文章编号：2095-2163(2020)04-0239-05 中图分类号：TP391.4 文献标志码：A

印刷维吾尔文识别后处理

贾钰峰[1]，章蓬伟[1]，邵小青[2]，张玉茜[1]
（1 新疆科技学院 工商管理系 ，新疆 库尔勒 841000；2 中国石油天然气运输集团，新疆 库尔勒 841300）

摘 要：主要是将隐马尔可夫模型运用于印刷体维吾尔文识别后处理的初步尝试。在隐马尔可夫模型运用中，通过对识别错误进行了比较，分析，分类，构造 B 矩阵，使文本文字也可以产生双重随机性，满足隐马尔可夫模型的使用条件。最后，完成对相同的五个测试文本进行四种后处理的测试。

关键词：印刷体；维吾尔文；隐马尔可夫模型；后处理

Printed Uygur Character Recognition Post-processing

JIA Yufeng[1], ZHANG Pengwei[1], SHAO Xiaoqing[2], ZHANG Yuqian[1]
（1 Xinjiang University of science and technology Department of business administration Korla, Xinjiang 841000, China;
2 CNPC, Korla Xingjiang 841300, China）

【Abstract】 The Hidden Markov Models initial attempt to applied to Printed Uygur character recognition post-processing. In the application of The Hidden Markov model, the paper compares, analyzes and classifies the recognition errors and constructs the B matrix, so that the text can also generate double randomness and meet the application conditions of the hidden Markov model. Final, this paper show that the test results of five samples by apply four post-processing.

【Key words】 Printed; Uyghur; Hidden Markov Models; post-processing

0 引 言

印刷体文字识别主要是针对印刷文字进行识别。文字识别系统由预处理，特征提取，模式匹配和后处理四大模块组成[1]。后处理的主要任务是提高系统的识别率。本文中印刷体维吾尔文识别后处理是采用统计方法实现的[2]，即将识别结果与存储这类字符文本所有词的数据库中的信息作比较或分析字符串出现的频率。

1 前期工作

本文的维吾尔文印刷识别后处理，是通过对识别错误建立误差模型，用模型中的错误分类与统计，得出错误字母与正确字母之间的概率，从而得出隐马尔科夫模型参数（A、B、π）中的 B 矩阵（观察矩阵），其流程如图 1 所示。最终使用 HMM 模型进行校正维文印刷体识别后处理，通过对结果进行观察，判断该方法是否可行。

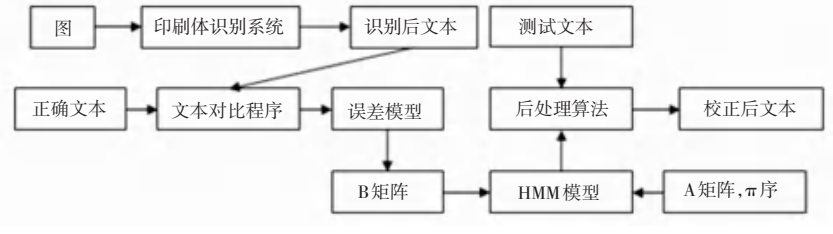


图 1 实验流程图

Fig. 1 Experimental flow chart

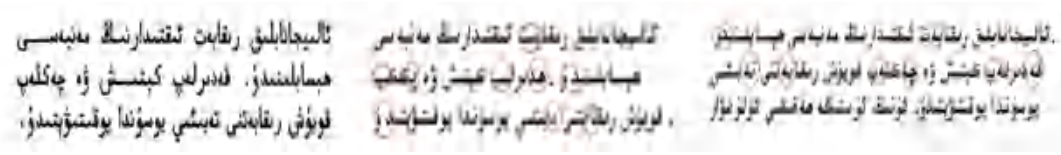
本文后处理实验前，通过对 54 个维吾尔文的图像进行识别，预先准备了错误语料信息。如图 2 所示。

误差模型是通过文本比对程序比对后，发现识别后错误的类型主要是切分的错误，空格的错误，近似字母和首字母的错误。对错误类型人工调整的更加合理后，再对错误语料进行处理得到的模型。

基金项目：自治州科学技术基金项目（2019018）。

作者简介：贾钰峰（1986-），男，硕士，助教，主要研究方向：图像处理与模式识别；章蓬伟（1985-），男，硕士，助教，主要研究方向：人工智能与模式识别、大数据。

收稿日期：2019-12-27



(a)待识别图像 (b)识别的结果文本 (c)正确文本

(a) Image to be recognized (b) Recognition result image (c) Correct result text

图 2 识别效果图

Fig. 2 Recognition effect picture

2 算法概述

2.1 隐马尔可夫模型

20 世纪初俄罗斯数学家 Markov 首次提出了一种随机过程。该随机过程是一种用参数表示的,用于描述随机过程统计特性的概率模型,是一个双重随机过程,由两个部分组成:马尔可夫链和一般随机过程。其中,马尔可夫链用来描述状态的转移,用转移概率矩阵描述。一般随机过程用来描述状态与观察值的关系,用观察值概率矩阵描述。对于 HMM 模型,其状态转换过程是隐含的,因此称之为“隐”马尔可夫模型。

HMM 中相关概念的定义如下[3]:

(1)隐含状态: $S = \{s_1, s_2, \cdots, s_n\}$,是可能出现的隐状态集合。其中 N 是状态的总数;

(2)不同的观察符号总数 M。观察序列被模型化后的输出,各个符号表示为: $V = \{v_1, v_2, \cdots, v\}$;

(3)状态转换概率 $A = \{a_{ij}\} X$. 其中:

$$a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i),\ 1 \leqslant i, j \leqslant N \quad (1)$$

实际应用中,根据具体情形添加相应的限制。模型的定义中,假设一个状态能达到任何状态(也包括本身状态);

(4)在某一个状态 J 中观察符号概率分布 $B = \{b_j(k)\}$,其中:

$$b_j(k) = P(v_k \text{ at } t \mid q_t = s_j),\ 1 \leqslant j \leqslant N, 1 \leqslant k \leqslant M \quad (2)$$

(5)初始状态分布:模型中的各个状态初始观察符号概率。

HMM 包含 3 个问题:

问题 1 对给予的观察序列 $O = o_1 o_2 \cdots o_T$ 和一个模型 HMM (π, A, B),如何有效地计算该观察序列的出现概率,也就是根据已给予的模型和观察序列如何计算此观察序列被此模型产生的概率。此问题可用前向算法或后向算法解决。

问题 2 对给予的观察序列 $O = o_1 o_2 \cdots o_T$ 和一个模型 HMM (π, A, B),如何选择一个理想的隐状态序列(隐状态序列最可能地产生所给出的观察序列),此问题可用维特比(Viterbi)算法用来解决。

问题 3 观察序列的出现概率尽量增大的情况下,如何优化模型的各种参数,使该模型最好地描述观察序列的产生,此问题可用前向-后向算法(BW 算法)来解决。

2.2 建立误差模型

通过统计分析错误语料中的错误,得到误差模型,误差模型主要工作:一个是查找错误单词队列,另一个是查找详细错误类型[3]。

首先,对错误文本 E 和正确文本 R 进行逐个字符的对照,当发现字符 r 与字符 e 不一致时,分别找出 R 中与字符 r 对应的单词 R_i,E 中与字符 e 相对应的单词 E_j。找出 R 中 R_i 后的所有单词 $R_{i+1}R_{i+2}R_{i+2}\cdots R_{i+n}$,$E$ 中 E_j 后的所有单词 $E_{j+1}E_{j+2}E_{j+3}\cdots E_{j+m}$,逐个单词的比较,当发现单词 R_{i+x} 与单词 E_{j+y} 相等时,则单词 $R_iR_{i+1}R_{i+2}R_{i+3}\cdots R_{i+x-1}$ 为正确序列,否则单词 $E_jE_{j+1}E_{j+2}E_{j+3}\cdots E_{j+y-1}$ 为错误序列,示意图如图 3 所示。

图 3 查找错误单词队列示意图

Fig. 3 Diagram of finding wrong word queue

错误类型分析模块将单词 $R_iR_{i+1}R_{i+2}R_{i+3}\cdots R_{i+x-1}$ 视作 P，将字符串 $E_jE_{j+1}E_{j+2}E_{j+3}\cdots E_{j+y-1}$ 视作 Q。对 P 与 Q 进行逐个字符的比较，当发现字符 P_i 与字符 Q_j 不等且字符 P_{i+x} 与字符 Q_{j+y} 相等时，则 $P_iP_{i+1}P_{i+2}P_{i+3}\cdots P_{i+x-1}$ 为正确字符，$Q_jQ_{j+1}Q_{j+2}Q_{j+3}\cdots Q_{j+y-1}$ 为错误字符。通过分析比较正确字符与错误字符得出错误类型，错误类型分析模块流程示意图如图 4 所示。

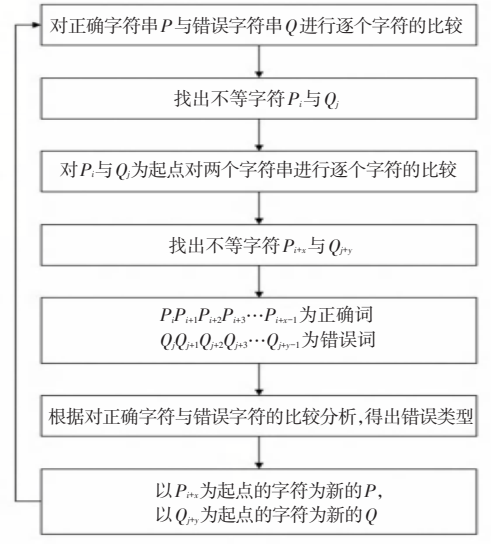


图 4　查找详细错误类型示意图

Fig.4　Schematic diagram of finding detailed error types

2.3　基于 HMM 算法的后处理算法

对于 HMM 模型而言，该模型的适用对象为双随机事件，比如语音识别，文字识别。对于识别后处理与文本校正，一般都认为是单随机事件，所以并不适用于 HMM 模型。但文本识别并不是完全意义上的单随机事件，它也有双随机特性，也可以适用于 HMM 模型。

首先，对 HMM 模型三元组（π，A，B）进行分析。A 矩阵为转移矩阵，即从一种隐藏状态转移到另一种隐藏状态的概率集合。B 矩阵为观察矩阵，即从一种观察值到一种隐藏状态的概率集合。π 序列为初始概率序列，即一种隐藏状态的初始概率集合[4-5]。

对于拼写文字，A 矩阵就是从一个字母转移到另一个字母的概率，即 a_{ij} 为 a_ia_j 这种组合在训练集中的概率，其中 a_i 为一个字母，a_j 为另一个字母。π 序列就是一个字母为单词首字母的概率。可以得出字母为隐藏状态，观察序列就是待校正的错误单词，由此得出观察值为校正的错误单词的字母。那么字母即为隐藏状态又为观察值，也不会产生冲突，因为隐藏状态对应于正确单词的字母，而观察值对应于错误单词的字母（包括正确字母与错误字母），虽然二者为同一个集合，但二者的意义不同。例子：单词 the 被识别成 tbe，单词 list 被识别成 bist，单词 bad 被识别成 bab，即字母 h、l、d，都有可能被识别成 b，当得到一个包含字母 b 的观察序列（待校正单词），字母 b 为观察序列中的一个观察值，从 b 可以得出那些隐藏状态。字母 b 可以由其本身识别得出，所以可得出隐藏状态字母 b，除了其本身外，字母 b 还可以由字母 h、l、d 错误的识别得出，因此还可以得出隐藏状态字母 h、l、d，通过观察值字母 b，可得出隐藏状态字母 b、h、l、d。由此，可以对误差模型中的错误分类进行分析，得出观察值与隐藏状态对应的 B 矩阵（观察矩阵）。维吾尔文字也属于拼写文字，也可以用该方法生成 B 矩阵（观察矩阵）。

生成 A 矩阵（转移矩阵）的流程图如图 5 所示。

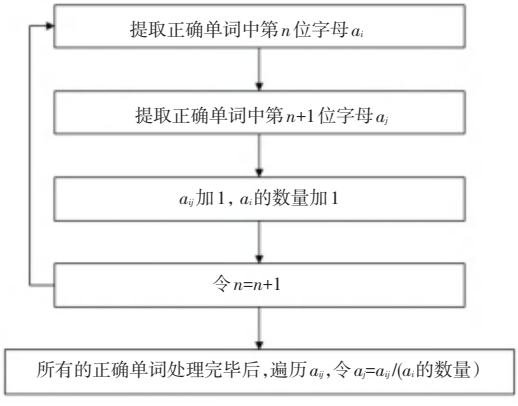


图 5　生成 A 矩阵流程图

Fig. 5　Flow chart of generating a matrix

生成 B 矩阵（隐藏矩阵）的流程图如图 6 所示：

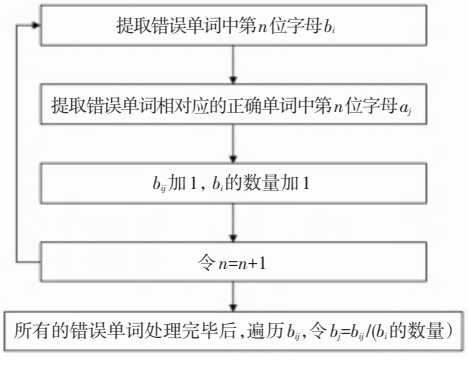


图 6　生成 *B* 矩阵的流程图

Fig. 6　Flow chart of generating *B* matrix

生成 π 序列（初始概率序列）的流程图如图 7 所示。

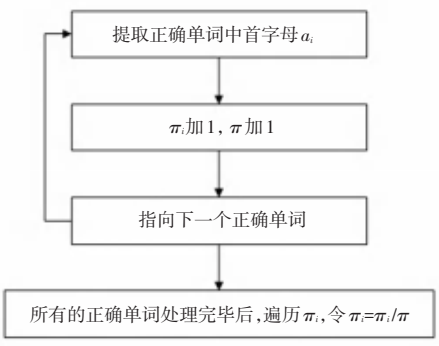


图 7 生成 π 序列的流程图

Fig. 7 Flow chart of generating π sequence

3 试验结果

3.1 建立误差模型

对于误差模型，编写了两个算法：(1)查找错误单词队列算法；(2)查找详细错误类型算法。将错误类型分为 5 种，分别为：删除错误(单与多)、替换错误(单与多)、插入错误(单与多)、空格错误、其它情况错误。

将查找错误单词队列算法执行部分结果，见表 1。

表 1 错误词与正确词对照表

Tab. 1 Comparison of wrong words and correct words

正确词	错误词
ريقابەت	ريقابت
هىسابلىنىدۇ قەدىرلەپ	هىسابلىنىدۇ ۋەدىرلەپ
چەغلەپ	پىكىپ
ريقابەتنى تەبىئي	ريقابتى ەبىئي
يوقىتۇپتىدۇ ئۇنىڭ	يوقىتۇپتىد ۇ ئۇنىڭ
تالانت	تالات
قەدىرلەپ كېتىش	ھەدىرلىپ كېتىش
چەغلەپ	پىكىپ
ئەرەپتىن	ھەرەپتىن
بولىدۇ يازغۇچى	بولىد ۇ يازغۇچى

将查找详细错误类型算法执行部分结果，见表 2。错误分类统计结果，见表 3。

3.2 基于 HMM 算法的后处理算法

本实验主要的有三个算法：(1)HMM 模型初始化算法；(2)前向-后向算法(BW 算法)；(3)维特比算法。

在误差模型的基础上建立了 HMM 模型，该算法的核心功能代码与建立误差模型算法的核心功能代码极其相似，区别在于建立误差模型算法的功能代码只对错误类型进行分析，统计后输出，HMM 模型初始化算法的核心功能代码将错误类型分析统计后，求出所需的各种概率。下面将给出四种后处理算法的测试结果。

表 2 正确和错误词的错误类型对照表

Tab. 2 Comparison of error types of correct and wrong words

正确词	错误词	词/错误类型
ريقابەت	ريقابت	词
ە	ى	替换错误(单)
چەغلەپ	پىكىپ	词
چ	پ	替换错误(多)
ل	ك	替换错误(多)
ريقابەتنى تەبىئي	ريقابتى ەبىئي	词
ە	ب ت	删除错误(单)
ت	ە	删除错误(单)
يوقىتۇپتىدۇ ئۇنىڭ	يوقىتۇپتىد ۇ ئۇنىڭ	词
	د ۇ	空格错误
ن	ت	替换错误(单)
تالانت	تالات	词
ن	ا	删除错误(单)

表 3 错误分类统计结果

Tab. 3 Statistical results of error classification

错误类型	次数	概率/%
错误总数	1 152	100
替换错误	767	66.58
空格错误	203	17.62
插入错误	67	5.81
删除错误	115	9.98

用维文印刷体识别系统识别实验中给出的测试集识别正确率为 85.393%，从文本库中随机抽取了五个文本，识别系统的识别率为 79.705%；经过未进行 BW 学习的 HMM 模型的后处理校正后，再进行测试，正确率提高至 80.272%；经过只统计错误词的未进行 BW 学习的 HMM 模型的后处理校正后，平均正确率也为 80.272%。原始识别文本与经过后处理后识别的准确率情况见表 4。

表 4 准确率统计结果

Tab. 4 Statistical results of accuracy

算法类型	正确率/%
原始测试集	79.705
基于未进行 BW 学习的 HMM 模型的算法	80.272
基于进行 BW 学习的 HMM 模型的算法	79.252
只统计错误词的基于未进行 BW 学习的 HMM 模型的算法	80.272
只统计错误词的基于进行 BW 学习的 HMM 模型的算法	79.252

算法对于每个文本的详细统计结果见表 5。

表 5 正确率详细统计结果

Tab. 5 Detailed statistical results of accuracy %

	文本一	文本二	文本三	文本四	文本五
无修改	74.857	100	77.143	71.111	77.005
无 BW	74.857	99.394	77.714	73.889	77.005
有 BW	74.286	98.788	77.143	70.556	77.005
错词无 BW	74.857	99.394	77.714	73.889	77.005
错词有 BW	74.286	98.788	77.143	70.556	77.005

(下转第 245 页)