中文文本校对技术的研究与实现*)

陈笑蓉[1] 秦 进[1] 汪维家[1] 陆汝占[2]

(贵州大学计算机科学系 贵阳550025)[1] (上海交通大学计算机科学与工程系 上海200030)[2]

Research and Implementation of Chinese Text Proofreading

CHEN Xiao-Rong[1] QIN Jin[1] WANG Wei-Jia[1] LU Ru-Zhan[2]

(Department of Computer Science, Guizhou University, Guiyang 550025)[1]

(Dpartment of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200030)[2]

Abstract In this paper, we analyze the cause and types of Chinese text errors. Based on the weak relation between the error words and its adjoining words, we detect Chinese text errors using bi-gram as computational model. In this model, bi-cooccurrence probabilities, mutual information and t-test are applied. With context relation and likelihood match, the correction of Chinese text errors is implemented. The result of our experiment show the precise rate of the correction is 46.5%.

Keywords Automatic Chinese text proofreading, Bi-adjoining relation, Likelihood match, Candidate word set

文本自动校对工作的计算机化是说所有的校对工作应在计算机上自动完成;具体地说是计算机应在文字处理上增加专为校对服务的功能,主要是查(侦)错和改(纠)错。实际上计算机的文本校对工作主要是解决键盘输入方式下的文本信息准确性问题,而不是通常意义下的校改作文的工作,所以才能有针对性的功效。要研究与实现汉语文本电脑自动校对的技术,必须要先分析使用电脑输入产生文字错误的原因。

1 中文文本错误原因、类型分析

(1)词错误:输入时将词中的某字错录入为另一个字。例如:"根据在大量真实预料中的切分结果",该句中把"语料"错录为"预料"。"现代汉语语法信息词典祥解",该句中把"详"错录为"祥"。

(2)缺字错误:输入时少键入一个或几个字,例如:"每前进一步都要付出代价"错录入为"每前进步都要付出代价"。

(3)多余字错误:输入了多余的字,例如:"同学们对这个问题都有各自己的看法",录入了多余的"己"字。"上下文本中出现的词汇的概率"录入了多余的"本"字,这种情况有时表现为该字与前后是一个词组。

(4)五笔字型输入方法:这种方法结构相近的汉字其编码也相近,以及邻近键位击错或少击。如:"(ty)称为",错键为"(tu)科为"。"齿(hwb)",错键为"(hw)具"。词组误操作,如"(fva)老奸巨猾",错键为"(fba)地区"。

(5)标点符号错误:比如双引号、括号、书名号不匹配,其它的标点符号多了或缺了等错误。

(6)数错误:输入阿拉伯数字对应位置错或漏数字,例如:"2001",错输入为"2002","23980",错输入为"2980",这类纯属内容错误,是文本校对系统无法解决问题之一。

(7)输入造成的语义错误:例如:"我要去上海交通大学图书馆查资料"录入为"我要求上海交通大学图书馆查资料",该句中的"要去"错录入为"要求"。全句语法没毛病,但意思截然不同,这造成语义错误,计算机单纯地从形式上无法解决这种类型的错误。

在网上,各式各样的论坛是文字错误较多的区域。我们选择了上海交大BBS的饮水思源、体育、交大快讯、计算机系和古典名著四个讨论区中各选取了50至100篇文章,作了一个初步的统计。在250篇文章中,共有错误84处。

同音错误	同形错误	形音错误	缺字错误	多字错误	其他错误
63	4	7	2	3	5
75.00%	4.76%	8.33%	2.38%	3.57%	5.95%

在我们对使用键盘输入方法调查情况中得出,一般在专业录入人员使用五笔输入方法比较多,大专院校和部分机关,用拼音输入法输入的人占92%以上,所以同音错误出现率比较高。有的拼音输入系统是以词为单位的拼音输入法,如输入"dj",就会出现"大家"、"单价"、"动机"、"大计"、"冬季"、"代价"等后选,这无疑也增加了自动校对的难度。

2 设计思想

通过对待处理文本中常见错误类型的统计分析,我们可以了解到实际上任何一个有意义的文本都不是由一连串不相关的句子组成,相反,它的句子和段落总是会围绕某个中心事件,也就是文本具有一致性,因而我们在处理真实文本的时候技术上就应该围绕这个特点进行考虑。文本中词的集聚是一种词汇间语义关系引起的凝聚力,这些词之间用统计和规则的方法可识别其关系。一般说来,正确句子中相邻词之间应具有较强的"词语接续关系"和"词性邻接关系"。针对文中出错的字词与相邻字词接续关系弱的特征,并利用语言本身所具有的结构特征及统计特征,我们拟采用词性邻接关系检查、词语接续关系检查、词性二阶Markov预测的综合查错策略进行查错,然后把被确定包含错误的输入串变换成无错误的输出串。无错误输出串集的建立原则是:根据输入文本中的错误

*)基金项目:贵州省自然科学基金资助项目。陈笑蓉 副教授,研究方向为操作系统、自然语言理解。秦 进 在读硕士,研究方向为自然语言理解。汪维家 研究方向为自然语言理解。陆汝占 教授,博士生导师,研究方向为自然语言理解、语义web、汉语语义计算。

类型采用似然匹配算法来进行纠错，然后构成错误的字词候选字词集，对侯选字词集进行最佳序列排定。改错时以第一候选字词加以改正。

3 算法描述

3.1 词性二元邻接关系

对于句子 $S=w_1w_2\cdots w_i\cdots w_m$ 可以用 $P(S=w_1w_2\cdots w_i\cdots w_m)$ 表示 S 出现的概率，其中符号 w_i 表示第 i 个字、词，则有：

$$P(w_1w_2\cdots w_i\cdots w_m)=\prod_{i=1}^{m}P(w_i|w_1w_2\cdots w_{i-2}w_{i-1})$$

若当前字词 W_i 的出现只与前 N-1个字词有关，句子的概率改为：

$$P(w_1w_2\cdots w_i\cdots w_m)=\prod_{i=1}^{m}P(w_i|w_{i-N+1}\cdots w_{i-2}w_{i-1})$$

这就是所谓的 N-Gram 模型，也称为 N－1阶 Markov 模型。由于计算的时间和空间的局限，建立的 Markov 模型只能是低阶的，即：

N=2，我们称之为二元模型，这时

$$P(w_1w_2\cdots w_i\cdots w_m)=\prod_{i=1}^{m}P(w_i|w_{i-1})$$

当应用二阶 Markov 模型时，当前字词的出现仅能与它的前一个字词有关。

在汉语文本的句子中，一个搭配的词是任意的，可重复出现的词对，有的词性之间有密切的搭配和接续关系，有的则不然，我们试图通过这种搭配和接续信息来进行校对。设输入的句子 $S=w_1w_2\cdots w_{i-1}w_iw_{i+1}\cdots w_m$，其对应的词性为 $C=C_1C_2\cdots C_{i-1}C_iC_{i+1}\cdots C_m$，若 w_i 出错，则 w_{i-1}和 w_i 的邻接关系较弱。通过对以词为单位的词性标注的语料库，抽取词性二元邻接信息，得到以下邻接表(表1)。

表1 邻接表

C_{i+1} \ C_i / $P(C_{i+1}\|C_i)$	dm	fm	sm	……
dm	0.5329982	0.5632581	0.237651	……
fm	…			
sm	…			

表中的 dm、fm、sm 为词的属性，$P(C_{i+1}|C_i)$表示词性的邻接关系，即给定 Ci 条件下的 C_{i+1}的出现概率，表中的每一项皆为从语料库中统计得出的词性相邻出现的概率。

$$P(C_{i+1}|C_i)=\frac{N(C_iC_{i+1})}{N(C_i)}$$

当 $P(C_{i+1}|C_i)>0$时邻接关系成立。当 $P(C_{i+1}|C_i)=0$时则认为邻接关系不成立。词性的二元方法是用于发现和改正具有固定模式的错误，例如词典中专有的词和成语中的错误。例如："这件事与那个人有观(应该是有关)"。这种出错能通过此方法发现和纠正错误。

3.2 单字词同现概率统计

笔者在对错误情况的分析中注意到，有不可能同时出现的单字词相接，如："研究员"错录为"研员"，其错是因为"研"和"员"同时出现的概率极小或为零。针对单字词进行统计，这是因为，一方面，据对3755个一级汉字的统计，二元同现概率矩阵中同现过一次的二元组数目仅占3.8%；另一方面，中文文本单字词和两字词的使用占绝大多数，当两字词出现漏字或键选错误时往往出现不合法的单字词同时出现的情况。

设一个阈值 δ，判断相邻两词 w_i，w_{i+1}是否接续，就看相邻两词在语料中的同现概率 $P(w_iw_{i+1})\geqslant\delta$ 是否成立，如果成立，表示 w_i 与 w_{i+1}接续。单字词同现频率可以用极大似然法得到：

$$P(w_iw_{i+1})=N(w_iw_{i+1})/NT$$

其中 $N(w_iw_{i+1})$表示词串 w_iw_{i+1}在语料库中出现的次数，NT 表示为所统计的语料库中总词数。

对于那些经常连用而并未收入词典中的单字词来说，这种字字同现频率就相当于词频，对词频 $P(w_i)$我们可以这样来定义：$P(w_i)=N(w_i)/NT$

例如"科教兴国"，"点击"，"得很"。

3.3 词的邻接判断

设有句子 $S=w_1w_2\cdots w_iw_{i+1}\cdots w_n$，其中 $w_1w_2\cdots w_n\in D$。当相邻两个词 w_iw_{i+1}在语料中的同现概率 $p(w_i,w_{i+1})\geqslant\delta_1$时($\delta_1$为给定阈值)，可判定词 w_i 与词 w_{i+1}邻接。

这里使用信息论中的互信息(Mutual Information)概念[1]，用互信息比较二元单词对(w_i,w_{i+1})结合关系，能够反映汉字对之间结合关系的紧密程度。如果把两个相邻的词 w_i 与 w_{i+1}二元互信息即为：

$$MI(w_i,w_{i+1})=\log\{\frac{P(w_i,w_{i+1})}{P(w_i)P(w_{i+1})}\}$$

其中 $P(w_i,w_{i+1})$是 w_i,w_{i+1}的邻接同现概率，$P(w_i)$，$P(w_{i+1})$分别代表 w_i,w_{i+1}的独立概率。

若两个词 w_i,w_{i+1}的同现概率大，即有可信的邻接关系，则 $MI(w_i,w_{i+1})\geqslant\delta_2$，($\delta_2$为大于零的阈值)；若 w_i 与 w_{i+1}之间没有接续关系，则 $MI(w_i,w_{i-1})<0$；如果 w_i 与 w_{i+1}的接续关系不明确，则 $MI(w_i,w_{i+1})\approx0$，此时，必须参照一定的上下文，通过上下文词对之间的比较进一步寻找判据。

一般来说，可以将词 w 的左右 k 个词看作 w 的上下文，针对不同的应用，词的上下文取法不同(在消歧时，有人将 k 取±50)，实际上上下文范围的扩大而信息量不会有明显地增加，只会带来不必要的计算开销。而与词类关系最密切的是 w 的左右两个词。为此，我们参考文[1]引入 t-测试来作进一步的判定。

对词串 $w_{i-1}w_iw_{i+1}$定义词 w_i 相对于 w_{i-1}和 w_{i+1}的 t-测试(t-test)为：

$$t_{w_{i-1},w_{i+1}}(w_i)=\frac{p(w_{i+1}|w_i)-p(w_i|w_{i-1})}{\sqrt{\partial^2(p(w_{i+1}|w_i))+\partial^2(p(w_i|w_{i-1}))}}$$

其中，$p(w_i|w_{i-1})$、$p(w_{i+1}|w_i)$分别是 w_i 关于 w_{i-1}、w_{i+1}关于 w_i 的条件概率，$\partial^2(p(w_i|w_{i-1}))$，$\partial^2(p(w_{i+1}|w_i))$则代表各自的方差。

互信息挂靠在两个词之间，而 t-测试是直接挂靠在词上的，特引入 t-测试差将二者的挂靠对象统一起来。对词串 $w_{i-1}w_iw_{i+1}w_{i+2}$，定义词 w_i 与 w_{i+1}之间的 t-测试差为：

$$\Delta t(w_i,w_{i+1})=t_{w_{i-1},w_{i+1}}(w_i)-t_{w_i,w_{i+2}}(w_{i+1})$$

当 $\Delta t(w_i,w_{i+1})\geqslant\delta_3$时($\delta_3$为一给定阈值)，我们判定 w_i 与 w_{i+1}接续。

词的二元邻接关系反映了词 w_i 附近词的接续关系，互信息 $MI(w_i,w_{i+1})$反映了 w_i、w_{i+1}之间的结合关系的紧密程度，而 t -测试差 $\Delta t(w_i,w_{i+1})$则考虑了四个词的耦合影响，综上所述，二元邻接、互信息和 t -测试差这三种方法具有互补性，综合使用可以弥补单一词性连接计算查错的不足。

4 错误纠正

对输入编码相同或相近的错误字词，用纠错候选字、词替

换出错字词，给出纠错的建议，对漏字、多字采用似然匹配法，将与出错的字词"相似"的字词作为纠错建议[2]。

设 $S=w_1w_2\cdots w_i\cdots w_m$ 为输入文本的一个句子，$S_{er}=w'_1w'_2\cdots w'_l$ 为查出的错误字串，$O=\{O_{i1},O_{i2}\cdots O_{ik}\}$，$O$ 为 w'_i 所有候选字集，纠正工作是要从 O 集中选择一个最适合该句上下文的字串 O_n，为此对错误字串 $w'_1w'_2\cdots w'_l$ 与词 $O_{i1}O_{i2}\cdots O_{ik}$ 采用似然匹配[2]，令函数：

$$xtpp(w'_i,O_n)=\begin{cases}1 & \text{当 } w'_i=O_{si}\\ 0 & \text{当 } w'_i\neq O_{si}\end{cases}$$

若当 $l\geqslant k$ 时，有：

$\Sigma xtpp(w'_i,O_{si})\geqslant l-1 \quad 1\leqslant i\leqslant l;w'_1=O_{s1}$

或 $\Sigma xtpp(w'_{i-i+1},O_{sk-i+1})\geqslant l-1$

$1\leqslant i\leqslant l;w'_l=O_{sk}$

或 $\Sigma xtpp(w'_i,O_n)+\Sigma xtpp(w'_{l-j+1},O_{sk-i+1})\geqslant l-1$

$1\leqslant i\leqslant l;w'_l=O_{s1};w'_l=O_{sk}$

当 m<n 时，有：

$\Sigma xtpp(w'_i,O_n)+\Sigma xtpp(w'_{l-i+1},O_{sk-i+1})=k$

$2\leqslant i\leqslant k-1;w'_1=O_{s1};w'_n=O_{sm}$

那么称 $w'_1w'_2\cdots w'_l$ 与词 $O_{s1}O_{s2}\cdots O_{sl}$ 似然匹配，记为：$xtppch(w'_1w'_2\cdots w'_l,O_{s1}O_{s2}\cdots O_{sm})$ 其中 $O_{s1}O_{s2}\cdots O_{sl}$ 是作为字串 $w'_1w'_2\cdots w'_l$ 的纠错建议。

例如：$xtppch$("歇四底里","歇斯底里")，$xtppch$("迎而解","迎刃而解")。

5 具体实现

5.1 系统组成

中文文本校对系统共有四部分组成。第一部分为人机界面，它主要处理录入文档的输入及查错结果的人机交互界面；第二部分为自动分词系统，它主要是识别录入文档的词汇。这部分的关键技术在于词的歧义切分和生词的处理问题；第三部分是查错和错误定位，就是利用词性邻接关系、词语接续关系检查、词性二阶 Markov 预测的综合查错方法，这部分的关键技术在于对错误定位的准确性；第四部分是纠错处理。利用似然匹配技术和给出错字或错词的侯选字词，改错时以第一候选字词加以改正，并提供用户作修改文本的选择。

5.2 词典组织

在本系统建立了以下词典：

① 含有5万词条基本词典(JBCT)，每个词条包含三个域，即词、词性和词频。

② 分级词典，根据 JBCT 形成4个分级词典，JB1为1字词条字典、JB2为2字长的词条字典、JB34为3、4字长的词条字典，JB5为5或5以上字长的词条字典，其中排序是按字词频排序，高频词放前。主要是提高查寻速度，在校对过程中便于候选词的寻找。

③ 同音字词典(TYCD)，是为了寻找文本中的同音字键选错误而设置的。将发音相同的所有汉字组织在一起。每项包含两个域，即读音和字。

④ 形近字词典(XJCD)，该词典是为了文本中的形近字键选错误而设置的，每项包含一组形近字。

5.3 实验结果

1. 实验样例。例1　政府/n 坚决/d 房子/n腐败/n。"房子"改为"防止"。例2：桥/n 上/f 一/m 群/q 我/r遥遥摆摆/d 地/u 走/v 过/u。"我"改为"鹅"例3：这/r 部/q 电影/n 对/p 大学生/n 的/u 影响/v 飞/v 同/d 小/a 可/c。/ "飞"改为"非"。这里仅举例三个。另外我们从网上下载了新闻报道类文章30篇，散文25篇，科普文章10篇，作了一个初步的统计。在65篇文章中，共有错误185处。同音错误89个，同形错误23个，形音错误17个，缺字错误16个，多字错误29个，其它错误21个。

2. 统计结果。系统测试出的错误数为217个，其中报错正确个数为131个，对上述的错误问题的正确识别：同音错误识别出70个，同形错误识别出17个，缺字错误识别出11个，多字错误识别出21个，其他错误识别出11个。对报错给出的第一改正后的正确个数为61个。我们对该自动校对系统评判指标用文[3]来统计查全率、查准率和纠错率。错误字词的查全率为70.81%，报错准确率为60.36%，纠错率为46.56%。

结束语　中文文本的自动校对是一个难度高而又复杂的课题，国内不少研究者对其表现出极大的关注，纷纷立项进行研究。从已得到的资料来看目前大多数的中文文本校对系统在查错的研究与实现方面进行得比较多，对纠错的报道较少，可能的原因是中文分词时产生歧义字段以及专用名词、人名、生词等问题，容易产生误判，使得纠错率不高。本文基于统计的方法对查错和纠错进行了探讨，下一步系统将把规则的方法也考虑进去，加强歧义字段、生词和专用名词的识别能力，提高自动校对的性能。

参考文献

1　孙茂松，黄昌宁，邹嘉颜，陆方，沈达阳. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义. 计算机研究与发展，1997.5

2　张仰森. 中文校对系统中纠错知识库的构造及纠错建议的产生算法. 中文信息学报，2001. 3

3　张磊，周明，黄昌宁，潘海华. 中文文本自动校对. 语言文字应用，2001.2

4　Manning C D, Schütze H. FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING The MTH Press，1999

5　Kukich K. Techniques for automatically correcting words in text. ACM Computing Surveys, 1992,24(4)