中 文 信 息 学 报
第 13 卷 第 4 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 13 No. 4

基于 HMM 的汉语文本识别后处理研究*

李元祥 丁晓青 刘长松
清华大学电子工程系 北京 100084

摘要 本文用 HMM(Hidden Markov Model)描述汉语文本识别后处理,将汉语语言和单字识别这两个概率模型结合起来,以充分利用单字识别器提供的信息。语言模型的参数由语料库统计得到;单字识别模型的参数为条件概率,经理论分析,它可转化为后验概率来求解。在分析训练样本集单字识别结果的基础上,提出一种统计方法估计候选字的后验概率。HMM 在脱机手写体汉语文本识别中的实验表明,后处理性能除取决于语言模型外,还取决于后验概率的精确估计。

关键词 汉字识别 后处理 语言模型 隐马尔可夫模型 后验概率

Post-processing Study of Chinese Document Recognition Based on HMM

Li Yuanxiang Ding Xiaoqing Liu Changsong
Department of Electronic Engineering Tsinghua University Beijing 100084
Email: lyx@ocrserv.ee.tsinghua.edu.cn

Abstract In this paper, a post-processing method using HMM(Hidden Markov Model) for Chinese document recognition is proposed. HMM combines language model with single character recognition(SCR) model to make the best of SCR output. The parameters of language model are derived from corpus, while the parameters of SCR model are conditional probabilities that can be converted into posterior probabilities by theoretic analysis. On the basis of SCR output analysis, posterior probabilities of candidates are obtained by statistical method. Experiments in off-line Chinese document recognition show that post-processing performance depends on efficient evaluation of posterior probability, besides proper language model.

Keywords Chinese Character Recognition Post-processing N-gram Language Model Hidden Markov Model Posterior Probability

* 本文系"全国中文信息学术交流暨工作会议"推荐的优秀论文
本文得到国家 863(编号 863-306-03-05-6)及国家自然科学基金(编号 69682003)的资助

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

一、引 言

汉语文本识别中,由于汉字结构复杂且变化性大,单字识别率受到一定的限制。为提高文本识别率,需在单字识别的基础上利用上下文相关信息进行后处理。随着 OCR 的不断发展,后处理技术愈来愈受到重视。

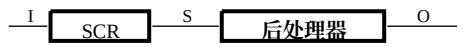
识别过程中,采用最近邻距离分类器时,单字识别器 SCR(Single Character Recognizer)除了输出候选字信息外,还输出与候选字相应的距离信息。显然,在后处理时应充分利用这两种信息来提高文本识别率。但是在以往的文献中,后处理时强调候选字上下文相关信息(即语言模型)的利用;候选字距离信息或被绕过[1],或只是简单地加以利用[2,3],缺乏理论依据。

HMM 是一种研究时间序列的随机方法[4]。HMM 描述的随机过程是双重的,其一是 Markov 链,这是基本随机过程,描述状态间的转移;另一随机过程描述状态和观测值之间的统计对应关系,状态隐含在观测值中。在文本识别过程中,汉字与其图象或特征间隐含着对应关系,即状态和观测值之间的隐含对应关系;汉字间的语言统计相关(Markov 链)表述了状态间的转移。因此,HMM 适于描述文本识别过程。

本文试图用 HMM 描述文本识别后处理,将自然语言和图象观测这两个随机过程结合起来。在对大量训练样本集单字识别结果分析的基础上,通过定义候选字可信度,将候选字的距离信息有效地用于候选字后验概率的估计中,以提高文本识别后处理性能。

二、用 HMM 描述汉语文本识别后处理

文本识别系统框图如下:



$I = I_1 I_2 \cdots\cdots I_n$ 为输入文本的一串字符图象(或其特征)序列,即观测值序列。

$S = S_1 S_2 \cdots\cdots S_n$ 为 SCR 的输出汉字序列(每个输出有多个候选汉字),即状态序列,状态集 C 为单字识别字符集。$I_i \rightarrow S_i \in \{ C_{ij} \quad j = 1, 2, \cdots, m \}$, $\{ C_{ij} \quad j = 1, 2, \cdots, m \}$ 为 I_i 的候选字集,其中 $C_{ij} \in C$。$O = O_1 O_2 \cdots O_n$ 为后处理器的汉字输出序列。

n 为句子的长度,m 为候选字集的大小。若仅考虑前十个候选字($m = 10$),则 S 共有 10^n 种可能的汉字序列;后处理时,要求从中选出最符合汉语语言规律的一种,这是典型 HMM 的第二个问题:选择最佳状态转移序列问题。这里

状态转移概率分布矩阵 $A = (a_{ij}) \quad a_{ij} = P(S_j \mid S_i)$

观察值概率分布矩阵 $B = (b_j(k)) \quad b_j(k) = P(I_k \mid S_j)$

初始状态分布 $\pi = (\pi_1) \quad \pi_i = P(q_1 = S_i)$ 为句首概率

基于 HMM 的汉语文本识别过程如下:

$$O = \arg\max_S P(S \mid I) = \arg\max_S P(S) * P(I \mid S) \qquad (1)$$

其中 $P(S)$,描述语言的统计概率分布,表示自然语言这一随机过程,由语言模型决定;$P(I \mid S)$ 描述文本的观测图象概率分布,表示图象观测这一随机过程,由 SCR 模型决定。

典型 HMM 中的模型参数由 Baum - Welch 算法迭代得到[4],它需要大量的观测值序列进行训练。但是,在文本识别后处理中,图象观测序列是极其有限的;因此,Baum - Welch 算法在这里是不适用的。实际上,模型参数可由语言模型和 SCR 模型分别求得。A 矩阵和初始状态分布 π 由语言模型决定,通过大规模语料文本统计可以得到。B 矩阵中的元素为条件概

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

率,难以计算。但是,通过下面的理论分析,条件概率可转化为后验概率;从而解决了 B 矩阵问题。

由于 SCR 对每个孤立的字符图象 I_i 进行识别,显然这种识别不依赖于上下文关系。故式(1)中条件概率 $P(I|S)$ 可表示为:

$$P(I|S) = P(I_1 \cdots I_n | S_1 \cdots S_n) = \prod_{i=1}^{n} P(I_i|S_i)$$

$P(I_i|S_i)$ 为单字的条件概率(HMM 中的 B 矩阵参数),S_i 为对应的候选字 C_{ij} 之一($j=1,2,\cdots,10$),所以

$$P(I_i/S_i) = P(I_i/C_{ij}) = P(C_{ij}/I_i) * P(I_i) / P(C_{ij})$$

这里,$P(I_i)$ 对 I_i 的每个候选字 C_{ij} 是一样的;$P(C_{ij})$ 为模式类的先验概率,在 SCR 中假定各个模式类出现的概率是相等的。故在实际计算中,$P(I_i)$ 与 $P(C_{ij})$ 这两项可不考虑。从而,求 $P(I_i|S_i)$ 就转换为求 $P(C_{ij}/I_i)$,$P(C_{ij}/I_i)$ 为后验概率。于是式(1)可表示如下:

$$O = \arg\max_s P(S|I) = \arg\max_s P(S) * \prod_{i=1}^{n} P(S_i|I_i) \qquad (2)$$

三、候选字后验概率估计

在 SCR 中,后验概率难以直接求出。最大后验概率判决一般转化为最近邻距离判决;后验概率越大,对应的距离就越小[5]。SCR 依照距离的大小排序,给出前 10 个候选字以及相应的 10 个距离值。后处理时,必须通过某种方式或映射将距离转换成后验概率。

文献[2,6]采用距离经验公式来估计后验概率。[2]中的距离经验公式如下:

$$P(C_{ij}|I_i) = \frac{SCORE_j}{\sum_{j=1}^{10} SCORE_j}, \quad SCORE_j = \frac{1}{d_j - d_1 + 1} \qquad (3)$$

[6]中的距离经验公式如下:

$$P(C_{ij}|I_i) = \frac{1/d_j}{\sum_{j}^{10} 1/d_j} \qquad (4)$$

上述两式中,d_j 表示第 j 个候选字对应的距离;它们虽然在一定程度上反映了识别的可靠性,但并没有与具体的识别器以及具体的样本特征空间分布紧密结合起来,经验的成分较多,缺乏理论依据。

3.1 对训练样本集单字识别结果的分析

距离的大小反映了 SCR 识别结果的相对可靠性,但由于汉字特征空间分布的不均匀,一些字特别是相似字间的距离可能会很小。所以距离小的候选字,其可靠程度不一定就高;就笔画简单的字而言,首选距离 d_1 较小,但第二选距离 d_2 也较小;例如,己:已 301574 己 327514。而距离大的候选字,其可靠程度不一定就低;对笔画复杂的字,d_1 较大,但 d_2 更大;例如,糕:糕 612232 拼 727456。

令距离差 $dif = d_2 - d_1$。大量实验统计表明,候选字的可靠程度随距离差的大小而单调地变化。距离差 dif 愈大,首选字的可靠程度愈高;反之,首选字的可靠程度愈低。当 dif 大于某一门限 TH 时,首选字的可靠程度极高,TH 可由识别系统确定。可见,距离差的大小能

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

够较好地反映候选字可靠程度的高低。但是,距离差反映的仍然是识别结果的相对可靠性,必须将它转换成能反映识别结果可靠性的概率度量。

3.2 候选字可信度的定义

令 x 为图象特征空间内的一点,分类器对 x 的判决为 $C_j(x)$(前 10 个候选字,$j=1,2,\cdots,10$),x 处的图象的真实类别为 $\omega(x)$,则定义 $C_j(x)$ 正确的概率:

$$CF_j(x) = P\{\omega(x) = C_j(x)\} \qquad j = 1,2,\cdots,10 \tag{5}$$

为该分类器在 x 处候选字的可信度。

上述定义中,$C_j(x)$ 是一个确定的类别;$\omega(x)$ 是一个随机变量,因为不同类别的图象特征提取后可能会落在特征空间内的同一点 x 处,从而使 $\omega(x)$ 表现出不确定性。

特别指出的是:$CF_j(x)$ 与后验概率是紧密联系的,它就是当一个图象落在特征空间中的 x 处时,该图象的类别为 $C_j(x)$ 的后验概率。即 $CF_j(x) = P\{\omega(x) = C_j(x)\}$。

下面给出一种求可信度的统计方法,将距离差(相对度量)转换成可信度(绝对度量)。

3.3 可信度估计

将区间 $[0, TH]$ 均匀量化,量化间隔为 Δ;$dif \geq TH$ 作为一个量化区间;这样,共有 $p = TH/\Delta + 1$ 个量化区间。在一个充分大的训练样本集上,对每个量化区间 Δ_i($i=1,2,\cdots,p$),统计落入该区间的样本个数 m_i,其中,第 j 个候选字是正确字的个数为 n_{ij},则由式(5)定义可知,n_{ij}/m_i 即为第 j 个候选字在该区间上的可信度 $CF_j(\Delta_i)$。由于 $n_{ij} \leq m_i$,显然 $CF_j(\Delta_i) \in [0,1.0]$。当 $dif \geq TH$ 时,对首选字,$CF_1(\Delta_p)$ 接近于 1.0;而对非首选字,$CF_j(\Delta_p)$($j=2,3,\cdots,10$)极小,趋近于 0。注意:量化间隔 Δ 不宜过小,应使可信度呈单调变化趋势。

这样就得到了各个候选字的后验概率:

$$P(C_{ij} \mid I_i) = CF_j(\Delta_k) \qquad j = 1,2,\cdots,10,\ k = 1,2,\cdots,p \tag{6}$$

受训练样本集(500 套 3755 个一级汉字脱机手写体样张)的限制,在实验中,我们将 SCR 的 10 个候选字分成首选和非首选两类,分别对这两类求可信度。求得的可信度存放于一张表中。后处理时通过距离差映射,查表直接得到各个候选字的可信度,简便易行。训练样本集很多时,可以统计得到每一级候选字的可信度。

首选字及非首选字的可信度与量化层次(距离差)的对应关系如图 1 所示(p = 21)。

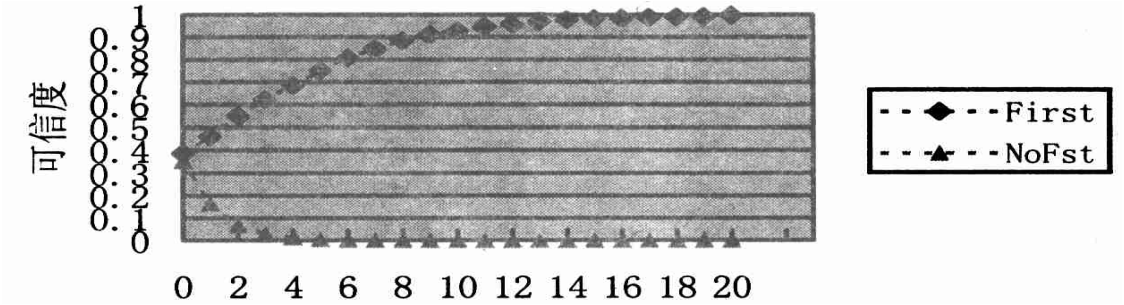


图 1 可信度曲线

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

四、N-gram 统计语言模型

假定汉语语言是一个 $N-1$ 阶的 Markov 链,即当前字、词只与前 $N-1$ 个字、词有关,而与更前面的字、词无关。此时的语言模型称为 N 元文法模型(N - gram)[8],故式(2)中的 $P(S)$ 可用 $N-1$ 阶 Markov 模型表示为:

$$P(S) = \prod_{i=1}^{N-1} P(S_i \mid S_1 S_2 \cdots S_{i-1}) * \prod_{i=N}^{n} P(S_i \mid S_{i-N+1} \cdots S_{i-1})$$

当 $N=2$ 时,有二元文法模型 Bigram。此时

$$P(S) = P(S_1) * \prod_{i=2}^{n} P(S_i \mid S_{i-1})$$

当 $N=3$ 时,有三元文法模型 Trigram。此时

$$P(S) = P(S_1) * P(S_2 \mid S_1) * \prod_{i=3}^{n} P(S_i \mid S_{i-2} S_{i-1})$$

字词间的转移概率由大规模语料文本统计得到。我们所用的语料是 1993 和 1994 两年的《人民日报》全文,约 4000 万字。其中,汉字占 87.6 %,非汉字符(数字、标点及其它符号)占 12.4 %。对于脱机手写体汉字识别,由于单字识别率不高,后处理一般是基于字的。对该语料库进行字一级上的加工,得到单字频、句首字频(HMM 中的初始状态分布 π)以及二元、三元字字同现概率矩阵(HMM 中的 A 矩阵)。

参照文献[1]采用了 3763 个标记,其中一级 3755 个汉字各为一个标记,国标一级汉字之外的所有汉字为一个标记;另外,将数字、字母及标点符号归为其它 7 个标记。

统计表明,同现过至少一次的二元同现对数目仅为 1100808(其中二级汉字有 2327 个同现对),占总数目(3763 ×3763)的 7.8 %。可见,二元同现对是稀疏的;用线性链表存储时,仅需 6.5M 字节空间即可。当删除同现次数少的二元组时,存储空间会更小。

三元同现对是极其稀疏的。至少出现一次的三元组数目为 2037200,仅占总数目(3763 × 3763 ×3763)的 0.0038 %,采用线性链表方式存储,约需 12M 字节空间。

五、实验结果及分析

HMM 用于脱机手写体汉语文本识别后处理时,以句子为处理单元,用 Viterbi 动态规划方法[4]从各候选字集中搜索最佳路径,作为文本识别的最终输出。

5.1 实验环境

利用 TH - OCR '97 综合集成汉字识别系统中的"脱机手写体汉字识别分系统"[7]进行单字识别。识别对象为 49 篇手写体文稿,共 19135 个字符,其中汉字 17324 个,每篇约 400 字,文稿书写质量差异较大,首选识别率最高达 96 %,最低仅为 50 %。后处理之前,文本平均识别率为 85.75 %;十选累计识别率是 96.15 %。

5.2 后验概率估计方法对后处理性能的影响

对上述 49 篇文稿进行后处理,语言模型为基于字的 Bigram、Trigram,分别用下面 3 种方法估计候选字的后验概率。

方法 1 采用经验公式(3),方法 2 采用经验公式(4),方法 3 是本文提出的统计方法,见式(6)。

表 1 为三种方法的后处理性能比较结果。其中,

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

文本识别率 = 1.0 - 处理后错误字符总数 / 总字符数

文本纠正率 = 1.0 - 处理后错误字符总数 / 处理前错误字符总数

十选校正率 = (处理后的文本识别率 - 原 SCR 首选识别率) /

(原 SCR 前十选累计识别率 - 原 SCR 首选识别率)

表 1 三种方法比较

方法	文本识别率		文本纠正率		十选校正率	
	Bigram	Trigram	Bigram	TrigramP	Bigram	Trigram
1	92.44 %	92.58 %	46.94 %	47.93 %	64.33 %	65.67 %
2	92.67 %	92.80 %	48.59 %	50.53 %	66.54 %	67.79 %
3	93.05 %	93.14 %	51.23 %	51.89 %	70.19 %	71.06 %

由上表可知,方法 3 的后处理效果最好,从而验证了用统计方法估计候选字后验概率的有效性。另外,基于字的 Trigram 模型的后处理性能总体上要好于基于字的 Bigram 模型,但改进不明显。

5.3 HMM 在文本识别后处理中的重要性

为了说明 HMM 模型在后处理中的作用,将 SCR 模型去除,利用纯粹的 Markov 语言模型进行后处理。语言模型为基于字的 Bigram、Trigram 时,上述 49 篇文稿的平均识别率分别仅为 76.89 %、75.07 %。显然,纯粹的语言模型会给文本识别后处理带来很强的负效应。可见,利用 HMM 进行文本识别后处理是十分重要的。

六、结 论

本文用 HMM 描述汉语文本识别后处理,将单字识别模型和语言模型两者结合起来,与只依靠语言模型相比,极大地改善了文本识别后处理性能。用统计方法估算单字识别后验概率较距离经验公式有明显的后处理纠错效果。汉语文本识别率的提高除了取决于单字识别首选识别率,还取决于:(1)精确地估计候选字的后验概率,(2)选择合适的语言模型。

参 考 文 献

[1] 夏莹. 基于统计的汉字文本自动后处理方法. 模式识别与人工智能,1996,9(2)

[2] Lee H J *et al*. A Markov language model in handwritten chinese text recognition. Proceedings of 2nd ICDAR, Japan, 1993

[3] Tung C H *et al*. Increasing Character Recognition Accuracy by Detection and Correction of Erroneously Identified Characters. Pattern Recognition. 1994, 27(9)

[4] Lawrance R Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 1989, 77(2)

[5] 吴佑寿, 丁晓青. 汉字识别的原理、方法与实现. 北京:高等教育出版社,1992

[6] Lei Xu *et al*. Methods of Combining Multiple Classifiers and their applications to handwritten recognition. IEEE System, Man and Cybernetics, 1992, 22(3)

[7] 陈友斌. 非特定人脱机手写汉字识别方法的研究[博士学位论文]. 北京:清华大学,1997

[8] Jelinek F. Self-Organized Language Modeling for Speech Recognition. Reading on Speech Recognition, 1990, 450～506

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net