中 文 信 息 学 报
第 15 卷 第 3 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 15 No. 3

基于二元接续关系检查的字词级自动查错方法*

张仰森[1,2] 丁冰青[2]
(1. 山西大学计算机科学系 太原 030006;
2. 中国科学院自动化研究所模式识别国家重点实验室 北京 100080)

摘要:本文探讨了基于字字同现、词性二元接续和语义二元接续的中文文本的自动查错原理和查错算法;给出了字词接续判断模型,并讨论了与接续判断模型相关的查错知识库的构造方法。通过对实验结果的分析和评测,证明本文所述方法是可行的。

关键词:中文文本自动校对;自动查错;二元接续关系
中图分类号:TP391.1

Automatic Errors Detecting of Chinese Texts Based on the Bi-neighborship

ZHANG Yang-sen[1,2] DING Bing-qing[2]
(1. Dept. of Computer Science, Shanxi University Taiyuan 030006; 2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences Beijing 100080)

Abstract: Automatic error detecting principle and algorithm of Chinese texts based on character-character co-occurrence, POS bi-neighborship and semantic bi-neighborship are discussed in this article. The models of judging character and word neighborship are presented, and the method of constructing error detecting knowledge bases which is related to these models is introduced. According to the analysis and estimation for experiment results, the method given in this paper is workable.

Keywords: Chinese text automatic proofread; automatic error-detecting; bi-neighborship

一、引言

中文文本校对系统的研究是一项亟待解决而又十分困难的课题[5]。由于汉语理论研究的局限性和汉语的特点,目前要达到完全的自动校对水平尚不现实,实现达到一定校对目标的人机交互式计算机辅助校对系统是我们研究的方向[3]。待校对文本是含有错误的真实文本,

* 收稿日期:2000-06-01;修改稿收到日期:2000-08-04
基金项目:山西省自然科学基金(981031)
作者张仰森,男,1962 年生,硕士,教授,研究方向为中文信息处理.

错误的来源一般有三种[3]:原稿错误、键盘录入错误或 OCR 和语音识别错误。这些错误一般破坏的只是词、短语等语言的表层结构。因此,对于键盘录入、OCR 识别、语音识别后的中文文本进行自动校对时,一般不需要采用全面的句法分析技术,利用字词间的接续关系和局部分析技术就能达到查错的目的[1,3]。本文就基于二元接续对的自动查错算法进行探讨。

二、字词接续判断模型

2.1 二元接续关系

接续关系是指有前后顺序的字词间的相邻关系[2]。二元接续关系是指在考察字串或词串 $X_1 X_2 \cdots X_{i-1} X_i X_{i+1} \cdots X_n$ 中 X_i 和相邻字词间的相邻关系时,根据语料库语言学中的二元模型理论,只须考察 X_{i-1} 和 X_i 以及 X_i 和 X_{i+1} 之间的关系即可。经过对大规模语料的分析处理,如果发现从 X_{i-1} 到 X_i 的转移概率 $P(X_i/X_{i-1})$ 满足一定的阈值限制,我们即认为 X_{i-1} 和 X_i 接续。在自动查错过程中,若要考察 X_i 是否出错,首先检查 X_{i-1} 和 X_i 是否接续,如果不接续,这时再检查 X_i 和 X_{i+1} 的接续关系(即检查从 X_i 到 X_{i+1} 的转移概率 $P(X_{i+1}/X_i)$),如果 X_i 和 X_{i+1} 也不接续,则判定的 X_i 出错。本文先对待校对文本进行字字接续检查,再对其进行词性接续和义类接续检查,从而得到词接续信息。综合考虑字接续和词接续情况,从而达到提高召回率和查准率的目的。

2.2 字字接续判断模型

假设句子 $S = Z_1 Z_2 \cdots Z_i Z_{i+1} \cdots Z_m$; Z_i、Z_{i+1} 为两个相邻的汉字。若在字容量为 N 的汉语语料库中,Z_i, Z_{i+1} 邻接同现次数为 $r(Z_i, Z_{i+1})$, Z_i, Z_{i+1} 独立出现次数分别为 $r(Z_i)$, $r(Z_{i+1})$,则定义:

$p(Z_i) = r(Z_i)/N$; $p(Z_{i+1}) = r(Z_{i+1})/N$ 分别代表 Z_i, Z_{i+1} 独立出现的概率。

$$p(Z_i, Z_{i+1}) = r(Z_i, Z_{i+1})/N \tag{1}$$

为 Z_i、Z_{i+1} 的邻接同现概率。定义 Z_i, Z_{i+1} 之间的互信息为:

$$I(Z_i, Z_{i+1}) = \log_2 \frac{P(Z_i, Z_{i+1})}{p(z_i)\, p(Z_i) * p(Z_{i+1})} \tag{2}$$

由式(1)知道,当 $r(Z_i, Z_{i+1}) = N * p(Z_i, Z_{i+1}) \geq \tau_0$($\tau_0$ 为一阈值),说明 Z_i, Z_{i+1} 的共现频率较高,我们可以判断 Z_i, Z_{i+1} 接续,但这是一个绝对指标,对那些在语料中共现频率低,但 Z_i, Z_{i+1} 分别单独出现的频率也低的接续对,可能会误判为不接续。

互信息是一个相对指标,可以解决这一问题,由式(2)可以知道,当:

①$I(Z_i, Z_{i+1}) \geq \tau_1$,$\tau_1$ 为远大于 0 的阈值,则 $p(Z_i, Z_{i+1}) \gg p(Z_i) * p(Z_{i+1})$,此时 Z_i, Z_{i+1} 具有可信的接续关系,并且 $I(Z_i, Z_{i+1})$ 值越大,接续强度越大。

②$I(Z_i, Z_{i+1}) \approx 0$,则 $p(Z_i, Z_{i+1}) \approx p(Z_i) * p(Z_{i+1})$,此时 Z_i, Z_{i+1} 之间的接续关系不明确,我们根据经典的统计理论 *Pearson* 的 χ^2 - 统计量检验 Z_i 和 Z_{i+1} 的独立性。

我们用二元组 (X, Y) 表示相邻的两个汉字,X 的取值范围是 $(Z_i, -Z_i)$,$-Z_i$ 表示取值不为 Z_i,Y 的取值范围是 $(Z_{i+1}, -Z_{i+1})$,$-Z_{i+1}$ 表示取值不为 Z_{i+1}。假设 $H_0: Z_i$ 和 Z_{i+1} 独立,从二元组 (X, Y) 中抽取字字同现频率矩阵中的 n 个非零元 (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n) 作为子样,用 n_{11} 表示取值为 (Z_i, Z_{i+1}) 的子样个数,n_{12}、n_{21}、n_{22} 分别表示取值为 $(Z_i, -Z_{i+1})$, $(-Z_i, Z_{i+1})$, $(-Z_i, -Z_{i+1})$ 的子样的个数,记 $n_{i.} = n_{i1} + n_{i2}$, $n_{.j} = n_{1j} + n_{2j}$ $(i = 1,2;\ j = 1,2)$,有 $n = n_{12} + n_{12} + n_{21} + n_{22}$。

$x^2(Z_i, Z_{i+1})$ - 统计量可定义为[5]:

$$\chi^2 = \frac{n * (n_{11} * n_{22} - n_{12} * n_{22})^2}{n_{1.} * n_{2.} * n_{.1} * n_{.2}} \tag{3}$$

由 Pearson 定理的推广可知,当 $n \to \infty$时,χ^2_n 的极限分布是自由度为 1 的 χ^2- 分布。选取显著水平 α,查 χ^2 - 分布上侧分位数表,得 $\chi^2_{\alpha,1} = \tau_2$,这样,$H_0$ 之 α- 水平否定域是 $\chi^2_n > = \tau_2$。若 $\chi^2_n < \tau_2$,不能否定 H_0,说明假设成立,Z_i 和 Z_{i+1}独立,即 Z_i 和 Z_{i+1}不接续;否则假设不成立,Z_i 和 Z_{i+1}不独立,即 Z_i 和 Z_{i+1}接续。

③$I(Z_i, Z_{i+1}) \ll 0$,则 $p(Z_i, Z_{i+1}) \ll p(Z_i) * p(Z_{i+1})$,此时 Z_i, Z_{i+1}之间基本没有接续关系,并且 $I(Z_I, Z_{i+1})$值越小,接续强度越弱。

由于在进行汉字间的互信息统计时,所用语料库规模有限,因而所得到的字字接续矩阵不能完全充分地反映字字之间的接续关系,为此,结合二元语法 $F(Z_{i+1}/Z_i)$ 和一元语法 $F(Z_{i+1})$,用退步法(BackOff Approach)计算从 Z_i 到 Z_{i+1}的转移概率:

$$P(Z_{i+1}/Z_i) = \lambda_1 F(Z_{i+1}/Z_i) + (1 - \lambda_1) F(Z_{i+1}) \tag{4}$$

$$F(Z_{i+1}/Z_i) = r(Z_i, Z_{i+1}) / r(Z_{i+1})$$

$$F(Z_{i+1}) = r(Z_{i+1}) / N$$

这里,λ_1 为经验参数,经试验,一般选取 $\lambda_1 = 0.8$。

当 $P(Z_{i+1}/Z_i) \geq \tau_3$ 时(τ_3 为一阈值),Z_i, Z_{i+1}具有可信的结合关系,且 $P(Z_{i+1}/Z_i)$值越大,接续强度越大。

以字字同现概率式(1)为主,结合考虑式(2)、(3)、(4),定义字字接续函数 $ZZJX(Z_i, Z_{i+1})$为:

$$ZZJX(Z_i, Z_{i+1}) = \begin{cases} 1 \text{ 当 } r(Z_i, Z_{i+1}) \geq \tau_0 .\text{OR. } I(Z_i, Z_{i+1}) \geq \tau_1 .\text{OR. } \chi^2(Z_i, Z_{i+1}) \geq \tau_2 \text{ OR. } P(Z_{i+1}/Z_i) \geq \tau_3 \\ 0 \text{ 其它} \end{cases}$$

当 $ZZJX(Z_i, Z_{i+1}) = 1$ 时,我们认为汉字 Z_i 和 Z_{i+1}接续,当 $ZZJX(Z_i, Z_{i+1}) = 0$ 时,我们认为汉字 Z_i 和 Z_{i+1}不接续。τ_0、τ_1、τ_2、τ_3 为相关阈值。

2.3 词接续判断模型

设句子 $S = W_1 W_2 \cdots W_{i-1} W_i W_{i+1} \cdots W_m$,其中 $W_i (1 \leq i \leq m)$ 为第 i 个词,$T_i[X]$为词 W_i 的可能标记,$Num(T_i[X])$为该词可能标记的个数,设 $T_i[a]$为该词的第 a 个标记,则 $T_i[a] \in T_i[X]$,$1 \leq a \leq Num(T_i[X])$;$T_{i+1}[X]$为词 W_{i+1}的可能标记,$Num(T_{i+1}[X])$为该词可能标记的个数,设 $T_{i+1}[b]$为该词的第 b 个标记,则 $T_{i+1}[b] \in T_{i+1}[X]$,$1 \leq b \leq Num(T_{i+1}[X])$。我们定义 $P(T_{i+1}[b]/T_i[a])$为从 $T_i[a]$到 $T_{i+1}[b]$的转移概率:

$$P(T_{i+1}[b]/T_i[a]) = R(T_i[a], T_{i+1}[b]) / (R(T_i[a]) * R(T_{i+1}[b]))$$

其中 $R(T_i[a], T_{i+1}[b])$为 $T_i[a]$与 $T_{i+1}[b]$二元词性同现次数,$R(T_i[a])$和 $R(T_{i+1}[b])$分别为词性 $T_i[a]$和 $T_{i+1}[b]$在统计语料中出现的次数。

当 $P(T_{i+1}[b]/T_i[a]) > \rho_{Tag}$时,我们认为标记 $T_{i+1}[b]$和 $T_i[a]$接续,设 $Tag = 1$、2、3 分别表示标记取词性、义类大类、义类中类,$CXJxSet[i]$、$DYLJxSet[i]$、$ZYLJxSet[i]$分别为 $T_{i+1}[X]$中所有满足 $P(T_{i+1}[b]/T_i[a]) > \rho_{Tag}$的标记组成的集合,$Num(CXJxSet[i])$、$Num(DYLJxSet[i])$、$Num(ZYLJxSet[i])$分别为相应集合中元素的个数。则定义:

1. 词性接续函数 $CXJX(W_i, W_{i+1})$

$$CXJX(W_i, W_{i+1}) = \begin{cases} 1 & \text{若 } Num(CXJxSet[i]) > 0 \\ 0 & \text{若 } Num(CXJxSet[i]) = 0 \end{cases}$$

2. 义类接续函数

记义类大类接续函数为 $DYLJX(W_i, W_{i+1})$,义类中类接续函数为 $ZYLJX(W_i, W_{i+1})$:

$$DYLJX(W_i, W_{i+1}) = \begin{cases} 1 & \text{若 } Num(DYLJxSet[i]) > 0 \\ 0 & \text{若 } Num(DYLJxSet[i]) = 0 \end{cases}$$

$$ZYLJX(W_i, W_{i+1}) = \begin{cases} 1 & \text{若 } Num(ZYLJxSet[i]) > 0 \\ 0 & \text{若 } Num(ZYLJxSet[i]) = 0 \end{cases}$$

当 $CXJX(W_i, W_{i+1}) = 1$ 时,我们认为词 W_i 和 W_{i+1} 的词性接续,当 $DYLJX(W_i, W_{i+1}) = 1$ 时,我们认为词 W_i 和 W_{i+1} 的义类大类接续,当 $ZYLJX(W_i, W_{i+1}) = 1$ 时,我们认为词 W_i 和 W_{i+1} 的义类中类接续。

3. 词接续判断函数的求取

根据词性接续、义类大类、中类接续的情况,以举手投票法确定 W_i,W_{i+1} 是否接续:

```
FUNC CIJX(Wi,Wi+1);
BEGIN
Agree = 0;
IF CXJX(Wi,Wi+1) = 1 THEN Agree = Agree + 1;
IF DYLJX(Wi,Wi+1) = 1 THEN Agree = Agree + 1;
IF ZYLJX(Wi,Wi+1) = 1 THEN Agree = Agree + 1;
IF Agree > = 2 THEN return 1;
ELSE return 0;
END
```

当 $CIJX(W_i, W_{i+1}) = 1$ 时,认为 W_i 和 W_{i+1} 接续,当 $CIJX(W_i, W_{i+1}) = 0$ 时,W_i 和 W_{i+1} 不接续。

三、自动查错知识库的构造

由上面的叙述可以知道,要想判断字字接续、词性接续以及义类接续,必须知道字字同现频率、单字频率、词性同现频率以及义类同现频率,为此,我们利用语料库语言学的基本思想,研究知识库的获取技术。通过对大规模语料的统计分析,构造字频向量、字字同现频率表、二元词性同现频率表以及二元义类同现频率表。

3.1 字字同现频率表的构造

1. 语料选取

语料库的选取应具有典型性、广泛性。我们直接对未经加工过的生语料进行统计,语料来源情况如下:

①1990,1994 年《人民日报》部分已分词、词性标注语料(917 篇,93 万字);②《读者》光盘语料(1102 篇,164 万字);③《计算机世界报》语料(343 万字)。共 600 万语料。

2. 字字同现频率表(ZZTX)的数据结构

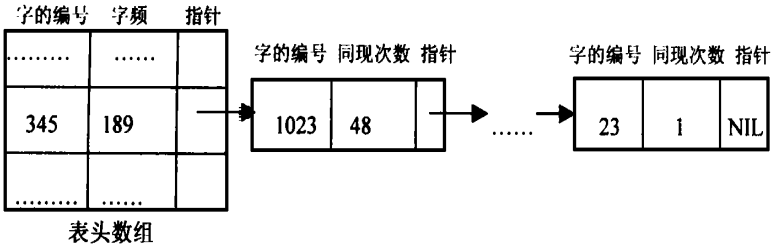
在实际文本中,除了国标一、二级汉字外,还有英文、数字和标点符号,为了便于统计,我们采用 6768 个标记集 Σ,它由以下几部分构成:

①6763 个一、二级汉字(记为 C)各为一个标记;②阿拉伯数字 N(0 - 9);

③英文单词 W(a - z,A - Z);　　④句首标记 H;

⑤句尾标点 T(。!?:;,、-);　　⑥其它 O。

由于汉字标记集 Σ 中含有 6768 个标记,所有可能出现的字二元对的个数为 4.5805824×10^7。若按矩阵构造字字同现频率表,其规模令人难以接受。其实,字的二元同现是很稀疏的,我们采用链表形式压缩存储字同现表:首先,给每个标记集中的每个标记赋一个唯一的编号,用 $Num(Z_i)$ 表示字 Z_i 对应编号;然后,按下列形式组织字同现频率表 ZZTX:



表头数组中存放着所有字(标记)的编号及其单独出现的次数,各数组元素链出的单链表中存放着所有与该元素同现的词的标记及其二元同现次数。

3. 二元字字同现频率表 ZZTX 的生成算法

算法 1　字二元同现频率表 ZZTX 的生成算法

Ⅰ. 初始化字二元同现频率表 ZZTX:生成表头数组,按从小到大的顺序放入所有字的编号,将字频全置为零,指针都置空。

Ⅱ. 若语料库目录中尚有未读文件,则打开其中一个文件,做 Ⅲ;否则,算法结束。

Ⅲ. 若当前文件未空,则从中划分出一句(以。!?:;- - 结束) $S = Z_1Z_2\cdots Z_i\cdots Z_m$,其中 $Z_i \in \Sigma$,接着做 Ⅳ.;否则,关闭该文件并转 Ⅱ.。

Ⅳ. i 从 1 到 m 做 i.、ii.:

i. 在表头数组中寻找 Z_i 对应的元素 k,置 ZZTX[k].CP = ZZTX[k].CP + 1;

ii. 从 ZZTX[k].NEXT 出发查找与 Z_{i+1} 对应的表结点,找到,则令同现次数加 1;

否则,链入一个新结点,往其中置入 Z_{i+1} 的编号,并将同现次数赋值为 1。

Ⅴ. 转 Ⅲ。

通过对 600 万大规模真实文本的统计,得到了二元字字同现频率表和单字字频表。

3.2　二元词性同现频率表和二元义类同现频率表的构造

我们采用的词性标记集共含 14 类标记,列表如下:

词性	标记	词性	标记	词性	标记	词性	标记
形容词	A	成　语	I	介　词	P	动　词	V
连　词	C	数　词	M	量　词	Q	语气词	Y
副　词	D	名　词	N	代　词	R		
叹　词	E	象声词	O	助　词	U		

通过对 1990,1994 年部分《人民日报》已分词、词性标注的熟语料进行统计,得到相邻词性标记之间的二元同现频率,生成一个二元词性标记同现频率表,同时生成词性同现三元组。按《同义词词林》体系,词义空间分为 12 个大类和 94 个中类,通过对 5 万经人工义类标注(排歧[6])的文本进行统计,得到相邻义类标记之间的二元同现频率,生成二元义类大类标记同现频率表和二元义类中类标记同现频率表。

四、基于字词二元接续关系检查的综合自动查错算法

字词综合查错算法,将字字同现和词接续综合考虑,把字接续检查与词接续检查进行叠加,发现待校对文本中尽可能多的错误,提高查错系统的召回率。基于字字同现和词接续的字词级综合查错算法如下。

算法 2　字词级综合查错算法

输入:句子 $S = Z_1 Z_2 \cdots Z_i Z_{i+1} \cdots Z_m$;经分词后为 $S = W_1 W_2 \cdots W_j W_{j+1} \cdots W_n$;

输出:句子 $R = Z_1 Z_2 \cdots \# Z_j \cdots Z_{j+k} \# \cdots Z_m$;其中 $Z_j \cdots Z_{j+k}$ 为被判定为不接续的单字或单字串。

设 Ticket1[k]为字字接续判断向量,Ticket2[l]为词接续判断向量,Strlen(W_j)为 W_j 的词长。算法描述如下:

/ *字、词接续向量的获取 */

1. 初始化:Ticket1[0..m] = 0;Ticket2[0..n] = 0

2. 循环,i 从 1 到 m - 1,做 3

3. 如果 ZZJX(Z_i,Z_{i+1}) = 1 则 Ticket1[i] = 1;否则 Ticket1[i] = 0

4. 循环,i 从 1 到 n - 1,做 5

5. 如果 CIJX(W_i,W_{i+1}) = 1 则 Ticket2[i] = 1;否则 Ticket2[i] = 0

/ *词接续向量的变换 */

/ *这里,len 为待处理词尾字在句中的位置,flag 为待处理词中的字与其前后字是否接续的标志,flag = 0 表示待处理词中所有字与其相邻字接续,flag = 1 表示待处理词中至少有一字与其相邻字不接续 */

6. i = 1;len = 0;

7. 如果 i < = m,做 8,9;否则转 12。

8. flag = 0;j = len;len = len + strlen(W_i);

9. 如果 flag = 0 .and. j < len,做 10,否则做 11;

10. 如果 Ticket1[j - 1] = 0 且 Ticket1[j] = 0,则 flag = 1,转 9;否则 j = j + 1,转 9;

11. 如果 flag = 1,则 Ticket2[i - 1] = 0,Ticket2[i] = 0;否则 i = i + 1,转 7;

/ *在不接续的字词前后加“#”标记 */

12. j = 1;

13. 如果 j < = n,做 14,否则结束。

14. 如果(Ticket2[j - 1] = 0 .AND. Ticket2[j] = 0),则做 15,16,否则做 19;

15. R = R +“#”;

16. 如果(j < = n. AND. Ticket2[j - 1] = 0. AND. Ticket2[j] = 0)则做 17,否则做 18;

17. R = R + W_j;j = j + 1;转 16;

18. R = R +“#”;转 13;

19. R = R + W_j;j = j + 1;转 13;

五、实验结果及查错实例

5.1 实验结果

我们采用 Visual C + + 语言在 Windows 环境下实现了一个基于二元接续关系检查的字词级综合查错实验系统,对 100 篇录入但未经校对的真实错误文本(还有 2041 个测试点)进行查错处理,测试结果如表 1 表示。

表 1　不同查错模型测试结果

实验编号	实验内容	召回率	查准率
实验 1	标记分词后的单字(串)为出错点	70.8%	19.6%

实验 2	基于字字二元接续关系检查	62.4 %	35.8 %
实验 3	基于词性、义类二元接续检查	56.4 %	33.3 %
实验 4	基于字词二元接续关系综合查错	71.2 %	35.1 %

其中召回率为错误被发现的比例,查准率为查出的错误为真正错误的比例。

5.2 查错实例(带下划线的值低于接续阈值)

经试验,我们选定

$\tau_0 = 100$; $\tau_1 = 2.3$; $\tau_2 = 0.02$; $\tau_3 = 10.828$; $\rho_1 = 0.000005$; $\rho_2 = 0.00002$; $\rho_3 = 0.00003$

例 1 我们认为可疑延长时间。

字串: 我 们 认 为 可 疑 延 长 时 间

$r(Z_i, Z_{i+1})$ 6583 326 3408 82 15 0 74 127 2355

$I(Z_i, Z_{i+1})$ 6.8014 4.1921 6.9713 -0.2086 2.7700 0.0 6.4397 1.8607 6.5841

$P(Z_{i+1}/Z_i)$ 0.2692 0.0523 0.0923 0.0043 0.0173 0.00001 0.0063 0.0059 0.2732

$\chi^2(Z_{i+1}/Z_i)$ 960605.5 79043.9 677156.4 27181.6 268033. 0.0 386112.0 448506.7 599839.9

*Ticket*1[*i*]0 1 1 1 1 1 0 1 1 1 0

分词: 我们 认为 可疑 延长 时间

词性: *R* *V* *A* *V* *N*

0.000005 0.000006 0.000005 0.000006

义类大类: *A* *G* *E* *I* *C*

0.000187 0.0000106 0.0000134 0.00094

义类中类: *Aa* *Gb* *Ed* *Ih* *Ca*

0.0000464 0.0000253 0.0000082 0.0000345

*Ticket*2[*i*] 0 1 0 0 1 0

变换后 0 1 0 0 1 0

输出 *R*: 我们认为#可疑#延长时间。

在本例中,通过字字接续判断,发现"疑"和"延"不接续,但不好确定对那一个字标错,而通过词接续判断,发现"可疑"与其前后相临的词接续关系较弱,故将其标出。

例 2 变贫空落后为文明富强。

字串: 变 贫 空 落 后 为 文 明 富 强

$r(Z_i, Z_{i+1})$ 4 0 0 229 20 31 256 0 12

$I(Z_i, Z_{i+1})$ 4.59 0.0 0.0 6.03 -1.78 -0.56 4.89 0.0 2.88

$P(Z_{i+1}/Z_i)$ 0.13 0.0 0.0 0.02 0.001 0.004 0.04 0.0 0.002

$\chi^2(Z_{i+1}/Z_i)$ 819178.7 0.00002 0.0003 8302.0 5067.18 1418763.0 1418762.9 0.000002 931020.7

*Ticket*1[*i*] 0 1 0 0 1 1 1 1 0 1 0

分词: 变 贫 空 落后 为 文明 富强

*Ticket*2[*i*] 0 1 1 1 1 1 1 0

变换后 0 1 0 0 1 1 1 0

输出 *R*: 变贫#空#落后为文明富强。

在本例中,分词后虽然通过词接续判断,认为"贫"、"空"和"落后"是接续的,而通过字字接续判断,发现"贫"和"空"、"空"和"落"都是不接续的,故将"空"标为出错处。

通过以上例子可以看出,基于字词接续的综合查错算法对提高查错系统的召回率和查准率确实有效。

综上所述,在目前对正确的大规模真实文本的处理尚不完善的条件下,利用二元接续关系检查进行局部字词级查错处理是有效的。将基于字和词接续查错模型结合起来,互相补充,可查出文本中潜在的尽可能多的错误,使查错召回率有进一步提高。

参 考 文 献

[1] 慕勇,孙才,罗振声. 汉语文本自动查错与确认纠错系统的研究. 见:计算语言学进展与应用,北京:清华大学出版社,1995

[2] 邱超捷,宋柔等. 大规模语料库中词语接续对的统计与分析. 见:语言工程. 北京:清华大学出版社,1997

[3] 张仰森,丁冰青. 中文文本自动校对技术现状及展望. 中文信息学报,1998,12(3)

[4] 于勐,姚天顺. 一种混合的中文文本校对方法. 中文信息学报,1998,12(2)

[5] 刘开瑛. 中文文本自动分词和标注. 北京:商务印书馆,2000 年

[6] 张永奎等,基于义类组合信息的义类排歧方法研究. 情报学报,1996. 12,增刊

[7] James L Peterson. Computer Programs for Detecting and Correcting Spelling Errors. Communication of the ACM 80. 12

IWPT2001

7th International Workshop on Parsing Technologies

http://www.icl.pu.deu.cn/iwpt2001**.html**

Call for Paper

The Institute of Computational Linguistics, Peking University, Beijing, China, will host the 7th International Workshop on Parsing Techonlogies (IWPT2001) form 17 to 19 October, 2001.

Topics of interest for IWPT2001 include, but are not limited to: theortical and practical studies of parsing algorithms for natural language sentences, texts, fragments, dialogues, ill-formed sentences, speech input, multi-dimensional (pictorial) languages, multimedia (web) documents, and parsing issues arising or viewed in a multimodal context. Both grammar-based and statistical approaches are welcome.

Prospecrive authors are invited to send submissions to the IWPT2001 programme chairman Giorgio Satta. Papers must be in the format given at the IWPT2001 home pages. Full papers should not exceed 12 pages; short papers should not exceed 1500 words. Submission is electronically, in postscript form. Only in case electronic submission is impossible, four send papers to:

iwpt2001 @dei. unipd. it

Giorgio Satta (IWPT2001programme chair)

Universita di Padova

Dipartimento di Elettronica e Informatica

via Gradenigo6/ A

I - 35131 Padova, Italy

Time schedule: Deadline for paper submission: June 5

Notification of acceptance: July 18

Final papers due: August 27