中文校对系统中的修改建议提供算法

郭志立 裘照明

IBM 中国研究中心 北京市海淀区上地六街 26 号 100085

摘 要： 本文介绍了对中文校对系统所侦测出来的错误字串提供修改建议的算法。这个算法已应用在 IBM 中国研究中心研究的中文校对系统 CEC 中。它首先根据字形、字音、字义或输入编码相近的原则整理出一个替换字表，然后结合主词典和二元语法的统计模型，通过加字或换字对侦测出来的错误字串提供若干修改建议并予以评价。这个算法能对 80%以上的错误字串提出正确的修改建议，并已经集成在 Lotus WordPro'97 这个字处理软件之中。

关键词： 中文校对 修改建议 替换字表

The Candidate Suggestion Algorithm in a Practical Chinese Error Check System

GUO Zhili, QIU Zhaoming

IBM China Research Lab

E-mail: guozhili@cn.ibm.com

ABSTRACT: This paper introduces an algorithm to offer correct candidates for the detected error strings in CEC, a practical Chinese Error Checker developed in IBM China Research Lab. Using a Hanzi substitution table which is previously collected based on the similarity of shape, pronunciation, meaning and/or input coding, a main dictionary and a bigram statistics model, it offers reasonable candidates for the error strings by adding or replacing one of its Hanzi. This algorithm is able to offer correct alternatives for 80% of the detected error strings and has been integrated into a commercial Chinese error check system.

KEYWORDS: Chinese Error Check (CEC), Candidate Offering, Chinese Substitution Table

1. 概论

在中文校对系统中，为侦测出的错误字串提供正确的修改建议，是校对系统的一个重要组成部分。侦错是为了检测出文本中的错误，修改建议则辅助用户订正错误。一个好的修改建议提供算法可以为用户修改文本提供方便，而且从一个侧面把造成错误字串的原因提示给用户，极大地改进了整个校对系统的用户友好度。

文献[1] 对中文校对做了深入浅出的探讨，但重点放在侦错过程上，对修改建议的提供

仅提到了 template 的概念。文献[2]则提出了一种利用语言模型评分来同时进行侦错和订正的算法。我们根据中文文本中的错误类型分布，在研究错误侦测技术的同时较深入地研究了提供修改建议的算法。计算机中文文本中的错误大致可以分为以下四类：

(1) 音同或音近的字：

例 1：“知识份子”中的“份”[应该是“分”]

例 2：“看风使驼”中的“驼”[应该是“舵”]

(2) 字形相近的字：

例 3：“作学问要经过三种境界”中的“要”[应该是“要”]

例 4：“军事干预”中的“于”[应该是“干”]

(3) 字义相近的字：

例 5：“故技重演”中的“技”[应该是“伎”]

例 6：“莫明其妙”中的“明”[应该是“名”]

(4) 输入编码容易混淆或键位相邻的字：

例 7：“她法部长后”中的“法”[应该是“当”]

例 8：“签团了自己的名字”中的“团”[应该是“署”]

(5) 输入过程中漏字造成的错：

例 9：“服器发生故障”中的“服器”[应该是“服务器”]

根据以上分析，对每一个汉字搜集整理其常见易混淆字，形成该字的替换字集。在 CEC 系统中这个集合叫做替换字表。

在错误字串中加字，或者用替换字表替换错误字串中的字，结合主词典和二元语法的统计模型判定所得到的新串是否具备一定的合理性，包括新串是否成为词条，新串的一部分是否成为词条，或组成新串的汉字之间是否有较高的共现频率。根据合理性高低予以评价和排序后放入到修改建议表中。这个算法已经应用到了 IBM 研究中心的中文校对系统 CEC 中，它能对 80%以上的错误字串提出正确的修改建议。

为了提高用户友好度，CEC 还提供了汉语易犯错误表和用户词典：

- 在易犯错误表中，收入了容易混淆而且在日常写作中出现较多的错误，如“按部就班”常常被写为“按步就班”，“莫名其妙”常常被误写为“莫明其妙”等；
- 在用户词典中，用户除了可以定义新词[用于减少 CEC 中的误报]外，还可以把错误形态及相应的正确写法定义到词典中，如把“丁青胶#丁腈胶”定义在词典中可以帮助 CEC 侦测出错误的写法“丁青胶”，并在修改建议中提示出正确的写法“丁腈胶”；而词条“郭志力#郭志立”可以避免在文本中把人名写错。

在这两种数据结构种定义的正确写法都会被修改建议提供算法取出，并优先放在修改建议列表之中。

2. 中文替换字表和语言模型

2.1 近似字的选取和中文替换字表的构造

首先按照上面所提到的音同或音近、形近、义近、编码或键位相近等汉字近似原则，分别构造近似字表：

- 音近表中收入所有与原字拼音相同或近似的字。这个表从操作系统中的拼音输入法中获得。如

原字	*近似字*
大	答达打搭瘩塔笪耷哒褡疸怛靼妲沓嗒鞑
耍	要唰

- 输入编码相近或键位相邻的字。上表中已经涵盖了目前简体汉字常用的拼音输入法中编码近似的字，所以本表主要针对另一种常用的输入法，即五笔字型输入法来整理。如

原字	*近似字*
大	百帮成春磁夺丰顾胡灰碱克磊厉历龙面三砂硕肆太套厅厦厌友右原
耍	要布厕碘碉而矾耐碳砚页矶

- 形近表和义近表由从事文字工作的专家搜集而来，如

原字	*近似字*
大	太人
耍	要

把这些表综合起来，然后根据单字使用频率以及与原字的近似程度把替换表中的字进行排序，作为系统所使用的汉字替换表。

在实际应用中由于内存和操作界面的限制，把这些表简单地综合排序所得到的表往往规模太大，所以还需要进一步舍弃某些汉字。我们现在采用简单的字频筛选法，也就是把使用频率低的字舍弃。通过实验发现这样做既可达到内存的要求，又能达到一定的精度。

2.2 带有词频信息的词典和二元语法统计模型

CEC 系统的侦错和修改建议采用了同一部含有词频信息的词典。词频信息从一个较大规模的语料库[102M，约五千万字]中统计而来，因而较真实地反映了汉语词条的使用情况。词频信息在建议项的判定、评价和排序中起着重要的作用。

在 CEC 的侦错过程是以词为单位进行的，所以在对语料进行字的二元语法统计时，首先对语料进行切分，之后统计单字切分单元与单字切分单元的共现频率，作为单字二元语法矩阵（文献[3]称这种统计为 inter-word character bigram）。

3. 建议算法

修改建议模块从应用程序中接收到包含错误字串的句子以及错误字串在句中的起始和终止位置，在清空建议缓冲区、把建议计数器初始化为 0 之后，便把错误字串从句子中取出，判定其类别。

对非汉字或不全是由汉字构成的错误字串，修改建议模块将予以特殊处理；如果错误字串仅仅包含单个汉字，目前的 CEC 只是简单地把该字替换表中的前 20 个放入建议缓冲区中，而没有结合上下文进行评价或调整；下面着重介绍 CEC 修改建议算法对由多个汉字构成的错误字串的处理算法。

算法框图在下页的图 1 。

在这个算法中，第 2 和 3 两个步骤为引用用户词典，这样用户保证自己经常使用的专业词汇或重点词汇不被误写。例如算法可以根据在用户词典中定义的“丁青胶#丁腈胶”和“郭志力#郭志立”等条目为错误字串“丁青胶”和“郭志力”提供出正确的修改建议“丁腈胶”和“郭志立”。

第 4 和 7 两个步骤检查错误字串是否出现在易犯错误表中。

第 5 步检查原错误字串是否由于在词条中漏字所致。若是，则把原词条作为修改建议；

第 8-14 步是对原错误字串进行替换。在实际对三个或更多字构成的错误字串的字进行替换时，放宽了替换用字的范围[第 8 和 10 两个步骤]，其他步骤均采用替换字表中的候选字。在判定替换后的新串是否可以作为修改建议时，除了新串是词典中的词条[第 8, 9, 10, 11, 13, 14 步]之外，如果新串的一部分成词[第 9 步骤，四字或四字以上的串] 或相应字在二元语法模型中的共现频率较高[第 12, 13, 14 步]，新串也可以作为修改建议。

4. 测试实例和实验结果

下面是针对不同错误类型的一些测试用例：

- 由于音近或形近而造成的错误[主要由拼音输入法造成]：句子中带双横线的斜体字是 CEC 侦测出的错误，句尾圆括号内的斜体字是提供的正确的修改建议。

  大陆统一是全体~~*港奥*~~同胞的心声（*港澳*）。

  他已厌倦了这种~~*按步就班*~~的生活方式（*按部就班*）。

  作学问~~*要*~~经过三种境界。学好外语也不能例外（*要*）。

  由于编码方案的设计~~*伤*~~缺乏基础理论的支持（*尚*）。

  在创造新词的群体中，以城市青少年~~*群题*~~创造的流行（*群体*）。

  计算机系的~~*程序射击*~~课程很受欢迎（*程序设计*）。

  ~~*知识份子*~~对社会应该有责任感（*知识分子*）。

- 由于形近或五笔字型编码的接近而造成的错误[主要由五笔字型输入法造成]：

  美国对伊拉克进行军事~~*手预*~~（*干预*）。

  近代学者王国维先生说，作学问~~*耍*~~经过三种境界（*要*）。

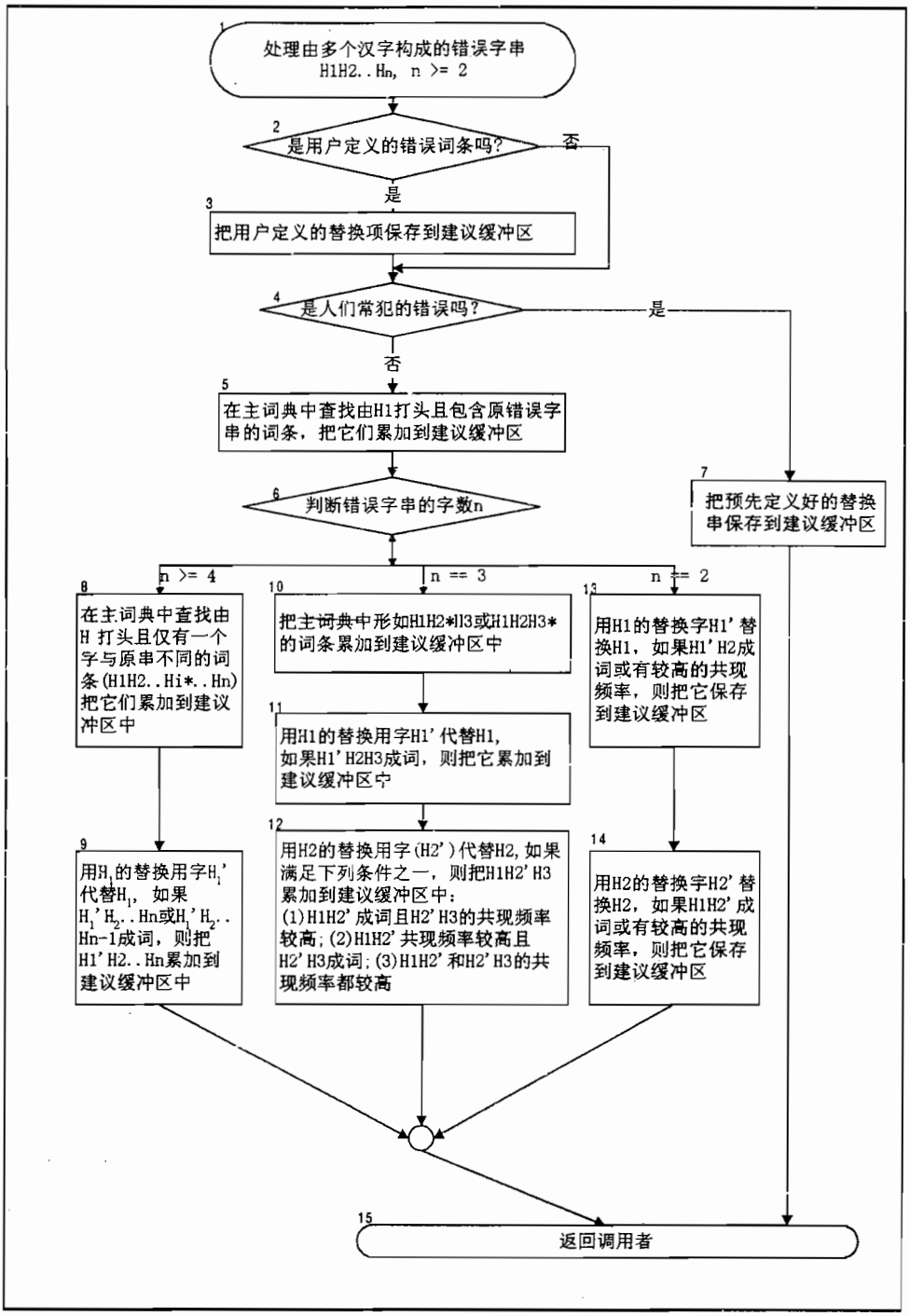

图 1　对多字中文字串提供修改建议的算法框图

~~*锡*~~我们力量，他说（*赐*）。

他父亲的身体~~*每况俞下*~~（*每况愈下*）。

技术是在人类的物质活动中~~*实旋*~~的一种方案（*实施*）。
美国人~~*伏着*~~财大气粗，开场白就来者不善（*仗着*）。
力争最大限度~~*的勇融*~~（*的通融*）。
我们要~~*高质量的完成*~~建设任务（*高质量地完成*）。

- 由于形近或五笔字型编码的接近而造成的错误[主要由五笔字型输入法造成]：

美国对伊拉克进行军事~~*于预*~~（*干预*）。
近代学者王国维先生说，作学问~~*要*~~经过三种境界（*要*）。
~~*场*~~我们力量，他说（*赐*）。
他父亲的身体~~*每况俞下*~~（*每况愈下*）。
技术是在人类的物质活动中~~*实旋*~~的一种方案（*实施*）。
美国人~~*伏着*~~财大气粗，开场白就来者不善（*仗着*）。
力争最大限度~~*的勇融*~~（*的通融*）。
我们要~~*高质量的完成*~~建设任务（*高质量地完成*）。

对一个含有 100 个错误字串的文本的测试结果如下[在评价正确的修改建议时，只计入原串的正确写法出现在修改建议列表中的前五个之中的情形]：

	产品一	CEC	产品二
检出错误	83	95	94
提供正确的修改建议	66	82	61

对另一个含有 235 个错误字串的测试文本的测试结果如下：

	产品一	CEC	产品二
检出错误	172	189	164
提供正确的修改建议	81	131	57

CEC 的修改建议提供算法比另外的中文校对产品的效果都要好。

5. 结论

利用较大规模的近似字集作为替换字表，结合主词典和语言统计模型，通过加字和换字的算法可以为侦测出来的大多数错误字串提出正确的修改建议。把它应用到字处理或编辑排版系统的中文校对模块之中，可以获得良好的效果。目前该算法已经作为中文校对模块的一部分集成到了 Lotus Word Pro'97 之中。

目前的算法尚不能对下面几种错误类型提供出正确的修改建议：

(1) 由于词语中漏字而遗留下来的错误单字：
*唉*叹气的生活*[唉声]*。
*膊*上阵是他的绝招*[赤膊]*。

对于这件事情，他感到*惭*万分*[惭愧]*。
我喜欢大自然，尤其是*澈*透明的泉水*[清澈]*。
今天是他的*诞*纪念日*[诞辰]*。
采用*恫*的手段是十分卑鄙的，我们应该光明正大*[恫吓]*。

(2) 词语搭配错误[一般是由智能输入法，如智能 ABC，导致的输入错误]：
无法理解这种*抽象型式[抽象形式]*。
明天将举行反对*充足歧视*的大游行*[种族歧视]*。
王老师认为必要时*可疑延长*考试时间*[可以延长]*。
总经理认为*必须制度*一整套管理制度*[必须制定]*。
这项工作*本身煤油*意义*[本身没有]*。
*宝剑食品*是一项新兴的产业*[保健食品]*。
她的*政治地委*上升很快，官越做越大*[政治地位]*。
中国*改革开方*的步伐很快，世界经济与科技也在飞速发展*[改革开放]*。

(3) 日期表达错误
在*1989年9月9*这一天，我上了大学*[1989年9月9日]*。
一九九三年十日，我去学校报到*[一九九三年十月十日]*。
一九二一年二月四五日，是一个令人难忘的日子*[一九二一年二月四日]*。
一九二一年二月四十五日，是一个令人难忘的日子*[一九二一年二月四日]*。

第(1)种情形，是由于目前的算法对单字提供修改建议的能力太弱所致。处理办法可以通过在词典中查找含有该字的二字词，考察这些词与句中相邻词的共现频率，取合适者作为修改建议，这需要预先进行词的二元语法统计；

第(2)种情形，也需要词的搭配知识[如词的二元语法]才能很好解决，这要求对语言知识资源进行搜集整理并用有效的数据结构在计算机中加以实现。这在目前受到语料库规模较小和机器内存不足的限制；

第(3)种情形，可以根据侦错原理把可能的错误原因提示给用户，而提供正确的修改建议必须首先使系统具备更高层次的语言理解能力。

[参考文献]

[1] GUO Jin, Automatic Chinese spelling checking, COLING'94 Tutorial Notes, 1994
[2] CHANG C.H., A pilot study on automatic Chinese spelling error correction, Communications of COLIPS, VOL. 4, No.2, 1994
[3] Lee L. S. et al. Golden Mandarin (II) - an improved single-chip real-time Mandarin dictation machine for Chinese language with very large vocabulary, Proceedings. Of ICASSP-93, 1993