**《中文文本信息处理的原理与应用》清华大学出版苗夺谦卫志华**

**中文信息处理原理与技术提要**

（一）、中文信息处理的主要任务

1、自然语言处理：语意理解——解释、翻译、文摘生成、回答问题。

2、自然语言处理：语音识别——音素、音节、单词、短语、句子；语音合成——单元、韵律。

3、多语种环境：建立适应多语种处理的计算机系统、网络系统、嵌入式系统环境。

4、多语种应用：多语种混合教学系统、通信系统、发布系统、理解和翻译系统。

（二）、中文信息处理理论及技术基础

1、语言学知识：

语音——音素、音节、单词、韵律；

语言——词素、单词、短语、句法；

语义——理解、上下文关系、表达、转化。

2、算法——统计与规则的结合

HMM (隐含马尔科夫模型)模型是语音识别中常采用的统计模型，状态的概率函数能够描述语音的重要属性特征的分布特点，以提供给模式分析/识别模块使用。

Gabor小波变换是一种多通道信号分解，能同时在时域和频域中较好地兼顾对信号分析的分辨率要求，快速Gabor变换应用。

SVM的分类方法，结构风险最小化原则的SVM技术具有全局最优性和较好的泛化能力。采用SVM技术能够提高分类性能。

Lattice采用基于音素的置信度与基于状态的置信度相融合的方法，提高评估准确性。

3、资料基础：语料库——音素库（三音子库）、音节库、词库、句法库，不同层次的自然语言原始语料素材库；语料库的标注——切分、主要属性标注；连接库——连接规则、概率。

（三）、中文信息处理（阿勒泰语系语言）的主要特征

1、语言特征：维吾尔语、哈萨克语属于阿勒泰语系突厥语种，是一种黏着语：名词有数、从属人称、格等语法范畴；动词有态、肯定否定、语气、时、人称、数、形动词、动名词、副动词等语法范畴。词组和句子有严格的词序：主语在谓语之前，限定语在中心词之前。维吾尔文用阿拉伯文字母为基础拼写。现行维吾尔文有8个元音字母，24个辅音字母。

2、文字特征：维吾尔文字自右向左横写；每个字母按出现在词首、词中、词末的位置，有4种不同的变形书写方式，32个字母共有128种字形；维吾尔文字母还有不等宽的特点。维吾尔文词根上的语音变化、附加成分与词干在元音和辅音上的和谐的语言特征也出现在文字上，同一个词根往往有不同的书写形式，同一个附加成分往往有几种变体。

3、与阿拉伯文的近似性。

（四）、中文信息处理（阿勒泰语系语言）依据的标准

1、基本标准：信息交换用编码标准、字形标准、术语标准，以及键盘标准、语料库标准。

2、现行编码标准：在ISO/IEC 10646-1:2000/Unicode 4 中的阿拉伯文字符集已包括了所有的维吾尔文、哈萨克文、柯尔克孜文常用的字符形式。但由于维吾尔文的变体显现形式与阿拉伯文不完全对应或不完全相同，为确定和规范维吾尔文、哈萨克文、柯尔克孜文字符集在ISO/IEC 10646-1:2000/Unicode 4 中的准确位置与编码，制定标准DB65/2190 —2005。本标准已为国际标准化组织接受。

另外还有“维哈柯文信息交换用字型标准”、“维哈柯文信息交换用界面标准”。以及键盘标准，哈、柯古文献文字国际/国家代码标准；维、哈、柯整词输入标准；维、哈、柯文自然语言处理语料库规则标准；维、哈、柯文信息处理用音节标准；维、哈、柯文信息处理交换用文献规范名词术语标准。

3、在确定了基本字形的基础上，制定了16种字体标准，以及维、哈、克文字的标准字库，以及矢量字库。

（五）、中文多文种处理基础系统研究

1、基础平台研究：重点解决操作系统平台对多文种的支持，要解决的问题——系统的输入与输出，桌面的编排（横排与竖排）、文字和句子方向（从左到右与从右到左）、不同字母和在不同位置是字符宽度选择；汉、英、其他多文种混合编辑技术。

在基础平台的研究方面也经历了外挂式结构（不改变操作系统结构，对输入多文种进行处理后进入显示系统），内核修改（利用操作系统的接口和函数完成多文种处理），以及现在按照国际标准，通过建立支持多文种系统模块，实现多文种处理能力的三个阶段。

目前在内部处理上较好的是使用坐标动态变换技术，在操作系统底层实现对多语种文字处理的支持，输入、编辑、排版完全符合多语种文字的使用习惯。用户界面如菜单、工具条等控件符合阿拉伯文靠右对齐的使用习惯。

2、跨平台多语种处理技术：基于Java的字母自动选形技术、双向光标控制技术、按音节断行、字母自动拉长、界面控件映射技术、支持多语言的应用集成技术等多语种处理关键技术。

通过具体研究开发对象，建立跨平台的多语种文字处理架构并集成多语种处理模块，实现跨平台处理。多种语言混合处理技术，实现不同风格文字混合编辑处理，文字方向、排版方式等都符合使用习惯，实现按照上下文环境智能识别、字符显示宽度自动调整。

3、多语种支持软件：多语种Office套件，图形、图像处理软件、支持多语种数据库软件，多语种系统电子邮件，电子词典软件，支持多语种杀毒软件。

4、多语种信息网络传输技术。基于互联网的多语种处理技术，采用IDS（信息动态服务）和CGI（公用网关接口）技术相结合的方法，解决了不同方向文字信息在互联网上的信息处理、网上检索、排序和统计技术。实现互联网上多种语言数据传输、存储与转换等操作的一致性与完整性。

（六）、中文多文种处理应用系统研究

1、多文种发布系统：多文种电子政务系统，办公自动化系统；公众服务系统：多文种电话查询应答系统——识别询问文种，自动进行信息匹配，将文本信息转换为语音文件，实时地将文本信息合成语音并通过电话发布。

2、多文种教学系统：多语种多媒体教学系统，实时交互、双屏或多屏、图像与语音的结合。

3、多文种语音识别和语义理解系统：建立用于多语种信息处理的大型关键性基础资源库(语种库、语音库等)，提高多语种机器翻译和基于内容的互联网多语种信息搜索，信息处理的速度，实现高速识别、语义分类、智能化高速处理。

4、多文种手写笔迹的识别和鉴别。采取了与文本无关的Gabor变换方法来提取笔迹图像的全局纹理特征是可行的方法。可采用的分类器有SVM分类器， k近邻分类器、基于欧氏距离的分类器以及神经网络分类器。可经过筛选比较使用。

5、嵌入式多语种多平台软件：例如手机输入法软件有广阔的应用前景。