藏语文本信息处理的历程与进展

江荻

中国社会科学院民族学与人类学研究所

论题

- 为什么要关注藏语文本处理
- 藏语文本统计与熵值计算
- 藏语文本资源与电子词典
- 藏语分词与组块识别方法
- 藏文的拉丁转写
- 藏语电子词典排序

- （本文讨论不包括：藏文编码、字形规范、藏文平台、文字识别、藏文排版等等应用或开发研究）

为什么要关注藏文文本处理

- 藏语是非常古老的语言
  - 始于公元7世纪藏民族先民建立吐蕃王朝时期
- 藏文文献浩如烟海，内容广泛
  - 碑文,岩刻,敦煌石窟手卷,竹木简牍,木刻文献
- 当代藏语和藏文使用人口超过500万
  - 藏民族重要的社会交流、教育和文化工具
- 信息化是藏族文明遗产保存和发展的途径
  - 藏语信息化和网络化是藏民族发展的需求
- 文本是自然语言处理的主流领域和基本范式

藏语文本处理起步阶段：静态统计

- 藏语文本处理起步于90年代初期
  - 江荻. 藏语动词音变现象的统计分析. 民族语文. 1992:(4):47-50
  - 江荻, 董颖红. 藏文信息处理属性统计研究. 中文信息学报. 1995:(2):37-44
    - 首次获得有关藏语字、词及字母符号的静态统计结果
    - 首次获得藏字平均字长与字母构词频度静态统计数据
  - 陈玉忠博士认为，“这一工作虽然只对一少部分藏字进行了静态的统计，但这一工作的意义则远远大于结果本身”

藏语文本处理起步阶段：动态统计

- 扎西次仁：<中华大藏经·丹珠尔>藏文对勘本字频统计分析，**中国藏学. 1997:(2)**
  - 1000万字<大藏经>历史文本统计，获得许多有价值的字频统计数据
  - 平均字符数为2.54，构成句的平均构件数为25个
  - 前15个高频字累积频率达到29.22%，比较汉语15.21%
- Jiang, Di. The Phonological Construction of Tibetan Words and Its Frequency Phenomena. **Waseda University. Tokyo, 1998**
  - 100万字现代藏语文本字频和结构统计
  - 不同字形数达到5581字
  - 藏字结构共有25类
  - 前40个藏字占全部语料频数的33%，又只占全部语料用字的0.7%

藏语文本熵值计算

- 字母及字丁熵值计算
  - 江荻. 书面藏语的熵值及相关问题, <1998年中文信息处理国际会议论文集>, 清华大学出版社.
    - 20余万字文本语料作小规模字母信息熵估算
  - 严海林等. 藏文大藏经信息熵研究. 中国少数民族多文种信息处理研究与进展. 呼和浩特. 2004
    - 大藏经藏语字丁(768) 信息熵计算

语种	符号数	零阶熵	一阶熵	二阶熵	三阶熵	极限熵	冗余度
英语	27	4.76	4.03	3.32	3.1	1.4	0.71
汉语	6763	12.72	9.71			7.64	0.519
藏文字母	40	3.99	1.25				0.766
藏文字丁	768	9.59	4.80	3.12	2.70	2.70	0.72

藏文信息熵的价值

- 从冗余度数据可以看出书面藏语的信源属性，藏文字丁的冗余度说明藏语文本信源有约72%的多余度。即书写藏文时，有72%是由语言文字结构（字、词、句）规定了的，可自由选择的可能只是28%。这也意味着藏文中有约3/4的字母符号不是用来传递消息的，而是用来保证这些字母的组合符合藏语的组词、构字及有关语法的规则。

- 藏文字丁的冗余度大还说明，藏文字丁虽然数量较大，但字丁的使用集中在少数字丁上，而且上下文关联较大，在实际的藏文信息处理中，这是有好处的。首先，藏文字丁的使用不同频且集中在少数字丁上，对藏文信息的压缩和编码是很有利的；其次，藏文文字的上下文关联较大，也就是有比较严格的语法语义规则，这除了对藏文信息的压缩和编码有利外，对藏文的文本自动处理也是很有好处的，譬如在藏文的自动分词、藏文识别以及藏文语音识别中都是很有好处的

藏语字或音节熵值计算

- 字或音节熵值计算
  - 王维兰, 陈万军. 藏文字丁、音节频度及其信息熵. 术语标准化与信息技术. 2004:(2)
    - 藏文字熵（音节熵）
    - 双字音节的相对熵为8.069，意味着在3601个双字音节排序表中选取一个特定的双字音节，需二分查找8.069次，显然比任意选取一个双字音节需要11.814次的二分法查找改善了许多。

字丁或音节	数量	相对熵值	绝对熵值	字丁或音节	数量	相对熵值	绝对熵值
字丁	521	5.8784022	5.8784022				
单字音节	475	5.7227485	2.2216804	三字音节	902	8.01503618	2.0700541
双字音节	3601	8.0692044	4.3840819	四字音节	896	5.7244898	0.2071008

藏语资源的性质与分类

- 周季文等, 藏语计算机统计用语料抽样文本的遴选. 《中国少数民族语言文字现代化文集》. 北京: 民族出版社, 1999
  - 藏语文本性质复杂，不能拿来就用
  - 藏语各地差别巨大，不能通话，书面藏语也差别很大
  - 佛教经典与日常用语差别很大
  - 古今文献差别很大

藏语文本多种分类

- 文本分类标准及平衡语料库
  - 以下分类是建立藏语文本语料库需要考虑的要素
  - 题材分类：1文学类，2政论历史类，3专门类（宗教、历算、因明、藏医药）
  - 文体分类：1散文体，2韵文体，3散、韵混合体。
  - 语体分类：1古体文言，2质朴文言，3藻饰文言，4浅近文言，5口语
  - 著译分类：1著述，2翻译（梵文或汉文）
  - 时代分类：8－9世纪；10－13世纪；14－17世纪；18－19世纪；20世纪
  -

早期藏语平衡语料库数据（98）

- 藏语初级平衡语料库（98），全部语料约500万音节字
  - 题材分类，文学类：占总语料的61.78%；政史类：占总语料的31.42%；专门类：占总语料的6.80%。
  - 文体分类，散文：占总语料的58.87%；韵文：占总语料的0.77%；散韵混合：占总语料的40.36%。
  - 语体分类，古体文言：占总语料的1.66%；质朴文言：占总语料的29.76%；藻饰文言：占总语料的17.02%；浅近文言：占总语料的48.94%；口语：占总语料的2.62%。
  - 著译分类，著述：占总语料的79.66%；翻译：占总语料的20.34%。
  - 时代分类，8－9世纪，占7.77%；10－13世纪，占1.91%；14－17世纪，占22.88%；18－19世纪，占16.07%；20世纪，占51.37%。

大型藏语平衡语料库数据（2003）

- 卢亚军, 马少平, 张敏, 罗广. 基于大型藏文语料库的藏文字符、部件、音节、词汇频度与通用度统计及其应用研究.《西北民族大学学报》2003, 24:(2)
- 4000 万音节字大型藏文语料库。
  - 该语料库将语料分为7类，
    - 报刊类、文学类、教育类、科技类、佛学类、历史类、传统文化五明类。
  - 在对语料进行领域分类的基础上，再按作品的文体、风格、年代、篇幅进行人工遴选

藏语电子词典（1）

- 卢亚军, 马少平, 张敏, 罗广. 基于大型藏文语料库的藏文字符、部件、音节、词汇频度与通用度统计及其应用研究.《西北民族大学学报》2003, 24:(2)
  - 以《藏汉大辞典》为蓝本
  - 目的是进行藏语文本统计
  - 该词典收入34124个词条

藏语电子词典（2）

- 陈玉忠, 俞士汶. 藏文信息处理技术的研究现状与展望.《中国藏学》2003:(4)
  - 目的是为藏汉机器翻译服务
  - 该词典总规模已达18万余条
  - 带词性标注

藏语电子词典（3）

- Jiang Di, Long Congjun, Zhang Jichuan. The Verbal Entries and Their Description in a Grammatical Information-Dictionary of Contemporary Tibetan. Natural Language Processing – IJCNLP 2005. springer
  - 目的是开展藏语分词以及句法分析
  - 收录12万词条的分词用大型词典
  - 拉萨口语词3万词条的语法信息词典
  - 带多达20余条词法和句法属性信息
  - 部分词语带释义例句

藏语分词：分词规范

- 罗秉芬等, 藏文计算机自动分词的基本规则.《中国少数民族语言文字现代化文集》.民族出版社, 1999
  - 最早提出藏语分词原则，早期的尝试
    - 藏语36条分词基本规则
    - 词分类及词类标记
    - 黏着语素词、屈折形态词处理方法
    - 重叠形式、四字格成语、音译外来语、名词前加成分a、数词中嵌垫音形式处理方法

藏语分词：分词方法

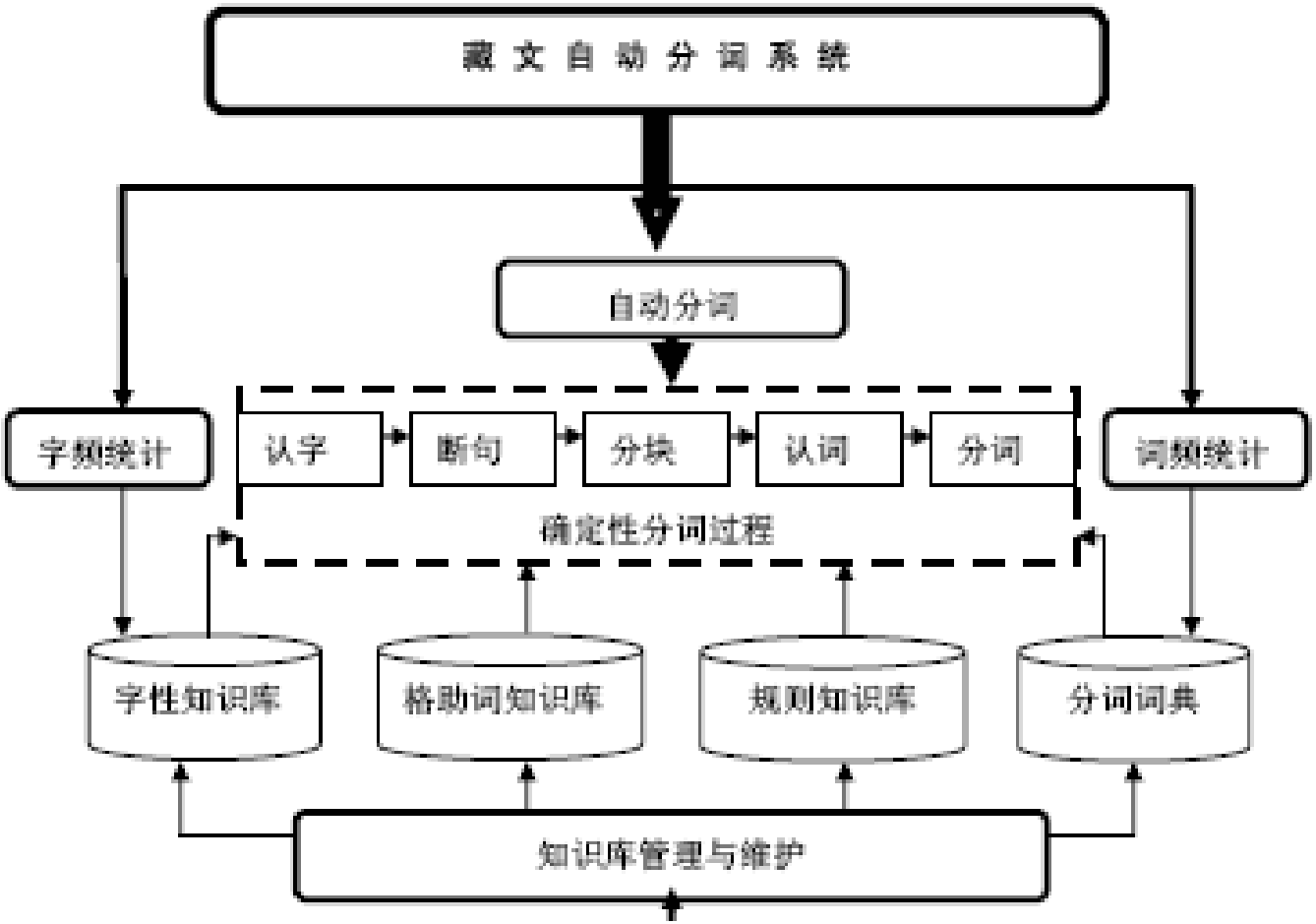
- 扎西次仁. 一个人机互助的藏文分词和词登录系统的设计《中国少数民族语言文字现代化文集》. 民族出版社.1999
  - 未预先建立匹配词典，而是在运行过程中通过人机互助逐渐增加词条，扩大词表，所以又称为词登录系统
- 江荻, 黄行. 藏语语料库语言学研究. 中华社科基金课题（97BMZ009）报告.2000
  - 单纯机械式的匹配方法，很难对藏语进行有效的分词。
  - 采用最大匹配法机械式扫描句子则会出现两类严重问题。
    - 一是经常会出现长词覆盖短词，造成切分盲点。
    - 二是扫描词串中包含了独立的句法标记，容易导致匹配上的误识
    - 实词与虚词造成的同形词影响切分准确性

陈玉忠分词方案

- 陈玉忠, 李保利, 俞士汶. 藏文自动分词系统的设计与实现.《中文信息学报》2003:(3)
- 陈玉忠, 李保利, 俞士汶, 兰措吉. 基于格助词和接续特征的书面藏文分词方案.《语言文字应用》. 2003:(1)
  - 这两篇论文是迄今为止藏语文本处理最重要的研究
  - 这个系统是目前唯一实现的藏语分词系统
  - 作者称该方案为基于格助词和接续特征的分词方案
  - 其实可以说该方案属于知识库语法规则分词方案

陈玉忠方案简述

该方案“利用字切分特征和字性库先‘认字’，再用标点符号和关联词‘断句’，用格助词‘分块’，再用词典‘认词’”，最终达到分词的目的



组块识别

- 江荻. 现代藏语的机器处理及发展之路.《汉语自然语言处理若干重要问题》.科学出版社. 2003
- 江荻. 现代藏语组块分词的方法和过程.《民族语文》2003
  - 组块识别主要指依据藏语各类句法标记切分短语或组块
    - 词格标记（14类词格）
    - 名词化标记（24种标记）
    - 结构助词（12类助词）
    - 体貌-示证标记（9类体标记，4类示证标记）
    - 单字词缀（复数后缀、敬语词缀、中嵌词缀等）
    - 部分封闭词类（如指示词、连词、语气词等）

组块类型

- 10类藏语句法组块类型及相关形式特征
  - 名词组块
  - 形容词组块
  - 非谓动词组块
  - 谓语动词组块
  - 前修饰语组块
  - 后修饰语组块
  - 小句组块
  - 从句组块
  - 游离语组块
  - 助词组块

组块识别

- 目前已开展的组块识别研究：
  - 谓语或非谓语组块
    - 江荻. 现代藏语谓语动词的识别与信息提取. 2003
    - 江荻, 龙从军. 藏语非谓动词的标记及自动识别策略，2003
    - 龙从军, 江荻. 现代藏语带助动词谓语组块的识别方法，2004
  - 形容词组块
    - JiangDi, HuHongyan. The Construction And Identification Approaches Of Adjectival Predicate In Modern Tibetan，2005
  - 名词组块
    - 黄行, 孙宏开等. 现代藏语名词组块的类型及形式标记特征，2005
    - 江荻：藏语拟声词研究，2006
  - 小句与助词组块
    - 江荻. 现代藏语致使动词句中宾语小句的边界识别，2006
    - 江荻. 藏语述说动词小句宾语及其标记， 2006

块内信息抽取与分词

- 特定组块蕴含了丰富的句法信息，例如谓语组块。抽取这些信息能辅助判断同形词
  - 谓语组块不同的体貌标记和示证标记能揭示谓语动词的句法语义类别
  - 谓语组块能指明主语或宾语是否带词格标记，或者带何类词格标记，
  - 谓语组块还能判断主语属于哪类人称代词或名词
  - ……

块内信息抽取与分词

- 目前的研究：
  - 江荻. 现代藏语动词句法语义分类及相关句式. 中文信息学报. 2006:(1)
    - 性状动词，动作动词，心理动词，感知动词，变化动词，趋向动词，述说动词，关系动词，领有动词，存在动词，互动动词，致使动词
  - 江荻. 现代藏语谓语动词的识别与信息提取. 2003
    - 谓语所蕴含的句法信息

文本研究的方法和技术（1）

- 虚词的形式化描述
  - 陈玉忠 俞士汶，面向信息处理的藏语虚词的语法信息表述研究，《中文信息学报》
  - 形式化表述框架
    - 格助词
    - 时态助词
    - 语气助词
    - 接续词
    - 词缀

文本研究的方法和技术（2）

藏语标注集

面向机器处理的现代藏语句法规则和词类、组块标注集。江荻，孔江平主编《中国民族语言工程研究新进展》社会科学文献出版社. 2005

- 14大类标记：词类，组块，词格，名词化…
- 183小类标记，例如“格”：施格，位格，向格，与格，领格，对象格，属格，从格，同类比较格，异类比较格，排他格，结果格，工具格，通格

– 该文近10万字，是面向自然语言处理的藏语语法体系框架

藏语分词的方向

- 组块识别与分词是基于藏语文本结构特征的方法
- 构建大容量句法和词法信息的电子词典是组块分词的根本基础
- 殊途同归
  - 陈玉忠的基于格助词和接续特征的分词方案本质上也是句法组块分析

藏文的拉丁字母转写

- 按照国际标准化组织的要求，所有非拉丁字母文字都应制定相应的拉丁字母转写标准和方案
- 目前各类转写方案均为非完全方案

藏文转写方案的基本原则

- 藏文文字系统的拉丁转写原则
  - 系统性
  - 逻辑关联性
  - 层次性
  - 可还原性
- 藏文转写中不存在梵文转写问题
  - 所有进入藏文系统的梵文字母都经过了藏文字形的改造、结构的变化和规则的更新，它们已经成为藏文系统的一部分，称为梵文藏字（Tibetan-transliterated character system from Sanskrit），或梵源藏文字母

转写内容

- 藏文转写对象
  - 藏文本体字符
  - 梵源藏字字符
  - 藏文以及梵源藏字组合字符
  - 结构和规则
  - 辅助性符号
    - 别义符
    - 标点符
    - 装饰符
    - 数字符

转写研究

- 研制藏文转写方案的基本原则与实践
  - 江荻. 藏文的拉丁字母转写方法----兼论藏文语料的计算机转写处理，民族语文. 2006:(1)

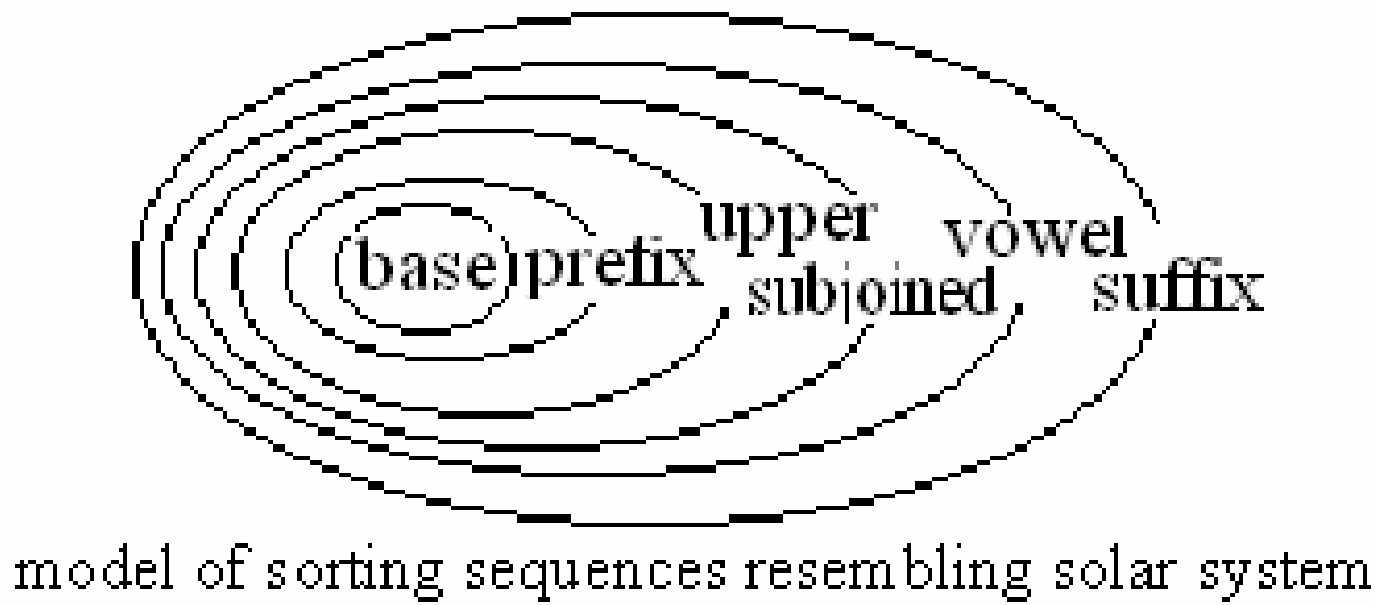
- 藏文转写规则
  - 二十二条基本转写规则
  - 转换准确率高达99.99%

藏语词典排序方法

- 藏语词典排序有何难度？可作比较
  - 英语按照字母序解决：简单约定序
  - 汉语按照专家制定顺序排列：非常约定序
    - 笔画
    - 部首
    - 拼音
    - ……
  - 藏语是字母序加传统约定结构序：复杂约定序

藏语词典序模型

– 这个排序模拟模型外观上类似于太阳系模型



model of sorting sequences resembling solar system

藏语词典排序方法研究

- 研究状况: 基本解决藏语排序问题
  - 江荻, 周季文. 藏语的序性及排序方法《中文信息学报》. 2000:(1)
  - 江荻, 康才畯. 书面藏语排序的数学模型及算法. 计算机学报. 2004:(4)
  - Jiang, Di (2006): The Current Status of Sorting Order of Tibetan Dictionaries and Standardization, *The 20th Pacific Asia Conference on Language, Information and Computation: Proceedings of the Conference*. Tsinghua University Press, 2006.

结束语

- 藏语文本信息处理经历了大约十年多一点时间，从简单的定量统计发展到语法信息词典构建，分词，以及句法语义分析，每一步都开拓一个新的视界，每一项都为藏语信息处理铺奠新的基石，我们相信，藏语自然语言处理的发展前景光明 ！

- 谢谢！