中文校对软件标准评测系统的构造

郑佩 杨力平 裘照明
(IBM 中国研究中心)

摘 要：本文介绍了一种针对中文校对软件的标准评测系统的构造方法，其中包括评测指标的确定及标准评测集的建造。该评测系统是由 IBM 中国研究中心在开发中文校对系统 CEC 时建立起来的，并贯穿应用于 CEC 系统的研制开发的全过程。最后，本文给出了利用该评测系统对 CEC 及其它两个有代表性的中文校对系统的评测结果及分析。

关键词：中文校对 标准评测系统 评测指标 标准评测集

The Development of a Standard Test System for Chinese Error Check System

Zheng Pei, Yang Liping, Qiu Zhaoming
(IBM China Research Lab)

Abstract: In this paper, a method to develop a standard test system for the Chinese error check system is introduced. The standard test system, including the test index and the benchmark, was developed and applied in the R&D progress of CEC(Chinese Error Check system developed by IBM China Research Lab). At the end of this paper, the test results of CEC and other two Chinese error check systems are provided and analyzed.

Keywords: Chinese Error Check(CEC), Standard Test System, Test Index, Benchmark

一、概论

中文校对软件的研究与开发是中文应用软件技术不可缺少的一部分，其应用不仅仅限制于计算机录入、出版领域，也可作为中文 OCR 系统、语音识别系统的后处理。自 90 年代初至今，已有不少科研机构，院校及公司进入这一领域，并发布了各自的中文校对产品。

从出错原因考虑，汉语文本错误主要来自计算机录入错误和原稿本身错误。录入错误包括由于字形，字音的接近引起的输入错误，重码或联想输入中的误选，以及漏字、多字、串行等。原稿中的错误主要是作者写稿时出现的各种搭配错误，词序不当，句子主、谓、宾及必要的虚词成分的残缺，以及不同句式杂糅造成的结构混乱。

从词法、句法、语义角度分析，常见的文本错误主要包括：

1. 错字、缺字造成的构词错误
2. 由词性搭配不当，关联词语搭配不当，句子成分残缺、位置不当等句型错误造成的句法错误
3. 语义搭配错误

对于不同处理阶段的文件，错误的种类和分布相差较大。例如，文献[1]提到，对《经济日报》和《科技日报》 12 篇文章 4.1 万字进行统计（校对前的毛样，错误率 4 ‰-11 ‰），得到录入文字错误共占 80%，标点符号错误约占 10%，其它错误的比例小于 10%。而根据“第一届首都 20 家报纸编校质量评比”提供数据，三校文样（错误率为 0.6 ‰），在 55 万 6 千 4 百字中，错误共 349 处，其中错字 54 处，漏字、多字、字符颠倒 21 处，标点错误 94 处，语病 167 处，知识性错误 13 处。

通过收集、分析、整理实际中出现的各种错误，并加以归纳、总结，可以构造一个较为全面的评测库。一个标准的中文校对评测系统正是以这个标准评测库为基础，通过对校对系统运行结果的分析，提供诸如检出率、误报率等分析指标，以达到如下目的：

1. 评定系统的设计思想和技术实现
2. 评定系统的改进方案
3. 为进一步的研究、开发工作提供信息
4. 向系统用户提供校对系统的工作质量、工作效率等方面的信息。

在中文校对软件研究、开发的初期，一些产品选用了方正校对系统的一份演示文件来展示系统的性能，这份文件中包含了 29 个错误，多数为音近字，形近字错误。但对于一个完整的软件系统而言，这个文件是远不能达到上述几点目的的。

IBM 中国研究中心在开发中文校对系统 CEC 的过程中，建立起一套标准评测集及相应的评测算法，此评测系统已贯穿应用于 CEC 系统的研制开发全过程中，并将在今后系统性能改进的研究中发挥更大的作用并不断完善。

二、主要评测指标

中文校对系统在研制过程中主要存在如下几方面的问题：

1. 提高错误的检出率；
2. 降低系统的误报率；
3. 提高机器处理速度；
4. 准确提供改错建议；
5. 用户词库、专业词库等附属功能的引入，以提高系统性能。

然而，检出率与误报率之间的矛盾，校对质量与校对速度之间的矛盾等是校对系统开发中不可避免的难题，也是人们评估多种校对系统的性能优劣或同一种校对系统的改进程度的重要指标。

CEC 的标准综合评测指标主要包括如下几点：

1. 检出率

此项指标主要统计校对系统对于一般性错误，漏字，五笔字形输入法错误，拼音法输入错

误（同音字，同音词，近音字（如南方音等））的查错结果。

2. 误报率

此项指标主要统计校对系统对于长篇的正确文字的检查结果。

3. 修改建议准确率

此项指标主要统计对于各种错误，检错和纠错的准确性。为方便用户，系统自身应对修改建议自动排队，评测系统中正确的修改建议仅指出现在前五个的正确建议。

4. 系统综合指标

此指标包括系统处理速度以及上述三项指标的综合评估。

在提供总体评测指标的同时，评测系统输出详细的分析文件，显示对每一处错误校对系统提供的修改建议的顺序排列，并标明如下三种情况：

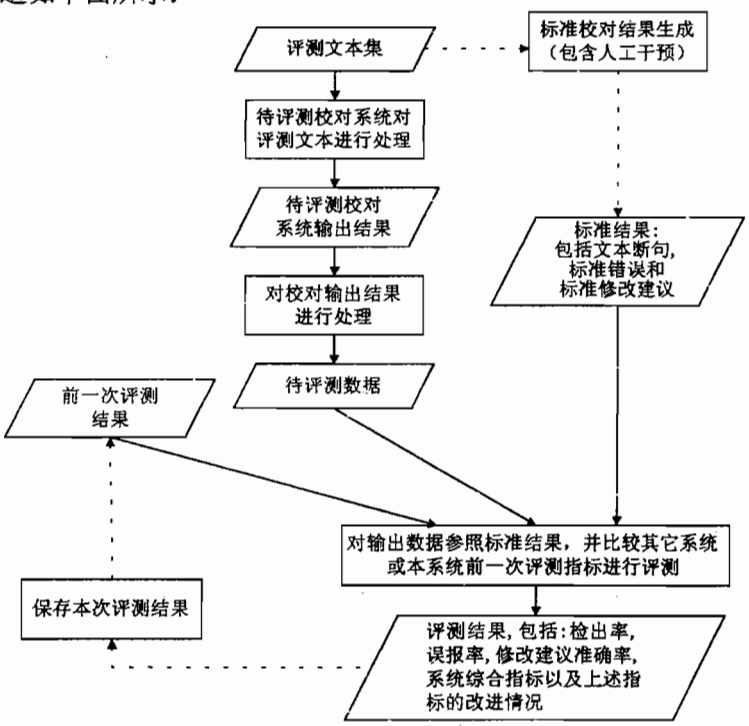
1. 检错正确并建议正确
2. 检错正确而建议不正确
3. 误报

同时提供正确答案供核对。

通过分析上述四项指标及输出文件，我们可以对被评估的校对系统有一个较全面的评价。

三、评测系统的构造

评测系统的构造如下图所示：



四、标准评测库的构造

虽然针对不同的评测指标，有不同的评测库，但总体上来说，评测库主要包括如下几方面的常见错误（其中，正确例句用来检查误报率，错误例句用来检查检出率）：

1. 音近字常见错误

音近字错误不仅包括使用拼音输入法易出现的错误，还包括语音识别的错误结果。例如：

- 进代学者王国维先生。（近代）
- 科教人员利用业余时间写作或传授技术取得一点报抽，被叫做“搞自留地”。（报酬）

2. 形近字常见错误

形近字错误不仅包括使用五笔字形输入法易出现的错误，还包括中文 OCR 的错误结果。例如：

- 美国对伊拉克进行军事于预。（干预）
- 他父亲的身体每况俞下。（每况愈下）

3. 句头、句尾出错

- 惕我们力量，他说。（赐）
- 唉叹气的生活。（唉声）

4. 漏字、多字

- 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）
- 于是就出现了处理汉字的的硬件，即汉卡。(多字：的)

5. 常见的易写错的词和成语

- 这是个神密的地方。(神秘)
- 一再告戒自己。（告诫）
- 家中还有两个熬熬待哺的孩子。（嗷嗷待哺）

6. 主要由智能输入法引起的错误

- 到达中央智慧部驻地佳县朱官寨的头一个晚上。（指挥部）
- 无法理解这种抽象型式。（抽象形式）

7. 数字及量词的搭配及识别

- 经过一年的努力，他的成绩从第二十名上升到了第一名。
- 工农业总产值为 483.5 亿元，棉花总产量为 2996 万担。
- 比 1995 年预计完成数增长 10.4 %。

8. 年月日等时间表示法错误

- 1993 年 12 我去了上海。(1993 年 12 月)
- 一九九三年十日，我去学校报到。(一九九三年十月十日)
- 1921 年 2 月 45 日，是一个令人难忘的日子。(1921 年 2 月 4 日)

9. 专有名词识别

9.1 人名识别

- 韩非子，关汉卿，安禄山都是古代人。

- 薄一波，步鑫生，郭沫若都是现代人。
- 白求恩是一位国际主义者。

9.2 称谓识别

- 杨副经理出差去了。这项任务交给郑科长来完成。
- 秦特派员是个专横的家伙。

9.3 地名识别

- 著名的平山县西柏坡革命纪念地是中国革命最后一个农村指挥所，
- 承德避暑山庄和遵化清东陵、易县清西陵分别是我国现存规模最大、保存最为完整的古典皇家园林和皇家陵寝群。

9.4 机构名识别

- 人大常委会外事委员会副主任符浩会见来访的《东京时报》千叶支社长本泽二郎夫妇。
- 美国伊利诺大学政治系主任、美中研究非洲交流委员会主席于子桥教授拜会了我协会会长李菊生，并与我协会会长李菊生、副会长刘延州、编委穆广仁、关云秋、交流委员李承曾、徐启新、学术委员袁晓利等进行了座谈

10. 标点符号

- ；在开始有一个错误标点。
- 有两个。。连续标点。
- 在创造新词的群体中，，以城市青少年群体为主。

11. 双字节英文识别

- Tom is the tallast boy in the family。(tallest)
- representitive (representative)

12. 不合理词性的搭配检查

- 他的脸很通红。（“很”不能用来修饰状态形容词“通红”）
- 很感触地说……（“很”不能用来修饰“感触”）
- 注目着那……的墙壁（不及物动词“注目”不能带宾语“那……的墙壁”）

13. 相邻或不相邻词间搭配检查（包括单字词搭配检查）

- 增加……体质(增强)
- 生活必须品(必需)
- 必需努力工作(必须)
- 看到一位老师深有体会地说(“看到”和“说”不一致)

14. “的”、“地”、“得”、“把”、“对”、“对于”和其他虚词的误用

- 对这个情况必须进行认真地调查(的)
- 他每天都起的很早(得)
- 决定对安全问题进行一次教育(“针对”或“就”)

15. 句子成分残缺

- 在老师的鼓励和帮助下，增强了她学好外语的信心，……（“增强了……”无主语）
- 听了李富英的话，使他心里感到很不舒服，就坐在一边不言语了。

("使他心里……" 无主语)

- 他的发言，表达了我们为把北京站建成现代化车站而努力奋斗。
("表达"应有一个名词性的宾语)

16. 句式杂糅

- 今年上半年报刊发行量比去年同期相比，增加了一千三百多份。

("比……增加了"与"跟……相比增加了"混淆)

- 难道这不是指的同样的东西又是指的什么呢？

("难道不……"与"不是…又是…" 混淆)

- 学习任务再重，越要坚持体育锻炼。

("越…越…"与"再…也…"混淆)

17. 构词法及词缀等（主要用于检测误报）

- 昵称：阿娟
- 拆词：吃了饭，洗个澡，快不快乐，抖一抖
- 重叠： 高高兴兴
- 词缀： -化，-性，-感

五、评测结果

在这里我们仅给出检出率，误报率，修改建议正确性评测结果

1、检出率

测试错误种类	错误总数	系统A	CEC	系统 B
1. 同音字	147	90	109	95
2. 同音词	102	4	32	27
3. 南方音	41	34	32	30
4. 形近字	101	78	85	69
5. 漏字	133	61	70	78
6. 一般错误	299	175	239	216

2. 误报率

评测文本	系统A	CEC	系统B
File A - 335K BYTE	3,124	1,424	2,510
File B - 33K BYTE	492	378	810
File C - 31K BYTE	207	89	146

*表中数字指总误报字数.

3. 修改建议

	系统A	CEC	系统B
检出错误数	172	189	164
正确提供修改建议数	81	131	57

*本测试文件中包含235处错误，主要是采用拼音输入法及五笔字型输入法时产生的音近字，形近字错误。

4. 综合性能测试

性能指标	系统A	CEC	系统B
处理速度(字符/秒)	30,000	13,000	200
检出率	83	95	94
误报率	93	60	62
正确提供修改建议数	66	82	61

上述性能指标测试是在32M RAM，主频为Pentium/100的PC机上的运行结果。

五、小结

标准评测集中涉及的部分句法错误及语义理解，知识性错误是当前各种中文校对系统尚不能解决的难题，且根据我们的了解，目前的各类产品均不存在真正的自学习性能。因此，目前对于中文校对这种属于计算机智能范畴的研究课题，提高各项性能指标只能从扩大知识库的数据量，改进数据质量入手。一个内容完整，考虑周密的校对软件评测系统是进一步提高校对软件性能的有效工具。

参考文献

1.欧阳龙根，“计算机汉语辅助校对的方法和目标”，计算机世界，60版，1996年4月15日。

2.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”，P100-105,计算语言学进展与应用，清华大学出版社，1995.10。

3.俞继东，王励，“中文校对 含苞待放”，计算机世界，56版，1996年4月15日。