基于语料库统计方法在汉字文本识别应用中的若干问题

夏莹 马少平 朱小燕 金奕江 姜哲 常新功
（清华大学计算机系）
（智能技术与系统国家重点实验室）

摘 要：为了提高汉字文本的识别率，本文讨论基于语料库统计方法在汉字文本识别后处理应用中的若干问题。分析语言模型的选择、统计标记集的选择、语料库的领域和大小、后处理的算法对处理质量的影响，并给出实验结果.

关键字：马尔科夫(MARKOV)模型 计算语言学 汉字文本识别后处理

Problems in the Application of Corpus-based Statistical Method to Chinese Text Recognition

Xia Ying, Ma Shao-ping, Zhu Xiao-yan, Jin Yi-jiang, Jiang Zhe,Chang Xin-gong,
(Department of Computer Science, Tsinghua University)
(State Key Laboratory of Intelligent Technology and System)

Abstract : In order to improve Chinese text recognition rate, some problems in the application of the corpus-based statistical method to the post-processing of Chinese text recognition have been presented in this paper. The effects of the selection of language model, the selection of statistial symbol set, the domain and the size of corpus and the algorithm of post-processing on the processing guality are analyzed. And the results of the experiments have been shown.

Key Words: Markov Model, Corpus Linguistics, Post-processing of Chinese Text Recognition

一、引言

目前，在汉字识别、语音识别、汉语拼音的整句输人、词性标注等研究中，基于语料库统计的马尔柯夫（MARKOV）方法都取得了很大成功，成为一个研究的热点。脱机手写汉字识别是文字识别最困难的一种，近年来，脱机手写汉字识别研究取得了很大进步，识别率在提高，前十选识别率可达98％以上，但首选识别率仍不够高。利用真实文本统计得到的上下文相关信息应用到文本识别后处理中，把一个汉语句子或短语作为一个处理单元，可以使汉字文本识别的正确率提高很多[8]...[14]，可达到实用的水平。该方法的框图如图1所示。把一个汉字文本识别系统视为两个部分：孤立汉字识别和自动后处理。一个识别系统将输人稿件经扫描、切分后，作整型归一化、特征抽取、匹配，识别出汉字及标点

*该项目得到国家自然科学基金69675004和863的资助

符号序列，并对每一个字给出多个特征相近的候选字和信度值；自动后处理则考虑可能所有汉字串序列，对每一字串进行概率计算，最后选择最佳的串为输出结果。并有"专业技术词表"机制，它的功能是允许用户建立自己的特定专业术语，如"断笔"、"脱机"等，其概率值与常用的同现概率值相当．在后处理时首先到该表中查找，然后再到同现概率矩阵中查找．

但是这种方法应用的效果如何？和很多因素有关，如采用的语言模型、统计标记集的选择、选用的语料库领域和大小、后处理的算法、文字识别的候选正确率、识别结果可信度与统计概率的配合等。我们希望真实文本统计的结果(相邻字字同现概率)占的空间不能太大、后处理的效果好、处理速度快，本文分析上述诸因素的影响，在不同的条件下应采用不同的策略。

二、语言模型的选择

基于语言模型统计的方法在汉字识别、语音识别的后处理和智能拼音输入法等中得到应用，语言模型是基于字还是基于词？我们认为要具体问题具体分析，在语音识别和智能拼音输入法中可能采用词间的模型较好，这是因为比较容易切分出词，而对于脱机手写汉字识别，由于识别率不够高，分词不可能都正确，我们认为基于字的语言模型较好．基于字的模型我们实验过三种：

1．基于字间二元语法模型[12]：仅考虑了相邻两个字之间的同现依存关系，对１５００万字的新华社通讯语料库进行统计，由结果来看，同现过至少一次的二元组数目仅占可能相邻两个字总数的３．８％。二元同现概率矩阵的数据约占４Ｍ字节。

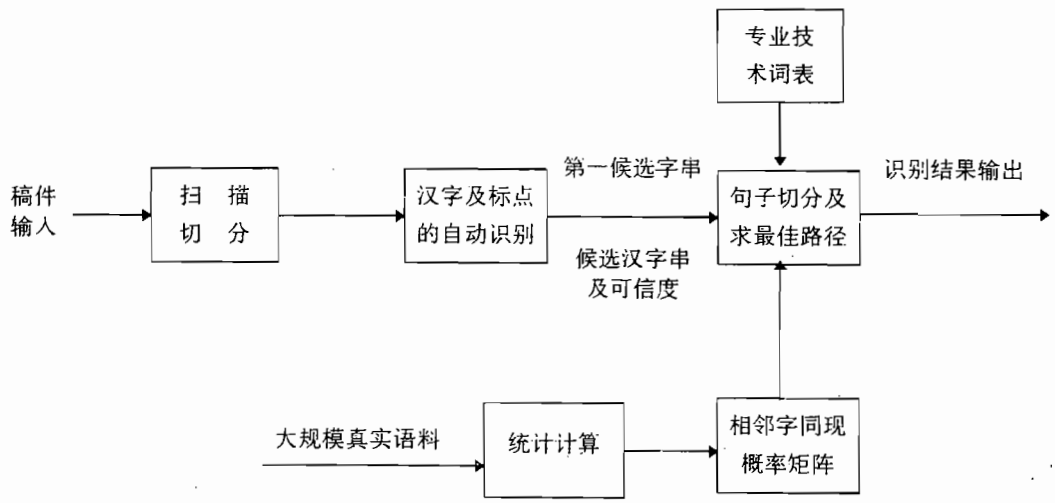


图 1　汉字文本识别后处理方法框图

二元语法模型利用相邻两个字之间的约束来确定最佳候选字，成功地解决了许多非首选的候选字确定。然而，在有些情况下仅有这种最近邻的约束关系是不够的。自然语言中同时存在近邻和远邻约束。 我们在实验中遇到的“字重迭复用”问题， 即一个字既与前相邻字组成一高同现概率两字词(或多字词中的两个相邻字)又同时与后相邻字组成另一

高同现概率两字词(或另一多字词中的两个相邻字)。例如: “赞成品”(赞成、成品), “空调整”(空调、调整)都是我们处理中遇到的字重迭复用。这种由于仅有二元近邻约束不能判断出高同现概率的二元组是否为两字词, 以及不能给出在不为两字词时其组成多字词的搭配范围。以一个实例来具体说明:

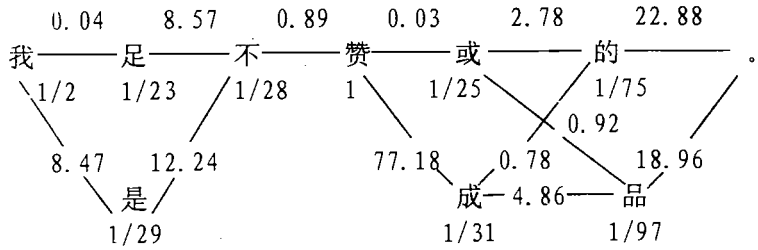


图2 路径图及数据

从上图实例的数据来看, 利用二元语法作后处理时, 很容易把“我是不赞成的”这句输出为“我是不赞成品”. 因这条路径有最大的累积同现概率。

用三元语法模型, 又多考虑了一个字, 能部分地解决这一问题。

2. 基于字间三元语法模型[12]: 考虑相邻三个字间的约束, 我们对1,500万语料作三元统计, 出现至少一次的三元组数目仅占总数目的0.03% 。其文件空间约15M字节。对某些问题比二元同现更可信, 后处理的效果更好. 上述的例子, 用三元语法模型处理时的参数如下:

(我 足 不)= 4.899704
(我 是 不)= 6.589819

(足 不 赞)= 2.849769
(是 不 赞)= 2.849769

(不 赞 或)= 1.231622
(不 赞 成)=11.129132

(赞 或 的)= 5.550718
(赞 成 的)= 8.389295
(赞 或 品)= 2.967883
(赞 成 品)= 4.750072

(或 的 .)= 7.174176
(成 的 .)= 9.561120
(或 品 .)= 8.086078
(成 品 .)= 9.261618

对上述数据概率计算, 很容易计算出三元同现概率累积数值最大的路径是“我是不赞成的”, 而不象二元语法模型那样得出“我是不赞成品”的错误结果。

虽然利用三元语法模型可以部分地解决上述问题, 但由于采用三元语法模型, 其数据空间庞大, 作后处理占的空间大, 处理时间加长, 用相同的动态规划法, 处理时间约加长20倍。

3. 基于词间字二元语法模型[14]: 统计是二元字字同现概率。后处理时以句子为单元, 对识别首选句子用最大匹配法分词,再用词间字二元语法模型求最佳路径。在识别首选句子可能存在错误的情况下, 分词不可能都正确, 只是把已经形成的词固定下来, 不会被误纠。在识别首选句子存在错误很多时, 基本无法分词, 相当是字间二元语法模型。例如:

后处理前: 更好地发挥工人阵级主力军钓诈同。
后处理后: 更好地发挥工人阶级三名军的作用。(没用最大词匹配法分词时,出现误纠)
更好地发挥工人阶级主力军的作用。(用最大词匹配分词时)

但是也有副作用，例如：识别“充分发挥职能部门的服务作用，”这个句子

后处理前：充分发挥职能部门的服劳作用，

后处理后：充分发挥职能部门的服劳作用，

“务”字的识别首选为”劳”字，”务”为第二侯选，因”劳作”是词，影响后处理时对”劳”字的纠正.

三、统计标记集的选择

为了满足真实文本的处理需要，即统计标记集应覆盖真实文本所有可能出现的符号，并使得统计标记集不至过大。我们曾经用两种标记集对语料库进行统计：

1. 采用3763个标记： 其中一级3755个汉字各为一标记。下边每组为一个标记：

①国标一级汉字之外的所有汉字

②阿拉伯数字（０－９）

③英文字母（Ａ－Ｚ）（ａ－ｚ）

④句边界类标点（。，！：；？）

⑤引用类标点左边部（　（〔〈〖“　等）

⑥引用类标点右边部（　）〕〉〗”　等）

⑦特殊数字标头符号（3.　(2)　⑧　等）

⑧其它符号

标记集只收录了一级汉字，这是考虑到一级汉字(3755个)可覆盖汉语普通文本的99.87%，其余所有的汉字为一个标记。另外，我们把不同的字母(A-Z)(a-z)，数字(0-9)和其它同类的符号分别归为一个标记是出于以下考虑：首先他们具有一致的同现搭配范围；再者它们对一确定的标记有一致的同现信息；把它们合并为同一个标记，不仅使得该标记与其它标记的同现概率更可信，而且可以有效地减少标记个数和计算参数空间。

2. 采用3801个标记： 其中3787个汉字是根据“现代汉语频度字典”选择频度高的汉字，各为一个标记，其余所有汉字为一个标记。这样作是由于一、二级汉字的划分并不十分合理，有些高频汉字并不在一级汉字中，一级汉字中有些汉字并不常用。英文字母、数字、标点符号等与上述1.同。

实验结果是第二种标记集效果更好些，高1%左右。

此外还有用基于字类（Word Class)的[6]，真实文本统计得到的同现概率距阵会小，但在作聚类时要花很多时间。

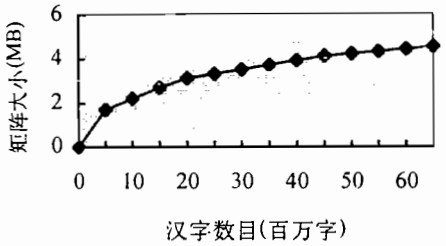
四、语料库的搜集和影响

基于语料库统计的方法，其统计得到的概率距阵与所用的统计语料库的大小、领域、时期、各种语料的比例等是有关系的，影响同现概率距阵的大小和质量。例如：

1. 同现概率矩阵与统计语料库大小的关系： 用计算机领域(90-94的年中国计算机报、计算机用户月刊、计算机世界周报、计算机世界月刊)的语料作字字同现概率的统计，随着语料库字数的增加同现概率矩阵的大小在增加，但矩阵达到4M以上时增加缓慢，如图3所示。

2. 统计语料库领域对后处理效果的影响： 用统计新闻领域语料库得到的同现概率距阵对计算机领域的文章作后处理，效果相对差一些，低1-4%. 如表1所示。主要是专业名词没有。例如：下面例句后处理的错误是因无" 句法" 词。

图3　同现概率矩阵与语料库大小的关系



后处理前：讲多荤若利闸结构模式识别方陷包括匈法模式识别方法，
后处理后：许多学者利用结构模式识别方法包括司法模式识别方法，(用新闻领域语料库得到的字字同现概率矩阵)
许多学者利用结构模式识别方法包括句法模式识别方法，(用新闻及计算机领域语料库得到的字字同现概率矩阵)

表1　统计语料库领域对后处理效果的影响

序　号	识别首选识别率	用新闻领域语料库得到的同现概率距阵作后处理	用新闻及计算机领域语料库得到的同现概率距阵作后处理
1	83.2%	93.1%	95%
2	89.4%	95.4%	98.1%
3	89%	94.1%	97.5%
4	88.9%	94.2%	96.1%
5	90.8	95.4%	96.5%

五、后处理算法的影响

利用二元、三元同现概率矩阵和汉字识别的可信度作后处理曾经采用过两种方法：

1. 动态规划法全程求最佳路径[12]：这种算法对脱机手写汉字文本识别的后处理的效果很好，尤其在第一选识别率不高而前十选识别率较高时，全程求最佳可以使连续多字识别首选不正确的情况而得到正确的后处理结果。例如：

后处理前：札会误干，
后处理后：在会谈中，

后处理前：是建设有中国特邑社余立叉伟大事业的主力军。
后处理后：是建设有中国特色社会主义伟大事业的主力军。

后处理前：但舍普质俘到明显改善.

后处理后：但合音质得到明显改善.

后处理前：曲守光照、
后处理后：由于光照、

2. 从首选字起始双向搜索局部寻优[10]：对于印刷文本识别作知识后处理，首选识别率已经很高时再进行语言后处理，虽然总体效果有提高，但存在误纠现象，用双向搜索局部寻优方法要比动态规划法效果要好一些，误纠现象可以减少。

当然，每种算法都有识别侯选字的可信值与同现概率的参数配合问题，只有参数配合适当时，后处理效果才好.

六、结论

用基于马尔科夫(MARKOV)模型的真实文本统计方法对汉字文本作自动后处理，是以一个汉字序列（多数情况为一个句子）作为处理单元. 要使实际的处理效果好，要在语言模型的选择、统计标记集的选择、选用的语料库领域和大小、后处理的算法等方面根据识别系统的情况进行分析，采用不同的策略，有时要采取折衷方案.

参考文献

[1]. George Nagy "At the Frontiers of OCR" Proceeding of IEEE Vol.80 NO.7 1992

[2]. R.M.K.Sinha "Rule-based Contextual Postprocessing For DEVANAGARI Text Recognition " Pattern Recognition Vol23 No5 1987

[3]. R.M.K.Sinha etc "Hybrid Contextual Text Recognition with String Matching " IEEE Trans. on PAMI VOL.15 NO.9 1993

[4]. 白栓虎 "基于统计的汉语语料库词性自动标注方法的研究与实现" 清华大学 硕士论文 1992

[5]. 施得胜等，基于统计的中文错字侦测法，Procedings 1992 International Conference on Chinese Information Processing(1), Beijing, China

[6]. Chao-Huang Chang, " Word Class Discovery For Postprocessing Chinese Handwriting Recognition", Proc. COLING 94, Japan, 1994

[7]. Xiang Tong and David A. Evans,"A Statistical Approach to Automatic OCR Error Correction in Context",Proceedings of The Fourth Workshop on Very Large Corpora,4 August 1996,Danmark

[8]. 常新功 夏莹 余骏 金奕江,"利用上下文知识的汉字文本识别系统","智能接口与智能应用'93"学术会议论文集，P. 45-48,1993 7,哈尔滨

[9]. Xia Ying Chang Xingong "Research on Corpus-based Text Recognition " NLPPR'93 Dec 1993 Japan

[10]. 常新功 夏莹，“基于N元语法的文本识别后处理中的局部寻优方法”，第五届全国汉字及汉语语音识别学术会议 论文集，P. 231-237, 1994.9,成都

[11]. 夏莹 马少平 常新功 朱小燕 金奕江，"脱机手写汉字文本识别",智能计算机接口与应用进展，P33-38, 清华大学出版社, 1995

[12]. 夏莹 常新功 马少平 朱小燕 金奕江，“基于统计的汉字识别文本自动后处理方法”，模式识别与人工智能，第9卷第2期，P. 172-178, 1996 6

[13]. 夏莹 马少平 孙茂松 朱小燕 金奕江 常新功，“脱机手写汉字文本识别的自动后处理”，Proceedings '96 of International Conference on Chinese Computing, P413-418,June 4-7,1996,Singapore

[14]. 李国华 夏莹 马少平 孙茂松 朱小燕 金奕江，“基于词间字二元语法模型的汉字识别后处理方法”第六届全国汉字识别学术会议 论文集，P.181-186,1996.9 重庆