吗如下。  至此，HMM模型已经训练完成和可以进行输入法预测功能。   1. 模型改进   在如上语料模型训练中，存在一些问题，最终会大幅度影响整个HMM模型的训练结果。其缺点主要有以下几点：  1.语料文件质量差。文中有大量非法字符，不得不用正则表达式去判断非法字符的位置，然后切断汉字串为两个汉字串。比如“我爱穿T恤，我好开心”，这里面有非法字符“T”和“，”，会将字符串切分为3个字符串段。  2.语料文件无拼音注释，第三方库添加错误率较高。因为语料文件没有拼音语料，所以不得不用第三方库Pypinyin来为所有的汉字串注音，但是此pypinyin可以的拼音准确度会直接影响我们的概率，同时其中有非常多的识别错误，也没有声母n、l跟韵母ü遇上的转换等过程。  3.语料文件中的汉字数量只有3000多字，而常用的汉字有8000字左右，相差较多，同理，其拼音数量也严重缺失，导致语言模型训练会出现大量的平滑。  由此，本文在网上下载了一些第三方较好的语料库文件，并训练好用来做出更好的改进。同时，本文也在网上找了更丰富训练集来测试模型。在此，感谢https://github.com/THUzhangga/HMM\_shurufa/tree/master/data的语料文件和测试集的帮助。 | | | | | | |
| 1. 实验结果及分析   1.基于实验语料文件的HMM训练与测试的结果  （1）基于toutiao\_cat\_data.txt文件训练后，对于实验给的测试集.txt文件进行测试如下图。可得到测试准确率在79.2%。    （2）基于toutiao\_cat\_data.txt文件训练后，对于github.com/THUzhangga的test\_set.txt文件进行测试如下图。可得到测试准确率在67.58%。    2.基于改进后的语料文件的HMM训练与测试的结果  （1）基于github.com/THUzhangga的语料文件训练后，对于实验给的测试集.txt文件进行测试如下图。可得到测试准确率在87.61%，可明显看到准确率大幅度提升。    （2）基于github.com/THUzhangga的语料文件训练后对于github.com/THUzhangga的test\_set.txt文件进行测试如下图。可得到测试准确率在75.56%，可明显看到准确率大幅度提升。    3.模型评价与分析  本文选取一些测试样例和搜狗输入法做比较，可以明显观察到长句的错误率较高，短句和常见的词组正确率一般都较高，同时搜狗输入法的长句错误率也较高。  此外，在运行时间和资源消耗的上，时间预测一个拼音字符串的主要时间消耗在数据文件的加载中，同时存储开销达到百兆的大小，并随着语料库的增加还会继续增大。  总体来说，本项目的正确率较高，但是项目的其他资源消耗太大，仅作为实验验证还是不错的。  表1 实验项目结果对比   |  |  |  | | --- | --- | --- | | **预测结果** | **搜狗输入法** | **正确答案** | | 虽然已经解决了建立新积分方法的首要问题 | 虽然已经解决乐见离心机芬芳发的首要问题 | 虽然已经解决了建立新积分方法的首要问题 | | 建立了交易办机上的策独立伦 | 建立了交易班级上的测度理论 | 建立了较一般集上的测度理论 | | 后面我们将成具有这种性质的函数为可测寒暑 | 后面我们将成具有这种性质的函数味可测函数 | 后面我们将称具有这种性质的函数为可测函数 | | 今天也是好天气 | 今天也是好天气 | 今天也是好天气 | | 中央已经决定了 | 中央已经决定了 | 中央已经决定了 | | | | | | | |