·成果简介·

# 蒙古文整词输入法重码词智能化选择输出方法与技术研究进展

S·苏雅拉图　　白双成　　六　月

（内蒙古社会科学院 MIT 研发中心，呼和浩特 010010）

［关键词］　蒙古文，词语组合，知识表示，人工智能，重码词选择，输入输出

在国家自然科学基金的资助下，“蒙古文整词输入法重码词智能化选择输出方法与技术”研究项目取得了很大进展。

## 1　研究成果

### 1.1　逐步建立和形成了庞大的蒙古文语料库

用课题组已有的《蒙古文整词输入法》输入完成的语料目前已可以满足各种不同目的的研究。按照文本文件计，语料库目前已达到几百种。按照不同语言语体计，语料库已达到几十种。按照整词字数计，已接近几个亿。按照词条计，已接近几万词条。

### 1.2　重码词获取方法的研究成果

由于蒙古文整词输入法发明采取的是“整词音节集合模糊编码方法”，所以产生重码词的概率较高，平均为全词的 10%。重码词的概念是：输入码一致、输出值不同的词。为了获得所有的重码词，我们研究实现了一个“重码词获取软件”，通过将其应用于蒙古文整词[1]输入法输入完成的大量文本文件数据（语料库）中，获得了所有的重码词，并按其不同的聚合值，编成了 11 本“重码词词典”。

### 1.3　重码词搭配关系的研究成果

利用已获得的“重码词词典”，确定了重码词 11 种不同的聚合结构关系范示。根据研究资源与课题批准完成时间，我们将其中的二值结构范示和三值结构范示确立为本次课题阶段的重点研究对象，确定了具体的计算域。

### 1.4　重码词知识表示的研究取得多项可持续成果

我们通过前一项自然科学基金课题持续研究产生的“蒙古语框架知识库”（该知识库在内蒙古蒙科立软件公司的各种产品中已得到广泛应用，正在取得良好的技术效益）的框架体系结构，提出了重码词复杂特征知识的表示方法。复杂特征具体给出方法：在“蒙古语框架知识库”已有的属性字段上再增加新的属性字段，实现了重码词复杂特征知识的表示。在“蒙古语框架知识库”中新增加的字段具体包括：多变体附加成分[2]字段（嵌入后台）；短语结构字段；词类标注字段；句法成分字段；标点符号字段；施事受事字段；重码词后项搭配规则字段；重码词前项搭配规则字段等。通过这一方法，丰富了“蒙古语框架知识库”的知识含量，为知识库完成后可能承担的任务储备了新的知识资源。

### 1.5　重码词选择方法的研究成果

通过研究，我们提出了词频概率关系选择方法、多变体附加成分组合关系选择方法、短语结构关系选择方法、句法组合关系选择方法、语义组合关系选择方法等 5 种不同的重码词智能化选择方法。

（1）词频概率关系选择方法。这种方法是通过重码词概率统计器，对整词生成器生成出来的重码词进行使用概率的统计计算，并根据计算结果对当前的重码词进行使用概率排序后，将结果交给输出端输出的智能化选择[3]。

（2）多变体附加成分组合关系排歧选择方法。这种方法是，通过对重码词后项词中是否可以搭配多变体附加成分，来确定选择或排除哪个的智能化选择方法。

（3）短语结构关系排歧选择方法。包括固定短语组合关系排歧选择方法、非固定短语组合关系排歧选择方法。前者是根据固定短语组合关系排除其

本文于 2005 年 8 月 31 日收到.

余重码词。

(4) 句法关系排歧选择方法。包括,(i) 句法成分排歧选择法,根据重码词在句子中承担的主要成分的特点,来确定选择或排除哪个;(ii) 词类组合关系排歧选择方法,根据重码词前后搭配何类词的特点,来确定选择或排除;(iii) 标点符号排歧选择方法,根据重码词后项是否搭配标点符号,来确定选择或排除哪个。

(5) 语义关系排歧选择方法。通过对重码词前后项搭配关系中是否存在"施事/受事关系"确定选择或排除哪个的智能化选择方法。

### 1.6 取得了重码词选择实现技术研究成果

经研究,实现了一个"屏幕抓词引擎"(工具软件),该引擎可以把已经被输出的重码词的屏幕信息以同步并行处理方式迅速传输给真实指针并行处理。

### 1.7 取得了《蒙古文整词智能化输入输出系统 V4.0》软件著作权

基于以上所取得的多项可持续应用基础研究成果,在课题预定的阶段里,完成了蒙古文整词输入法重码词智能化选择输出技术,取得了《蒙古文整词智能化输入输出系统 V4.0》软件。蒙古文整词输入法重码词智能化选择输出技术的最终成果表现为一软件模块,该模块现被集成在《蒙古文整词输入法》,构成了内蒙古蒙科立软件有限责任公司的《蒙古文整词智能化输入输出系统 V4.0》,已获得国家软件著作权,并列入 2005 年度重点新产品计划。相关新技术正在申报新的发明专利。

## 2 存在问题

经过两年的研究,我们发现,重码词智能化选择输出方法与技术研究,是一项特别复杂的人工智能方法与技术研究,因此在较短时间内、用较为少量的资金投入、用较为有限的技术力量来解决全部问题,还有巨大的困难,甚至很大的技术风险。从课题目前的技术储备、研究条件等各方面综合情况看,主要存在以下几个问题:

(1) 在技术层面上目前只解决了"二值和三值聚合关系重码词",还未解决"四值聚合关系"以上重码词的选择方法与技术难题。

(2) 大部分选择集中在第一层次关系(父关系)上,第二层次关系(子关系)还有待进一步扩展。

(3)"重码词后项前项混合搭配规则"的计算因其计算复杂性[4,6]原因,系统开销方面还存在一定的技术问题。

(4) 面向少数民族自然语言信息处理[5]的自然语言研究远没有成熟,例如:词类分类理论与体系、句法理论[7]与体系、语义研究成果[8]等基础研究方面,可利用的东西不多,由此给研究带来了很大的困难。

(5)"蒙古语框架知识库"还需进行进一步的深入研究和完善。

## 参 考 文 献


[1] S·苏雅拉图. 蒙古文整词编码研究. 中文信息学报,2001,15 (2):57—64.

[2] S·苏雅拉图. 蒙古文多变体附加成分智能化处理研究. 中文信息学报,2000,4:59—64.

[3] S·苏雅拉图. 蒙古语动词计算机生成研究. 计算机学报,2002,11(11):1200—1205.

[4] 冯志伟. 数理语言学. 上海:上海知识出版社,1985.

[5] 冯志伟. 自然语言的计算机处理. 上海:上海外语教育出版社,1996.

[6] 张鸣华. 可计算性理论. 北京:清华大学出版社,1984.

[7] 清格尔泰. 蒙古语语法. 呼和浩特:内蒙古人民出版社,1991.

[8] 诺尔金. 蒙文原理. 呼和浩特:内蒙古教育出版社,1987.


# RESEARCH ON DEVELOPMENT OF REPEATED CODE WORDS INTELLIGENT SELECTION OUTPUT METHOD IN MONGOLIAN WHOLE WORDS INPUT METHOD


S. Soyolt Bai Shuangcheng Liu Yue

(*Research and Developing Center of Natural Language, Inner Mongolian Social Science Academy, Huhhot* 010010)



**Key words** Mongolian language, words combination, knowledge expression, artificial intelligent, option of repeated code words, output and input