方略知识管理系统

学科导航4.0暨统一检索解决方案研讨会

清华大学少数民族文字识别系统研制成功

http://www.fristlight.cn 2007-01-31

[作者] 科技日报
[单位] 科技日报

[摘要] 蒙古文、藏文、维吾尔文等六种少数民族文字的纸出版物要转换成电子出版物，今后不再靠人工录入，只要经“统一平台少数民族文字识别系统”处理，印刷文档的扫描图像就会自动生成可编辑检索的电子文档。这是记者2007年1月29日在清华大学举行的“多体蒙古文（包括混排汉英）印刷文档识别暨统一平台少数民族文字识别系统”技术鉴定会上获悉的。

[关键词] 清华大学;少数民族文字;文字识别系统

蒙古文、藏文、维吾尔文等六种少数民族文字的纸出版物要转换成电子出版物，今后不再靠人工录入，只要经“统一平台少数民族文字识别系统”处理，印刷文档的扫描图像就会自动生成可编辑检索的电子文档。这是记者2007年1月29日在清华大学举行的“多体蒙古文（包括混排汉英）印刷文档识别暨统一平台少数民族文字识别系统”技术鉴定会上获悉的。据项目研制主持人，清华大学丁晓青教授介绍，该系统能识别多种印刷字体的蒙古文字符和文档，并能识别蒙汉英混排的文档，是集版面分析、文本行字切分、识别、纵向文档图文对照编改等技术于一体的蒙古文文档识别实用系统，解决了多字体蒙古文汉英混排文本切分和识别问题。在实际的多字体蒙汉英文档测试集上，文本识别率可达96.89%。据介绍，该系统是全球首款在统一平台上支持我国主要少数民族文字文档的识别系统。系统在汉字和英文文档识别的基础上将四种类型六种文字的少数民族文字，即蒙古文、藏文、维吾尔文、哈萨克文、朝鲜文和柯尔克孜文（混排汉英）。文档识别综合集成在一个统一的平台系统中，使我国最主要的少数民族文字文档能够自动识别输入计算机。该系统软件产品采用国际标准编码，系统结构具有良好的扩展性，还支持阿拉伯文的识别。由倪光南、何新贵、戴浩院士组成的鉴定委员会认为：该项目解决了实用的多字体印刷蒙古文文档及其混排汉英的识别问题，实现了在统一平台上蒙、藏、维、哈、柯、朝（混排汉英）文档识别的综合集成，其主要技术指标达到了国际领先水平，对促进我国少数民族语言文字的信息化建设具有重要意义。

北京雷速科技有限公司 Copyright © 2003-2008 Email: leisun@firstlight.cn