中文自动查错与人机交互纠错系统的研究与实现－简介语科中文自动校对系统

吴岩

哈尔滨工业大学计算机科学与技术系 哈尔滨 150006

E-mail: vanwu@langcomp.com.hk

蔺荪

香港城市大学中文、翻译及语言学系 香港九龙达之路 83 号

E-mail: ctslun@cityu.edu.hk

摘　要：本文介绍了一个针对任意中文文本的自动查错与人机交互纠错系统－语科中文自动校对系统，查错原理是基于汉语语言的语法语义分析，同时结合统计方法对文本的错误进行确认，然后根据原文使用的输入法给出候选词组。本系统可嵌入到 WIN WORD 中作为语言工具使用。

关键词：中文自动查错 中文自动校对 自动分词 语法语义分析

Research and Implementation on Automatic Checking and Man-machine Interactive Correction for Chinese Electronic Files

Dr. WU Yan

Dept. of Computer, Harbin Institute of Technology, Harbin 150006

E-mail: vanwu@langcomp.com.hk

Caesar LUN

Dept. of CTL. City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

E-mail: ctslun@cityu.edu.hk

Abstract: In this paper. LangComp Chinese Automatic Correction System, a hybrid automatic checking and correction system for the Chinese language is presented. At the beginning, a checking process that is based on multi-level analyses of the Chinese language combined with the statistical method is used to find error positions and possible candidate words. Then, the man-machine interactive method is applied to correct the error parts of the input file. The system can be embedded into MS WORD as a language aid.

Keywords: Automatic checking and man-machine interactive correction, Chinese segmentation, syntax and semantic analysis.

1 引言

随着信息技术的飞速发展，人们对电子文本的利用率越来越高。所以，文本的自动校对不仅适用于出版业，也可应用于任何使用汉字的机构。由于汉语本身的特点，中文自动校对系统始终没能像英文校对系统那样得到较大的应用。这主要有几方面的原因，一是中文电子文本的错误并不是写错字或词，而是用错字或词，是所谓“别字”而不是白字、非字，有时候，这就会造成似是而非的异体词。而英文文本恰恰相反，主要是针对不存在的非词。因此，要想使中文校对系统达到很高的精确度，语法甚至是语义的分析是必要的，这一点尤其适用于非出版业电子文本的校对。但由于汉语使用非常灵活，目前并没有完整的语言学的知识可直接应用到计算机中；二是中文的词语之间不像英文有明显的分割标记，因此处理中文电子文本之前必须进行分词或类似分词的处理，而目前并没有100%准确的分词算法，因此会使错误延续，影响后续处理的精确率；三是汉字输入计算机是使用不同的编码方法实现的，不同的校对系统会引起不同的错误情况，所以，校对系统难以用统一的方法实现查错和纠错。

目前的中文自动校对方法可分成两大类：基于语言学知识的分析方法和基于语料库语言学的统计方法。前者首先利用语言学知识对待校文本进行多层次的语法和语义分析[1]，然后确定错误位置及候选词串；后者[2]用大规模语料库得到所需的语言模型，然后应用某种概率算法实现查错和纠错。如参考文献[2]就是应用语料库得到三元语言模型，然后用校对算法实现查错和纠错。两种校对方法各有优缺点，前者存在语言学知识不足的问题，后者由于应用概率统计的方法，所以对一些很少出现的语言学现象的处理不够理想。因此，如何将两种方法结合，将是科学工作者应研究的问题。

本系统将理解分析与统计方法有效地结合起来，根据具体的实际情况用不同的查错方法，将各种方法相结合，并经实验证明是有效的。本文首先铺陈系统的总体结构，然后介绍系统的查错原理和纠错方法，最后通过实验分析得出评价系统的结论。

2 系统结构

一般来说，校对系统包括两个功能：查错和纠错。前者应用某种方法确定错误位置及候选词组，后者应用这些信息纠正错误。这里查错和纠错可以交替进行[5][6]；也可以分别进行[3]，即先查错，后纠错。

本系统应用后一种方法，即首先确定错误位置，得到候选词，即替代词，然后纠正。其中确定错误位置和寻找候选词是校对系统的关键。本系统采用分析与统计相结合的方法实现对错误字词的定位，首先应用统计的方法对待查汉字串进行分词与词性标注[9]，然后确定错误位置。这里的关键是如何确定字或词有错，本系统采用三方面的原则确定错误位置：

1) 出现了不成词的汉字串；

2) 不符合语法规则；

3) 不符合语义规则。

只要符合上述任何一点，我们就认为可能有错，然后通过计算语言学的方法得到候选词组。系统的总体结构见图 1。

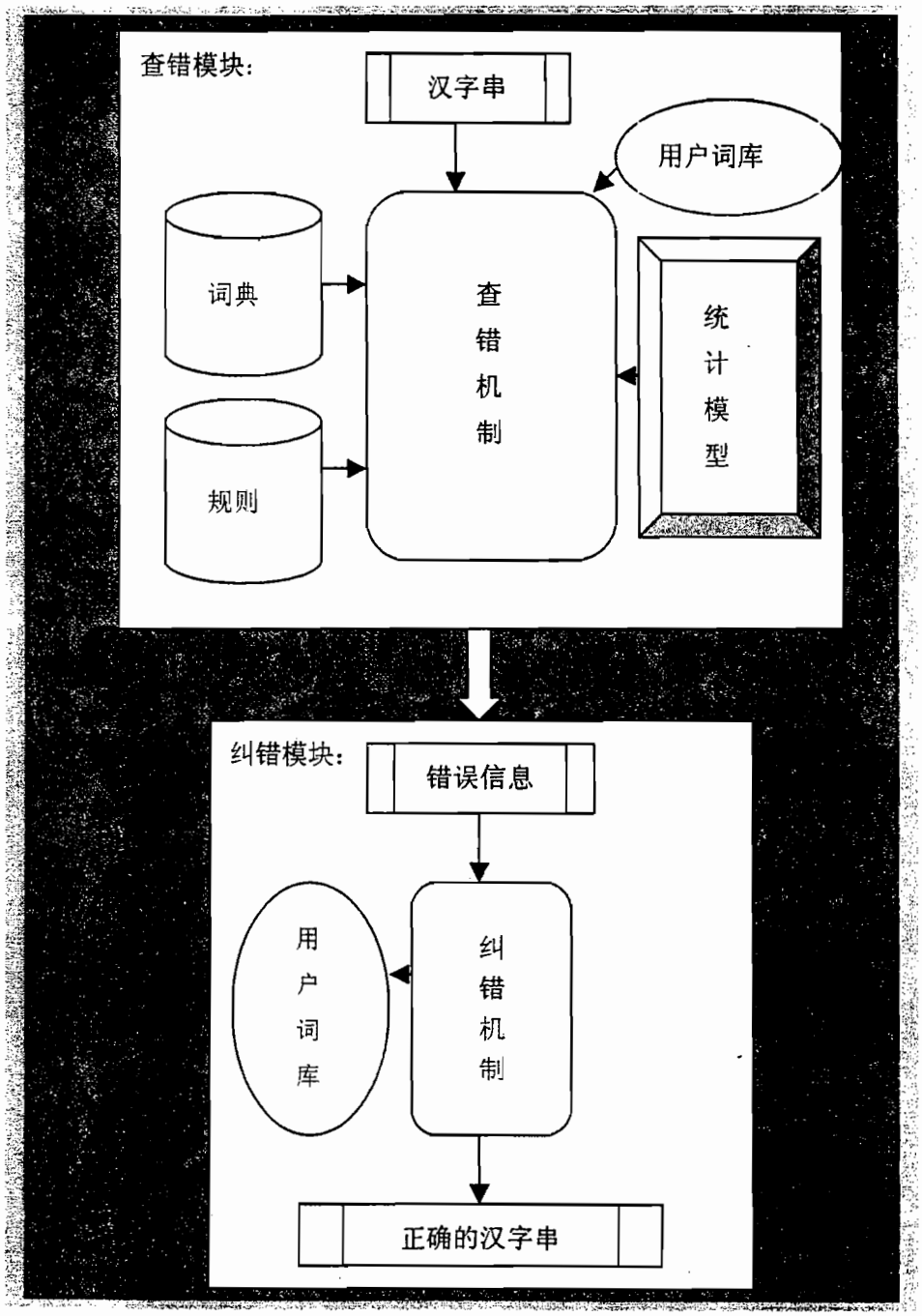

图 1 语科中文自动校对系统

3 查错原理

3.1 错误定位

本系统采用分析与统计相结合的方法实现错误的定位，首先应用统计的方法对待查汉字串进行分词与词性标注[9]，然后确定错误位置。

设 $c_1,c_2,\ldots c_n$ 是待查的纯汉字串，Ψ代表查错机制，E代表错误信息，则有：

$$E=\Psi(c_1,c_2,\ldots c_n)$$

$$\Psi(c_1,c_2,\ldots c_n)=\Gamma_3(\Gamma_1(c_1,c_2,\ldots c_n))+\Gamma_4(\Gamma_2(\Gamma_1(c_1,c_2,\ldots c_n)))+\Gamma_5((\Gamma_1(c_1,c_2,\ldots c_n))$$

其中，Γ_1 为分词模块，Γ_2 为词性标记模块，Γ_3 为散串处理模块，Γ_4 为语法分析模块，Γ_5 为语义分析模块。

3.1.1 分词模块Γ_1

本系统的分词方法采用正向最大匹配的原则，对于歧义问题，如“会诊断”中，“会诊”和“诊断”同是汉语词。为缩小系统的查错范围，本模块采用二次分词的原则，第二次分词是针对第一次分词结果的不成词串进行，如“会诊断”在第一次分词后会留下“断”为不成词串，那么经过二次分词后，“断”与前面的“诊”成词，所以，“会诊断”全部成词。由于汉语语言使用的灵活性和多样性，所以虽然本系统的词库达到 20 万条以上，但仍然不能涵盖所有的语言现象，所以生词的处理也是本系统着重考虑的问题之一。本系统采用参考文献[10]的方法实现分词的处理。

3.1.2 词性标记模块Γ_2

本模块为语法分析做准备，本系统采用二元文法的MARKOV模型实现词性标注[9]。

3.1.3 散串处理模块Γ_3

散串即在 1.1 处理后得到的不成词的汉字串，本系统把散串视为有错误的可能。如语句“我们参觐了香港展览馆”，其中“参觐了”为散串。对于散串，本系统利用错误过滤机制逐步排除串中没有错误的字。其关键是利用汉字的输入编码或同音字计算散串与词库中的词的近似度，计算公式如下：

$$SIMI\_DEGREE(S,W)=\max_{i=1}^{m} SIMI\_DEGREE(c_i,c_j)$$

$$SIMI\_DEGREE(c_i,c_j)=10*\sum_{k=1}^{n}1,\ if(codei[k]==codej[k])+$$

$$5*\sum_{k1=1}^{ni}\sum_{k2=1}^{nj}1,\ if(codei[k1]==codej[k2]) \qquad \text{EQ.1}$$

其中，S 为散串，W 为词库中等长的词，c_i,c_j 分别为 S,W 中的汉字，$codei,codej$ 分别为 c_i，c_j 的输入码，n 位 c_i 和 c_j 输入码 n_i,n_j 的最小长度。

实际系统中，如果 $SIMI_DEGREE(c_i,c_j)$超过某个域值(实验结果表明，域值为 40 时效果最好)，系统确定此串有错。如上例“觇”和“观”的仓颉码分别是“TGBUU”“GCBUU”，则按公式 1 计算近似度 *SIMI_DEGREE(觇, 观)*=60，因此，本串有错误。

3.1.4 语法分析模块Γ_4

语法分析是利用词性搭配原则寻找待校语句中的错误，因此，本过程是在词性标注的基础上进行的。如语句“他是公司的新雇员美丽”，按 1.3 的方法，并不能找出错误，但经过句法分析后，可以确定这里是有错误的。

3.1.5 语义分析模块 Γ_5

本过程用于找出语句中语义搭配不当之处，如不适当的关联词搭配，以及不恰当的量词与名词搭配等。如：“一个牛”，“虽然…，也许…”。

3.2 候选词确定

本系统通过三个层次的分析确定错误位置，那么相应错误串的候选词组也是用不同的方法得到的。

对于步骤 1.3 中确定的错误串，候选词是词库中近似度大于 40 的词；而步骤 1.4 的错误串，目前并没有给出适当的候选词；错误 1.5 的情况可通过系统的语义搭配规则库得到候选词。

4 纠错方法

本系统的纠错过程是通过实用的人机界面人机交互实现的(见图 2)。

图 2 的人机界面类似 MS WORD 的校对界面，所以用户在使用时更方便快捷。除此之外，为增加系统的学习功能，本系统为用户设置了个人词典，如果用户认为原文错误框中的内容是对的，可以在对话框中点击“新增到词库”按钮，从而将原文错误框中的内容增加到个人的词库中。

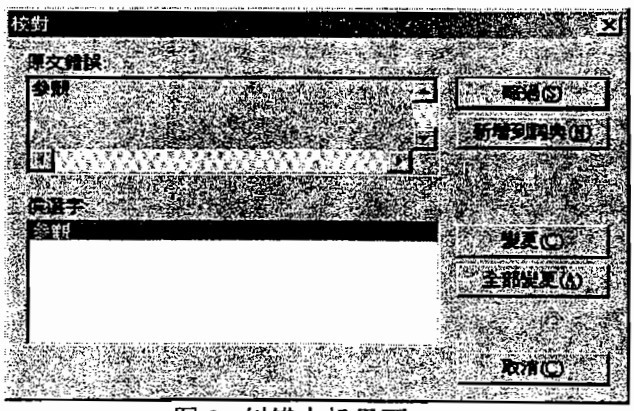

图 2. 纠错人机界面

5 实验分析

本系统是一种嵌入式软件，可嵌入到MSWORD上执行，并实现了简繁体汉字同时校对的功能。系统的中文词库由2字到7字词共20万词，规则库含有规则约500多条。

系统可校对的错误类型如下：

1. 中文字或词错误：如错字、别字、多字、少字、异体词等；
2. 中文语法错误：语句不符合语法规则；
3. 中文语义错误：不恰当的字或词的搭配，如量词与名词，配对关联词的不适当搭配或缺少其中的一部分；
4. 标点符号错误：如缺少配对的标点符号或错误连用标点符号；
5. 数字错误：如年月日数字的不恰当的应用，如“18月89号”；
6. 重复句、段错误。

为了验证本系统的校对能力，我们初步拣选了几篇样本文章，大约有2,314字。经系统测试后，召回率为93%，精确率为77%。大量的实际文本的测试还在进行中。

以下是部分运行的实例：

例1. 原文“香港城市大学位与香港九龙”，
校对结果：错误串：“位与”；候选词：“位于”。

例2. 原文“虽然香港与深圳相邻，两地的情况不同”，
校对结果：错误串：“虽然”；候选词：缺“但是”。

例3. 原文“在野的国民党称陈水扁宣布不连任就是台湾人最好的礼物。。”，
校对结果：错误串：“。。”；候选词：“。”。

例4. 原文“他于8月34号出生”，
校对结果：错误串：“34号”；候选词：日期数字错误。

例5. 原文“SARS考验台湾人的道德心和责伕感”，
校对结果：错误串：“责伕”；候选词：“责任”。

例6. 原文“这头马很健壮”，
校对结果：错误串：“头”；候选词：“匹”。

例7. 原文“另据悉，另据悉，”，
校对结果：错误串：“另据悉，另据悉，”；候选词：“另据悉，”。

例8. 原文“他们互相相帮助”，
校对结果：错误串：“相”；候选词：“”。

例9. 原文“他在教书香港”，
校对结果：错误串：“教书香港”；候选词：“香港教书”。

例10. 原文“他对长辈必恭必敬”，
校对结果：错误串：“必恭必敬”；候选词“毕恭毕敬”。

6 总结

本系统应用规则和统计相结合的方法实现了中文简繁体的文字校对，由于可嵌入到MS WORD上运行，所以，具有很好的兼容性和易用性。此外，由于配置了粤语词库，以及互动式的词典调配选项，本系统既可校对书面语文稿，也可校对粤语口语文件。我们未来的方向是要在词库标注系统、语法语义规则库、个人风格化及语料深广化方面加强研究，进一步努力实现更优化的校对功能。

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，《计算语言学的进展与应用》》，全国第三届计算语言学联合学术会议，1995.10。

[2] Li Jianhua, Wang Xiaolong & Sun Yuqi, The Research on Chinese Text Proofreading Algorithm, *High Technology Letters*, 2000, 6(1):1-7.

[3] Meknavin Surapant. Combining Trigram and Winnow in Thai OCR Error Correction. In *Proceedings of COLING'98*. Montreal, Canada, 1998, pp.836-842.

[4] Kukich. K. Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys*. 1992, 24(4): 377- 439.

[5] Mays Eric. Damerau F. J. and Mercer Robert L. Context Based Spelling Correction. *Information Processing & Management*. 1991, 27(5): 517-522.

[6] Golding Andrew R. and Schabes Yves. Combining Trigram-Based and Feature-based Methods for Context-Sensitive Spelling Correction. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, CA, 1996, pp.71-78.

[7] Kukich. K. Spelling Correction for the Telecommunications Network for the Deaf. *Communication ACM*. 1992, 35(5): 80-90.

[8] Oflazer Kemal. Error-tolerant Finite State Recognition with Applications to Morphological Analysis and Spelling Correction. *Computational Linguistics*, 1996, 22(1):73-89.

[9] 吴岩、刘挺，王开铸， 英文词性标注的抽象模型，计算机应用研究学报， Vol.15, No.3, 1998。

[10] 王开铸，李俊杰，吴岩，五词典自动分词的研究，《计算语言学的进展与应用》》，全国第三届计算语言学联合学术会议，1995.10.