中文字字同现概率统计及应用

夏莹　常新功　马少平　朱小燕　金奕江
清华大学计算机系(100084)

摘要：中文字字同现概率反映中文文本中汉字的相邻关系，这种中文上下文相关信息不仅对汉字文本识别、语音识别、中文计算机校对等领域有很大作用，而且对语言学研究也有一定意义。本文介绍中文字字同现概率统计的方法，并分析统计结果。将该字字同现概率用于汉字文本识别，把具有确定性边界的一个汉字序列（多数情况为一个句子）作为一个处理单元，利用统计获得的字字同现概率，采用动态规划方法，对手写汉字识别文本进行自动后处理，获得了令人满意的效果。

STATISTICAL METHOD AND APPLICATION OF CO-OCCURRENCE PROBABILITIES BETWEEN CHINESE CHARECTERS

Xia Ying, Chang Xin-gong, Ma shao-ping, Zhu xiao-yan, Jin yi-jiang
Department of Computer Science, Tsinghua University

ABSTRACT: The Co-occurrence probabilities between Chinese characters represent the adjoining relations of Chinese characters in Chinese text. The contextual informations play an important pert not only in Chinese text recognition, speech recognition, computer revise, but also linguistics research. In this paper, the statistic method and the results of the co-occurrence probabilities between Chinese characters have been introduced. The statistic results have been applied to Chinese text recognition. A bounded sequence of Chinese characters (more often, a sentence) is processed as an unit. And the co-occurrence probabilities between Chinese characters and dynamic programming strategy are employed. For Chinese text, a post-processing is automaticly processed. The satisfactory Chinese text recognition results are acquired.

1. 引言

近年来，汉字ＯＣＲ识别研究取得了很大成就，许多商品化的系统已成功地推入市场并获得了很大的经济效益和社会效益。然而，人们已看到单纯的单字（Isolated character）识别的方法对整个文本识别是不够的，而合理地利用上下文相关信息来提高汉字文本识别率是一个很重要、很有意义的研究方向〔１〕。集成自然语言上下文知识和其他知识在国外文本识别研究中也越来越受到重视〔２〕〔３〕〔４〕〔５〕。国内前几年，一些学者对汉语文本识别处理中加入联想、词组信息做了有益的探索。在计算语言学方面的词性标注、语义排歧等方面，基于统计的MARKOV语言模型方法都取得了很大成功，成为一个研究的热点。中文字字同现概率可以反映文章中汉字的相邻关系，它不仅对汉字文本识别、语音识别、中文校对等领域有很大作用，而且对语言学研究也有一定意义。汉字的数量很大，建立汉字相邻关系的MARKOV模型比英文困难得多，我们建立了汉字文本统计系统，获得二元、三元字字同现概率矩阵。并把这种统计的结果应用到汉字文本识别处理中，

实验结果证明了这种统计的方法对于文本识别处理同样是很有效的。

2. 文本统计系统的组成

文本统计系统主要由三大模块组成：文本统计计算模块、统计数据管理模块、二元三元字字同现概率矩阵。其中文本统计计算模块要在工作站上运行；后两个模块则在工作站（ＵＮＩＸ操作系统）和微机（ＤＯＳ操作系统）上各有一套版本。全部代码用Ｃ语言写成。系统的各模块关系如下图１所示。

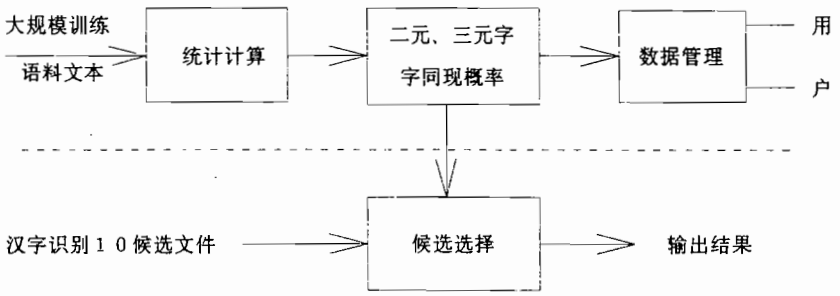

图1　　文本统计系统组成及应用

. 文本统计和概率计算部分

本部分首先对待统计处理的文本做了必要的预处理(包括去掉不必要的通讯稿头信息等等)规范化；然后统计单字出现次数，二元、三元组字字同现次数；再利用统计所得次数计算出同现概率值。

. 统计数据管理部分

本模块提供用户一个管理、查询、增删同现概率的接口。

主要功能有:

. 查看某一、二、三元组的同现概率;
. 浏览某一汉字与其它汉字的所有同现过的二、三元组及概率值;
. 浏览所有同现概率值大于任一域值二、三元组;
. 添加二、三元组或增加其同现概率值;
. 删减二、三元组;
. 专用术语表机制;

对大规模真实文本进行统计是基于统计的知识后处理的第一步。我们分别做了二元字字同现和三元字字同现的概率统计工作。我们用该系统对1,500万汉字的真实文本新华社通讯稿进行统计。

从我们的统计处理实验中发现,统计获得的中文字字同现概率矩阵随统计文本的领域不同而有很大不同。因而，我们从普通领域文本(新闻通讯稿)中统计的数据就不能满足特定专业领域的需要。从科技领域文本的处理来看，专业名词和术语是一个关键问题 。例如,在一篇关于汉字识别的文章的处理中，在处理“断笔”、“微处理器”、“硬件”等常用词时都遇到了问题。因而，对于二元语法模型，我们实现了一个 专用术语表机制，可为许多不同领域的用户建立其自已的专业术语表。在处理该领域的文本时只需加载相应术语表即可 。

若用户加载了专用术语表则在计算处理过程中任何读取访问字字同现统计数据的操作首先要到该表中查找，在一定程度上解决了大量的专业名词术语由于在普通文本统计所得的统计数据库中有较小的同现概率而导致的结果不可信问题。

3. 文本统计

3.1 统计标记集的定义

为了满足真实文本的处理需要，即统计标计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。我们的统计采用了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 一级之外的所有汉字
② 阿拉伯数字（０－９）
③ 英文字母（Ａ－Ｚ等）
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

我们的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%。另外，我们把不同的字母(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。

3.2 二元同现概率

如上所述，３７６３个标记的统计需要一个３７６３×３７６３大小的矩阵来存储任何两个标记组合的同现信息。当每一个元素仅占两字节时（而实际上，同现概率为浮点型，至少需４字节），则需至少２５Ｍ字节内存。通过小规模（１０万字）文本的统计实验，我们发现二元同现概率矩阵是一个稀疏矩阵。对1500万语料库的统计结果来看，同现过至少一次的二元组数目仅占3.8%　。对稀疏矩阵我们采用了链表的结构来记录同现数据，采用此结构内存仅占不到4M.

3.3 三元同现统计

3.3.1 问题的提出

二元同现仅考虑了相邻两个字之间的同现依存关系，在有些情况下仅有这种最近邻的约束关系是不够的。三元字字同现考虑相邻三个字间的约束，对某些问题较二元同现更可信。例如：在二元同现统计中，有“阿拉”二元组同现次数达2240次，而同时“拉”字打头的高频二元组有很多，如：拉克(1306次)、拉伯(1601次)、拉萨(1041次)、拉夫(878次)

、拉瓜(759次)、拉法(494次)、拉丁(169次)、拉开(292次)等。若仅考虑二元同现约束，则“阿拉克”、“阿拉萨”等三元组都被认为有很强的同现依存关系，然而若用三元同现约束，则仅有“阿拉伯”、“阿拉法(特)”、“阿拉斯(加)”和“阿拉塔”四个三元组是合法高频三元组。故而，我们在二元统计的基础上又进行了三元字字同现统计。

3.3.2 统计方法和数据结构

我们的标记共有3763个，在进行三元统计时，其可能的三元组个数为3763×3763×3763个，若每一元素同现信息需2个字节存放的话，则需要近100G的内存容量。我们采用了两个阶段分步统计的方法，并用带索引的链表结构来记录统计信息。第一步首先统计二元同现；第二步，所有出现过至少一次的二元组做为统计矩阵的列元素，这里即为链表头指针数组元素；以3763个标记做为矩阵的行元素，进行三元组的同现统计。这时采用了索引，因使得做为矩阵列元素的二元组的查找、定位速度很快。 三元组(tag1 tag2 tag3)同现的次数，对1,500万语料的三元统计，我们发现出现至少一次的三元组数目仅占总数目的0.03%。采用上述链表结构，只需30M左右的内存资源。三元组统计数据的磁盘文件结构类似于二元组统计文件。其文件空间只有15M字节。

4. 统计结果分析

我们对统计所得同现概率矩阵进行了分析，发现统计有效地反映了中文文本中汉字的相邻关系和某些规律。

. 较好地反映了常用词条信息:

把统计中出现的高频二、三元组与词典对比，我们发现汉语中常用的词，其对应的元组都有较高的同现概率。不仅两字词、三字词，多字词的相邻字两两同现概率也同样很高。因而可以说统计获得的同现概率矩阵包含了常用汉字的前联想、后联想汉字集合。例如，“暂”字为首的词条，在《新编实用汉语词典》中收入了“暂时、暂且、暂行、暂缓”4个两字词词条，该4个词条在统计中都出现过很多次，而“暂停、暂不”也具有很高的同现概率。这说明“暂停”等类似的未收入词条在现代新闻通讯中用得很频繁。

. 反映出非词的常用搭配知识

在统计结果中同现概率很高的二元组中除词条知识外还包括了许多常用搭配知识。在中文中有些虚词可与某些字结合出现，没有实词含义，仅起语法作用或表示程度、方式、状态等。例如：“把这”、“得很”、“成了”、“不可”、“我们的”等元组就具有很高的同现概率。诸如“把… ”，“被 …”，“… 的”等部分常用搭配可由统计结果反映出来。

还有一类字常与数字搭配。例如，“仅···”，“有···”，“长···”，“几分之···”，“···千”，“···年”，“···公里”等字。这类搭配关系同样也可反映出来。

. 一定程度上反映了词词间的同现关系

在统计语料中，常出现的词间同现在统计数据中反映为前一个词的尾字与后一个词的首字之间有较高的同现数据值。这种二元字字或三元字字同现元组就部分地反映了词间同现，然而，这种词间依存的反映是不完全可信的。

. 常用句头字、句尾字信息

我们的二元统计结果能得出一级汉字中有1270个字（如峦、虑、肪等）和二级所

有汉字不做句首字。而最常出现于句首的字（平均至少每一万字文本中做一次句首字）有85个（如并、不、从、而等）。同样，统计结果也反映了句尾字类似特征。

5. 汉字文本识别

我们把汉字同现概率的统计结果应用到汉字文本识别中，方法如下。

5.1 汉字文本MARKOV模型

把一个文本识别系统视为两个部分：ICR（Isolated Character Recognition）和LM（Language Model）。一个识别系统要从一输入符号串S（S1 S2 S3……Sn）识别出汉字序列W（W1 W2……Wn），ICR部分首先做特征抽取、分类，对每一个Si给出对应的多个特征相近的候选字集；而LM部分则考虑可能对应的所有汉字序列，对每一字串进行概率赋值，最后选则最佳输出。

我们来形式化地描述：

S=<s1 s2 … sn>；　为一可确定边界的字符串；

WS=<w1 w2 … wn>；　表示一可能的汉字串；

P(WS|S)　表示输入S时，结果为WS的概率；

当 P(WS^|S)=MAX(P(WS|S))，WS^ 即为识别结果。

由BAYES公式得：

$$P(WS^{\wedge}|S)=MAX(P(WS|S))=MAX(P(WS)*P(S|WS)/P(S)) \quad (1)$$

由于输入S已定，故P(S)项不影响选择，可不考虑。P(S|WS)项一般用ICR给出的信度值CF(Wi)等来代替（如后边的例子所示）。我们考虑下式：

$$WS^{\wedge} = P(WS)*CF(Wi) \text{ 为最大的WS} \quad (2)$$

我们这里视中文句子为MARKOV源（即一个状态的发生概率仅与其以前的状态有关），（2）式中的P(WS)可写为如下：

$$P(WS) = \prod_{i=1}^{n} P(w_i|<w_1 \cdots w_{i-1}>) \quad (3)$$

若采用一阶MARKOPV模型，即二元语法，$<w_1 \cdots w_{i-1}>$等价于w_{i-1}，即：

$$WS^{\wedge} = MAX \prod_{i=1}^{n} (P(w_i|w_{i-1})*CF(W_i)) \quad (4)$$

5.2 数据稀疏和同现概率计算

在训练语料不足或参数空间庞大的情况下(我们的字字二元同现、三元同现统计即为此情况)，会遇到数据稀疏(data sparsness)问题：即许多合法的在未来的文本中要遇到产标记同现现象在统计语料中从未曾出现过。因而当遇到新的同现标记n元组时，将出现零概率。对于合理地平滑处理数据稀疏的估值算法有很多，一种较简单的解决此问题的方法是：在统计数据很充分时，我们直接用n元组同现概率进行计算；若统计数据不充分时，确切地说，不可信时，我们宁可回到n-1元组甚至n-2元组来计算。实际应用中，是用它们的线性组合来实现的。在我们的模型中即采用了此方法来计算同现概率。

对于二元语法模型，其计算公式如下:

$$P(w_i|<w_{i-1}>) = \lambda_1 * F(w_{i-1}w_i) + \lambda_2 * F(w_i) \qquad (5)$$

其中 $\lambda_1 + \lambda_2 = 1$;

上述F( )的定义为:

$$F(W_{i-1}W_i) = N(W_{i-1}W_i) / N(W_i)$$

$$F(W_i) = N(W_i) / NT$$

N( ): ( ) 在语料库中出现的次数;

NT: 训练文本规模（总字数）;

λ_i: 我们系统中从工程角度出发，由试验选择。

5.3 动态规划法求最佳路径

结果句子的选择目前常采用的搜索算法是称为Viterbi的动态规划方法。其基本思想是把求解整个问题的最佳解归结为求解其子问题的最佳解。

5.4 实验结果

利用统计方法的文本识别最终识别率与候选字识别率的关系很大。因此我们用校正率、处理后的识别正确率、处理正纠率、处理误纠率等指标来评定文本的处理效果。

$$\text{校正率} = \frac{\text{处理后的识别正确率} - \text{原ICR首选正确率}}{\text{原ICR候选正确率} - \text{原ICR首选正确率}}$$

其中，处理后的识别正确率是指经语言学处理后的输出文本正确率；正纠率与误纠率之差是识别率提高（处理后的识别正确率－原ICR首选正确率）的百分点；处理正纠率是指把原首选错误的字纠正的百分点；处理误纠率是指把原首选正确的字改为错误的百分点。利用2元语法模型下的动态规划算法对脱机手写汉字文本作自动后处理，其识别正确率平均能提高15%，校正率平均在70%左右。

参考文献

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

〔2〕 Yukiyasu IIDA "Knowledge processing for an OCR" REVIEM of the Eltrical Communication Laboratories Vol.32 No.5 1984

〔3〕 C.J.Wells L.J.Evett "Fast dictionary lookup for contextual word recognition" Pattern Recognition Vol.23 No.5 1990

〔4〕 E.M.Riseman "A Contextual postprocessing system for error correction using binary n-gram" IEEE Trans. on Computer Vol 22 No5 1974

〔5〕 R.M.K.Sinha "Rule-based Contextual postprocessing for DEVANAGARI text recognition " Pattern Recognition Vol.23 No.5 1987