中文自动校对系统的研究与实践

刘　挺　王开铸

哈尔滨工业大学计算机系

The Study and Practice on Chinese Automatical Proofreading System

Liu Ting　Wang Kaizhu

Dept. of Computer Science　Harbin Institute of Technology

校对是报刊、图书出版工作中的一个重要环节。其任务是根据原稿核对校样，订正差错，以保证出版物的质量。目前校对工作完全需要人工进行，是出版业自动化的瓶颈所在。校对的原则是“忠实于原稿”，校对的目的是消除校样中不同于原稿之处，校对的过程就是校样与原稿的比较过程。

笔者通过对人工校对方法的分析，提出了相应的机器校对模型：

1　扫描原稿（机器折校）

用 OCR 扫描原稿，然后将扫描稿与从键盘录入的校样逐句调入内存缓冲区进行比校，二文本相同之处可视为正确文本，不同之处是否为错误文本需借助词典进行判决。

2　朗读校样（机器唱校）

语音合成器将校样由文字形式变换为语音形式，校对人员听校样、看原稿、改正错误。语音合成前可进行韵律修饰，自动调节语音的长短、强弱，使语调尽量符合人的朗读习惯，在有可能出错的地方重读、慢读。

3　专家系统（机器默校）

计算机脱离原稿，根据语言学知识、校对知识以及世界知识自动推理查出部分错误，并提供侯选词，请用户确认。本模型模拟校对人员不借助原稿却能基本准确地改正校样的能力。

305 个真实错例的人工分析表明：校样中文字错占 91.1%；标点错占 6.6%，数字错占 2.3%，因而排错的重点应放在文字错上。

笔者在 PC 机上开发了一个中文自动校对专家系统，该系统以小句为单位，以全文为背景，综合运用生词的自动识别、词的模糊匹配、简单的语法分析、中文人名自动辨识等技术，经过对机内校样的三遍扫描，逐步将正确字串捆扎起来，剩余字串被判定为误。目前查错率为 50%，改错率（第一候选命中）为 30%，虚报率为 50%，由于系统具有学习功能，虚报率在使用过程中将逐渐下降。

实验表明：当查错率提高时，往往导致虚报率以更大比例增长。为解决这一棘手问题，需进行更深入的语法语义分析，并充分考虑上下文知识和背景知识。