汉语文本校对字词级查错处理的研究

孙才　罗振声
清华大学中国语言文学系　100084

摘要　本文对于汉语文本查错的问题和现状进行了介绍，对于查错与纠错的策略和方法进行了分析，并针对键盘输入文本中的错误设计了一个自动查错和确认纠错系统。

Research On the Lexical Errors in Chinese Text

Sun Cai　Luo ZhenSheng
Department of Chinese language and literature , TsingHua University

Abstract This paper present a survey on the automatic detection and correction of Chinese text errors . and an experimental system to detect and correct errors in keyboard-in Chinese text .

一、引言

汉语文本的计算机辅助校对系统的研究，是当代出版与办公自动化技术迅速发展与现代化提出的亟待解决的重要课题。然而，通过近几年来的实践，人们发现这是一个具有相当难度的课题。

原因是多方面的，汉语理论研究，特别是自然语言技术研究在许多方面尚不成熟，而对基于错误的文本困难则更多；汉语言及汉字的特殊性，远远不同于西方拼音文字，后者拼音校对系统已达到实用化水平；汉语文本错误产生的原因与类型则具有自己独特的规律，但尚未被完全认识；最后，文本校对工作在计算机内部不可能有正确的文本供以对照检查，而只能采取从语法，语义等语言学分析与统计方法进行间接判断查错处理，本身就很难高准确度地再现原来文本。

因而，完全的自动校对的研究则更是困难的。当前研制一种人机交互式的辅助校对系统，达到一定校对目标，并不断提高水平，还是可取的。

我们通过不同手段，包括句法与句子成分分析，词性邻接关系检查，语义测试等，目前看来效率较低。根据文本录入常常引起局部性错误的事实，本文重点讨论了基于分词的字词级查错处理策略。尽管目前工作尚不能令人满意。然而，课题在不断发展，这就是希望。

二、文本中常见错误的分析及统计

1 、文本常见错误分析

国外一些研究者曾对于英文的拼写错误进行了详尽的统计和分析，并对基本的拼写错误进行了分类（ Pollock & Zamora 1983 ）。借鉴这种分类方法，我们将收集到文本中的错误进行了类似的分类统计。对于字词级的错误进行了以下分类：其中带下划线的部分是错误的部分，其后括号中为正确文本。

1)错字：将某字误录为另外一字

例：戈林一生所作所作（为）充分体现了希特勒帝国的暴力、复仇、种族灭绝的强盗法则。

十分同情劳动大众的不境（幸）遭遇。

错字有一种情况是错误单字与前一字或后一字相同：

例：最后一页未翻还给作作（者）并在上面写了“很好”的评语。

为这一时刻的到来，我们已经待（等）待很长时间了！

2)多字：录入了多余的字

例：运用他的天才智商和其他它的能力。

能同时要当好维也纳的长官如此重要的职务吗？

多字有一种情况是多余单字与前一字或后一字相同：

例：然后他他拔出腰里的手枪，朝着镜子连打三枪。

社会主义从空想到科学是由马马克思和恩格斯来完成的。

3)少字：漏录了原文中的字

例：而是时代变和世界矛盾转移所造成的合乎逻辑的结果。

为了加速我国经济和社会事业的发展，提了对外开放政策。

4)易位：将原文中相邻的两个字换位

例：他喜欢假装他是一个待人温和人的。

隆美尔整理了他的讲课记录，然后又改写了为一部激动人心的书。

5)多字词替换：

例：原有的地区冲突和热点总理（问题）相继解决和即将解决。

对于大规模真实广西（文本）的研究。

6)其它：不能归类于以上几种基本类型的错误

例：如何处理好马克思主义趔（原理）与我国具体国情、书本与实际之间的关系。

后来他又把矛头指出和（向）同盟军和苏联。

2、错误类型分类与统计

本文把从出版社及照排部门等得到的刚录入完，但尚未经人工校对的语料文本中的常见错误进行了分析与统计。其结果如下：

表一　　常见错误类型统计

错误类型	错误数	错误百分比	
错字	628	60.8	62.01
错字（重字）	12	1.16	
多字	66	6.4	9.5
多字（重字）	32	3.1	
漏字	175	16.86	
易位	8	0.7	
多字词替换	22	2.71	
其它	42	8.14	

由上表分析，对汉语文本中的常见错误，大致有以下规律：

(1) 文本中字词级错误，大部分由单个字错误引起，即属于前四种基本类型的错误，约占总错误出现率的 88.7%
(2) 在单个字错误中，错字是最常见的，约占总错误出现率的 62.01%
(3) 在多字错误中，重字是一种比较常见的情况，占多字错误的 32.65%
(4) 易位错误在汉语文本中很少发生。这是相对于英语中字母间常发生易位错而言的。

对上述错误进一步分析，还可看出以下特点：

(1)错字通常不与相邻的字或词构成词。

(2)错字通常由两种情况造成，一种是由于输入码相近；另一种是由于看错了字。在前一种情况下，输入码基本上都是由字母级的基本错误，即错字母，多字母，少字母或字母易位。

(3)不同类型的输入方法，如拼音类与形码类，其典型的错误差别较大。

三、查错与纠错策略

对于文本中一个给定的句子，如何才能分析并断定其是否包含有错误是非常困难的。显然，在计算机中不可能有原文进行对照检查分析的情况下，唯一的可能只有通过对句子进行语言学分析来观察其是否在某方面存在异常。运用规则或统计的方法来对句子进行分析，从而找出其可能出现错误的地方。因此需要对错误产生后引起句子在词法、句法、语义等各个层次上的异常分别进行分析。然后才能给出查错与纠错的策略。

1、错误引起的异常

在文本中当错误发生以后，通常会对周围上下文环境产生一定的影响。使句子在某个层次上产生异常。本文对于收集到的文本中的实例进行考察后，总结出以下关于错误产生后的现象：

1)非词单字（句子分词后）

例：他 到 南美 找 了 一 个 自 认为 相当 笱 职 的 位置 。

2)连续的单字词（句子分词后）

例：保卫 和平 的 伟大 力量 和 中 硫 砥 柱 。

3)相邻字词之间同现频率很低

例：五 名 妇女 充当 了 牺牲 吕。

4)相邻的重复单字词

例：在 新 的 攻击 之 之 带领 士兵 顺利 地 撤 出 阵地，

5)句子结构不完整，成分残缺

例：他 是 这样 （说）的 ，也 是 这样 做 的。

6)句子成分之间搭配不当。

例：不应 由 一 个 人 承担 ，而且 应 由 整个 民族 来 承担。

在以上几种情况中，前四种属于局部性的异常，而后两种则属于全局性的异常。在实际出现的由输入引起的错误中，引起局部异常的情况占绝大多数。

2、查错的策略

由以上分析，异常分为局部性和全局性两类，实践证明引起局部异常，是文本错误中的主要现象。对局部性异常，目前一般采用字，词级查错处理，其主要方法是通过对输入文本进行分词及词性标注处理，然后应用规则的和统计的方法进行分析处理。这是当前汉语文本校对技术采取的主要技术手段和方法。而全局性错误，如成分残缺，句法语义异常，分析与处理情况就复杂得多，困难较大。实践表明，目前相对来说效果差，效率低。这主要是由于当前基于正确文本的句法，语义分析技术不成熟，不过关。运用于错误文本的校对，尚有不少问题要深入探讨。因此，本文主要讨论前一种情况。

1)规则的方法

在拼音文字中，对于“上下文相关”错误的分析技术采用的是所谓“基于松弛”的分析。其原理是对于给定的句子进行分析时，如果分析失败，则认为句子有错。对于错误的定位则是通过放宽对于某条规则的限制，从而使得句子的分析能够获得成功来进行的。然而，汉语的情况比较复杂、比较特殊，对于错误文本的句法、语义分析尚有待进一步的探索。

本文认为，对于某些语言现象，规则的方法还是很有效的。本文对以下情况进行了处理：

i)数词－量词－名词短语成分残缺或搭配不当

例：他亲自为一年青犯人行刑。

ii)连词搭配不当

例：空想社会主义者之所以是空想社会主义者，因为在资本主义生产还很不发达时代，他们只能这样。

iii)介词短语成分残缺或搭配不当

例：在错误的，他越走越远。

2)统计的方法

统计的方法是一种不依赖于知识推理的方法，因而有其独特的优势。而运用统计的方法是一种基于概率的方案，因而误报是不可避免的。在统计应用中，一般是在若干个候选方案中进行选择时用统计的方法计算不同方案之间的优劣，比如在词性标注， OCR 等。而在本文所针对的键盘输入文本却不存在这种侯选，针对这种情况采取了一种设定阈值的方法。

本文在查错时采用了以下几种统计信息，其中括号内为统计用语料规模：

1)字频 (约五千万字语料)

2)二元字字邻接矩阵（约五千万字语料)

3)二元词性邻接矩阵（约四百万词经词性标注过的语料）

4)二元词间字邻接矩阵（约五百万词经自动分词的语料）

设文本中的句子为 S = C1C2 …… Cn ，经分词后 S = W1W2 …… Wm 。其中 Ci 为第 I 个字， Wi 为第 I 个词。

对于字段 Ci …… Cj 计算以下数据：

字段平均字频： F1 = （ ZF(Ci)+……+ZF(Cj)）/(j-I+1)，其中 ZF(Ci)为第 i 个字的字频;

字段转移概率： F2 = P （ Ci|Ci+1 ）*……*P(Cj-1|Cj) ，其中 P(Ck|Ck+1) = R(Ck|Ck+1)/(R(Ck)*R(Ck+1)), 其中 R(Ck|Ck+1) 为二元字字同现次数，R(Ck)为字频。

对于词段 Wi …… Wj 计算如下数据：

词间字转移概率： F3 = P （ Ii|Ii+1 ）*……*P(Ij-1|Ij) ，其中 P(Ik|Ik+1) =

R(Ik|Ik+1)/(R(Ik)*R(Ik+1)), 其中 R(Ik|Ik+1) 为二元词间字同现次数， R(Ik) 为词间字频。

词性转移概率： F4 = P （ Ti|Ti+1 ）*……*P(Tj-1|Tj) , 其中 P(Tk|Tk+1) = R(Tk|Tk+1)/(R(Tk)*R(Tk+1)), 其中 R(Tk|Tk+1) 为二元词性同现次数， R(Ck) 为词性在统计语料中出现的次数。

对于一个字段计算其 F1 值及 F2 值，对于一个词段计算其 F3 及 F4 值。如果某一值低于正常范围，则怀疑此字段或词段有错，并进行进一步的分析与判断。这里问题的关键是确定“正常范围”。本文采用的方法是对于每个结果值设定--个阈值，如果计算所得结果低于此阈值，则认为其可能有错。

阈值的选择是一个很重要的问题。在实际系统中，选择阈值是根据实际效果不断调整的，以在系统查错能力和误报率之间取得平衡。阈值偏高会导致系统查错能力下降，偏低则导致较高的误报率。

3)进一步的处理：

由以上判断所得到的结果是很粗略的，需做进一步的分析。设对于某一给定字段，其 F1 或 F2 值低于正常阈值。则对字段中的每一个字寻找可能的候选字，以决定是否存在一个更合理的近似字段，如果存在就认为其是正确字段。

对于字段中每一字 Ci，其可能候选是由以下途径产生的：

i) 常见错误字

例： （地 的 得）（奉 俸 捧 棒）

每个括号之间是互相容易错的字

ii) 输入码近似的字

如在五笔字型中，以下几组字：

（境 fuj 幸 fuf 培 fuk 增 ful ） （录 vi 当 iv ）

（抛 rvl 招 rvk 挪 rvf ）

iii) 在待测试文本中与前一字或后一字共现频率高的字

本文在实验系统中维持一个临时的字字邻接频率表，以记录动态的邻接关系。

对于字段的每一可能替换用动态规划的方法寻找最优的新的 F1 及 F2 值，如果明显好于原 F1 及 F2 值，则认为此替换为正确文本。

很明显，这样在查错的同时给出了可能的纠错方案。要说明的是这种方法只有对错字这种情况有效。而对于多字、少字等情况基本上没有校错的能力，而只有报错的能力。

4)针对性的处理

对于某些有特殊规律的错误，本文还采取了一些针对性的处理机制。主要是：

i) 在分词阶段采用模糊匹配以处理长词的校对

主要是为了增强系统处理长词错的能力，在自动分词进行最大匹配时对于三字以上词采用模糊匹配。如果模糊匹配成功，则认为在录入时产生了错误，实际应该是该长词。

例：充分暴露出资本主义手产关系已经不适应生产力的发展。

对“手产关系”与“生产关系”进行模糊匹配，则认为是在录入“生产关系”时，将“生”录入为“手”造成的错误。

ii)由于多字或错字造成的重字

这种情况主要是为了增强系统对这一类比较常见错误的处理能力。实际中错误的产生有着一些认知心理方面的规律很值得去探讨，其中重字就是一个很明显的例子。汉语中有一部分单字是可以叠字成词的，而大部分单字是不能这样的。

本文对于构成重字的单字进行检查，如果其不能构成叠字词则认为是由于错误造成的。

iii)由于少字而造成的组词能力强，独立成词能力差的单字词

这种情况主要是指在录入多字词时由于少字而造成的这类单字词。如：

例：博尔曼的凶残和粗俗尤为突。

在词频统计中，“突”字作为单字词，其词频为1；而做为多字词的一部分共出现了602次。本文对于可成词单字分别统计了其作为单字词及作为多字词中的一部分的频率。并把其中作为单字词的频率远低于（10倍以上）其组词频率的单字归为这一类情况。若文本在经自动分词后出现了这种单字词，就认为其是由于错误造成的。

系统对于多字，少字这样的基本错误处理能力较弱。后两种方法作为对基本方法的补充，而且很有效。

四、实验结果

本文在 Pentium90 上，采用 Visual C ++语言在 WINDOWS 环境下实现了一个实验系统。本文对采集到的真实文本抽样测试了约十万字的语料，其校对速度为 25 字/秒。其中实际包含 102 个文字性错误。系统共报错 323 次，其中 71 次为真正的错误；提供纠错侯选 212 次，其中 54 次为对于真正错误的纠错，并且有 49 次为第一候选正确。现给出系统的测试数据：

报错召回率 = 文本中的错误被发现的比例 = 71/102 = 70%
报错精确率 = 报出的错中真正为错误的比例 = 71/323 = 22%
纠错精确率 = 纠正的错误中为正确的比例 = 49/212 =23%
纠错召回率 = 对于真正的错误的正确纠正比例 = 49/54 =90%

在系统的精确率和召回率中，本文主要追求的是召回率。应该说系统的误报率还是比较高的。

本文工作还是初步的，有许多问题还需要深入探讨，特别是句法，语义分析技术的研究及其在错误文本处理中的应用，更是提高课题水平的重要环节。

参考文献

[1] Karen Kukich 1992 Automatically Correcting Words in Text *ACM Computing Surveys* 24(4) 376-439

[2] Lance A.Ramshaw 1994 Correcting real-word spelling errors using a model of problem-solving context *Computational Intelligence* 10(2) 185-211

[3] J.J Pollock & A.Zamora 1983 Collection and Characterization of Spelling Errors in Scientific and Scholarly Text *Journal of the American Society for Information science* 34(1) 51-58

[4] J.J Pollock & A.Zamora 1984 Automatic Spelling Correction in Scientific and Scholarly Text *Communication of the ACM* 27(4) 358-367

[5] Masaki Nagata 1996 Context-Based Spelling Correction for Japanese OCR *In Proceedings of COLING-96* 806-811

[6] 慕勇，孙才，罗振声 1995 汉语文本自动查错与确认纠错系统的研究 《计算语言学进展与应用》清华大学出版社 100-106

[7] 黄晓宏 1996 汉语文本自动查错与确认纠错系统的研究 清华大学硕士论文

[8] Luo ZhenSheng 1997 Automatically Detecting and Correcting Errors in Chinese Text *In proceedings of ICCPOL-97*