少数民族语言机器翻译资源与技术瓶颈

李 淼
中国科学院合肥物质科学研究院 智能机械研究所
2012-09-20 西安

报告提纲

一、少数民族语言机器翻译现状

二、资源瓶颈

三、技术瓶颈

四、实验室工作简介

五、总结：一些可借鉴的工作

一、少数民族语言机器翻译现状

- 研究主要集中于藏、蒙、维等少数几种语言
  - 藏
    - 1995年，陈玉忠、李延福等实现了汉藏科技机器翻译系统的原型系统；
    - 2000年，青海师范大学实用化汉藏机器翻译系统基本达到实用化水平；
    - 2004年，祁坤钰构造了初步适应英藏机器翻译的藏语语义分类体系。
  - 蒙
    - 在蒙古语词类自动标注系统、蒙汉词典、面向政府文献的汉蒙机器辅助翻译系统、蒙古语语法信息词典框架设计、英蒙机器翻译系统、蒙古文信息处理平台研究（软件、语料库）等多个方面取得成果。
  - 维
    - 在大型英汉维三向通用电子词典、维汉机器辅助翻译系统等应用软件的开发方面取得显著成绩；
    - 提供各类民文信息处理用规范与标准制定、民文信息处理语法、句法、词法规则库等。

一、少数民族语言机器翻译现状

一些指标

- 机器翻译评测最好成绩（**CWMT2011**）

语言	领域	BLEU	单位
蒙汉	日常用语	0.2275	自动化所
藏汉	政府文献	0.5843	自动化所
维汉	新闻领域	0.5144	计算所
哈汉	新闻领域	0.3988	计算所
柯汉	新闻领域	0.4391	自动化所

一、少数民族语言机器翻译现状

一些指标（续）

民文自然语言处理相关指标

内容	指标	来源
汉藏句对齐	准确率81.11%	于新等. 基于词典的汉藏句子对齐研究与实现. 中文信息学报, 2011.
汉蒙词对齐	准确率为81.4%，召回率为59.2%	雪艳. 汉蒙词语对齐及相关技术研究，内蒙古大学博士毕业论文, 2009.
蒙文词性标注	准确率为96.7%	张贯虹等. 融合形态特征的最大熵蒙古文词性标注模型. 计算机研究与发展, 2011.
蒙文形态切分	未登录词处理后，准确率96.94%	李文等. 基于短语统计机器翻译模型蒙古文形态切分. 中文信息学报, 2011.
维文分词	规则词准确率达到 95%	古丽拉·阿东别克. 维吾尔语词切分方法初探. 中文信息学报, 2004.
汉维词对齐	准确率为42.8%，召回率为37.3%	刘建明等. 基于统计机器翻译的汉维词对齐研究. 计算机应用与软件, 2011.
维文词性标注	准确率为79%	牛洪梅等. 现代维吾尔语的词性标注校对技术研究. 伊犁师范学院学报, 2007.

一、少数民族语言机器翻译现状

近几年少数民族语言机器翻译相关的项目支持

- **国家自然科学基金（7项）**
  - 重大1项：基于本体的多策略民汉机器翻译研究（2012）
  - 面上6项：
  - ①面向汉藏机器翻译的大规模双语语料库构建技术研究（2011）
  - ②融入语言学知识的汉蒙统计机器翻译研究（2011）
  - ③互译语言形态非对称的统计机器翻译模型构造方法研究（2011）
  - ④基于短语的维汉统计机器翻译关键技术的研究（2011）
  - ⑤纳西-汉语双语语料库构建与翻译方法研究（2012）
  - ⑥基于机器翻译的汉-维哈蒙多语种电子病历的研究（2012）
- **国家火炬计划（1项）**
  - LUOZANG-LZ828 藏汉英电子词典硬件产品（2010）
- **中国科学院西部行动高新技术项目（1项）**
  - 基于Linux的跨平台藏文信息处理系统（2003）
- **国家科技支撑计划（1项）**
  - 协同式多语言云翻译服务平台与应用（2012）

二、资源瓶颈

语料资源瓶颈

已建设语料（CWMT2011）

语言	规模	语料提供单位
汉蒙	6.8万句对/982,135蒙语词	内蒙古大学
汉藏	10万句对/约1,280,837藏语词	青海师范大学，厦门大学，西北民族大学
汉维	5万句对/1,091,903维语词	新疆大学
汉哈	5万句对/965,570哈语词	新疆大学
汉柯	5万句对/1,175,823柯语词	新疆大学

通用语料建设仍然不足

- 项目支持力度不足：2011-2012国家自然科学基金仅2项
- 语料使用机制不完善，导致语料建设者费力不讨好
- 已有语料扩展、更新缓慢

各领域专业词库/语料库稀缺

- 综合型人才稀缺
- 需要各领域专家和语言学专家合作

二、资源瓶颈

人才资源瓶颈

- **研究人员不足**
  - 已有研究人员数量仍显不足
  - 项目支持力度不够，今后的人才培养难以保证
- **人才资源分配不均衡**
  - 已有研究人员往往集中于有限的几个单位
  - 研究方向往往集中于某几个点，重复研究较多
- **各方面的人才合作仍需要深入**
  - 由单个单位申请的项目占绝大多数
  - 语言学专家和计算机专家的合作方式有待改进
  - 合作内容需要具体化，而不仅仅是作为参与者

三、技术瓶颈

译文质量的瓶颈

机器翻译离实用仍然相差甚远

- 中国数学家、语言学家周海中曾在论文《机器翻译五十年》中指出：要提高机译的质量，首先要解决的是语言本身问题而不是程序设计问题；单靠若干程序来做机译系统，肯定是无法提高机译质量的。在人类尚未明了“人脑是如何进行语言的模糊识别和逻辑判断”的情况下，机器翻译要想达到“信、达、雅”的程度是不可能的。这也是制约机器翻译质量提高的一大瓶颈。
- 美国发明家、未来学家雷科兹威尔最近在接受《赫芬顿邮报》采访时预言，到2029年机器翻译的质量将达到人工翻译的水平。对于这一论断，学术界还存在很多争议。
- 事实上，不论哪种方法，影响机器翻译发展的最大因素在于译文质量。就已有的成就来看，机器翻译的译文质量离实用仍相差甚远。

三、技术瓶颈

翻译引擎的瓶颈

自动翻译引擎技术短期内难以有大的突破

- 机器翻译运用语言学原理，让机器自动识别语法，然后调用存储的词库，自动进行对应翻译，对于语法、词法、句法的不规则性变化出现翻译错误更是在所难免。
- 少数民族语言由于其丰富的语言特征，导致汉民语言互译中出现的问题更加多样化。
- 实际上，机器翻译引擎技术的改进一直是个世界性的课题。要让计算机翻译自然语言，需要建立复杂的数学模型。以目前的人工智能技术来看，自动翻译引擎的技术已经到了一个瓶颈，短期内很难有大的突破。

四、实验室工作简介

基本情况

- 研究人员：高级职称2名、中级职称8人、初级职称1人
- 培养人才：硕/博研究生40余人
- 研究内容：
  - Web平行语料挖掘研究
  - 语言处理技术研究
  - 统计机器翻译模型构造方法研究
  - 源语言重排序方法研究
- 目前正在承担的课题：
  - 国家自然科学基金

    “互译语言形态非对称的统计机器翻译模型构造方法研究”
  - 国家科技支撑计划

    “民族语言文字信息应用管理与服务体系研究”
- 对外合作：内蒙古大学、新疆大学、西藏民族学院等

四、实验室工作简介

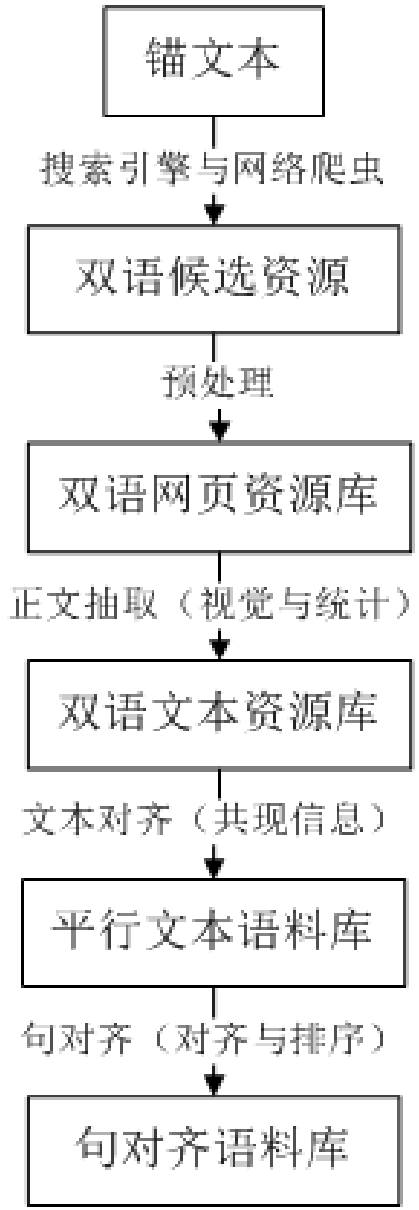
Web平行语料挖掘研究

- 锚文本->双语网页资源库：
  搜索引擎、网络爬虫、预处理

- 双语网页资源库->双语文本资源库：
  正文抽取（视觉与统计）

- 双语文本资源库->平行文本语料库：
  文本对齐（共现信息）

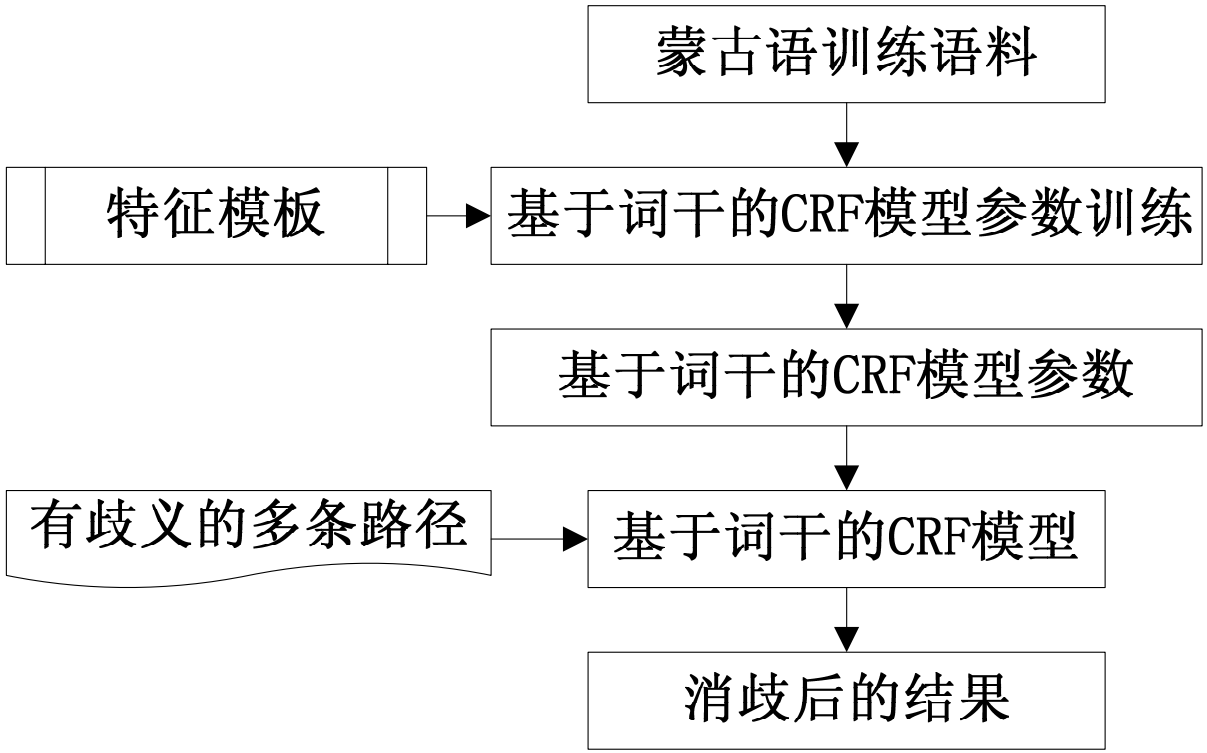
- 平行文本语料库->句对齐语料库：
  句对齐（对齐与排序）



四、实验室工作简介

蒙古文形态分析技术研究

- 基于转换方法或**CRFs**进行蒙古文词性标注以及消歧

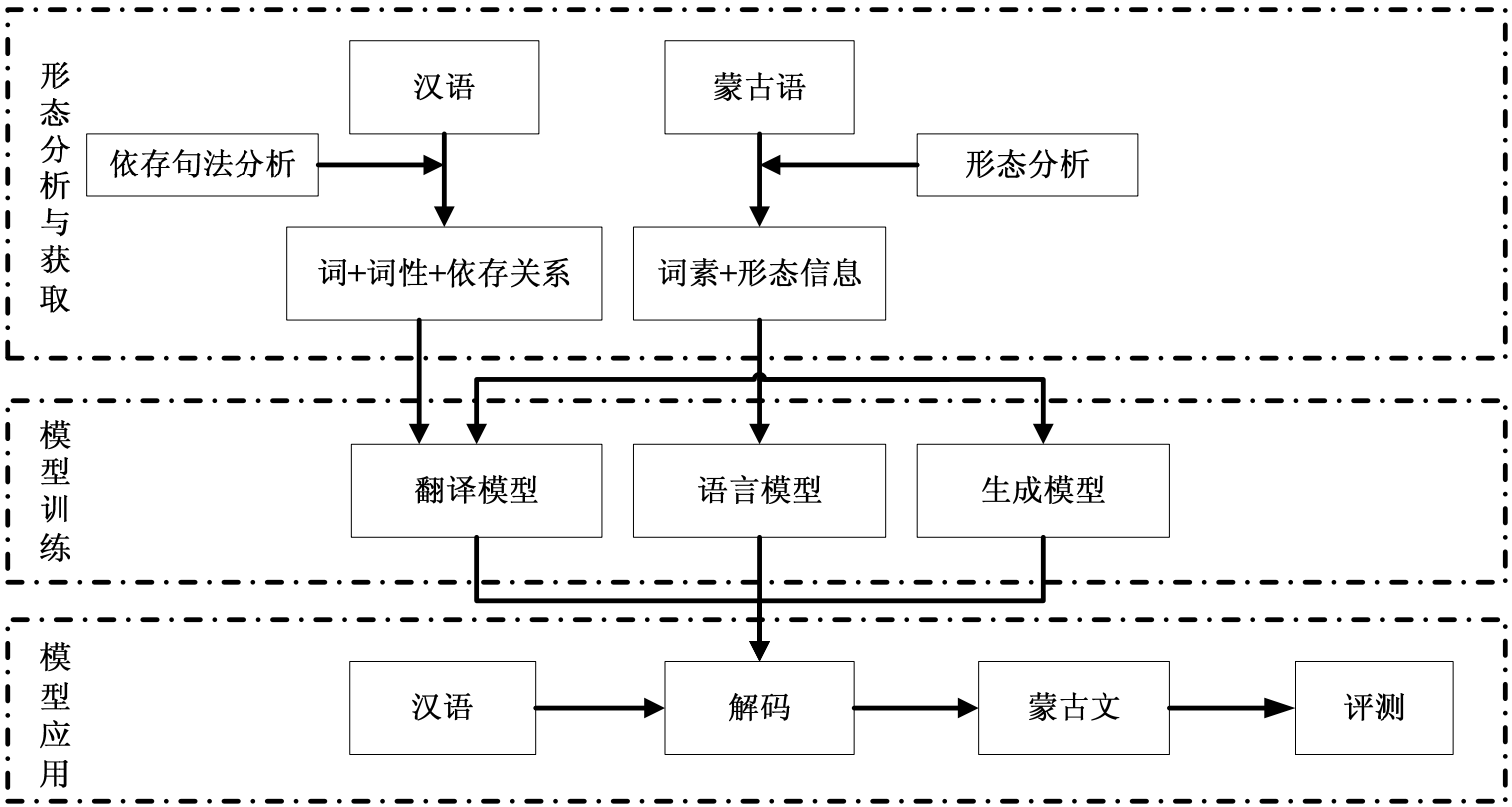


基于词干的CRF模型参数训练及消歧处理

四、实验室工作简介

形态学方法与因子化翻译模型研究

- 提高形态非对称语言互译质量的有效模型
- 形态信息以因子的形式进行融合
- 多个翻译模型、语言模型、生成模型的有机结合

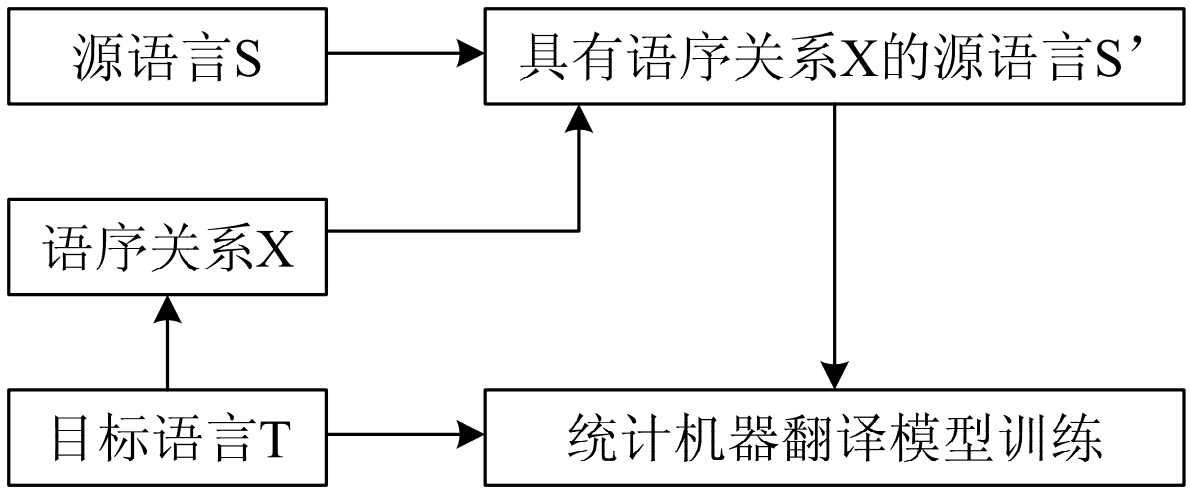


因子化翻译模型总体框架

四、实验室工作简介

源端重排序技术研究

- 分析目标语言T的语序关系X
- 利用X对源语言S进行重排序，得到S’
- 在S’与T的基础上进行统计机器翻译构型的构造

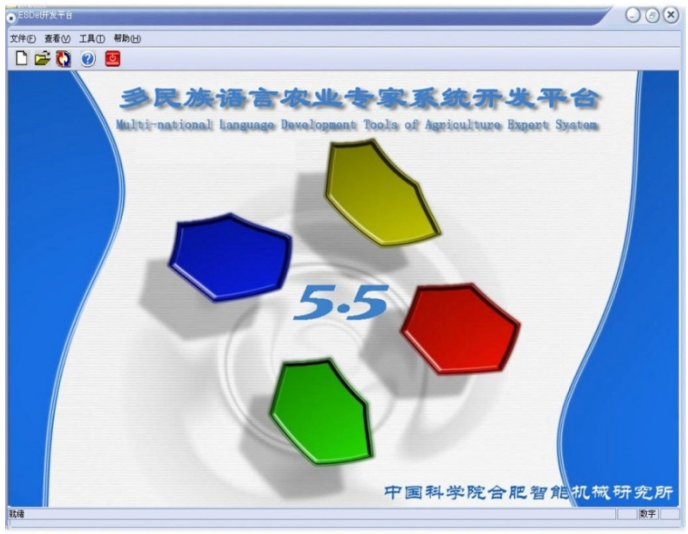


源端重排序方法示意图

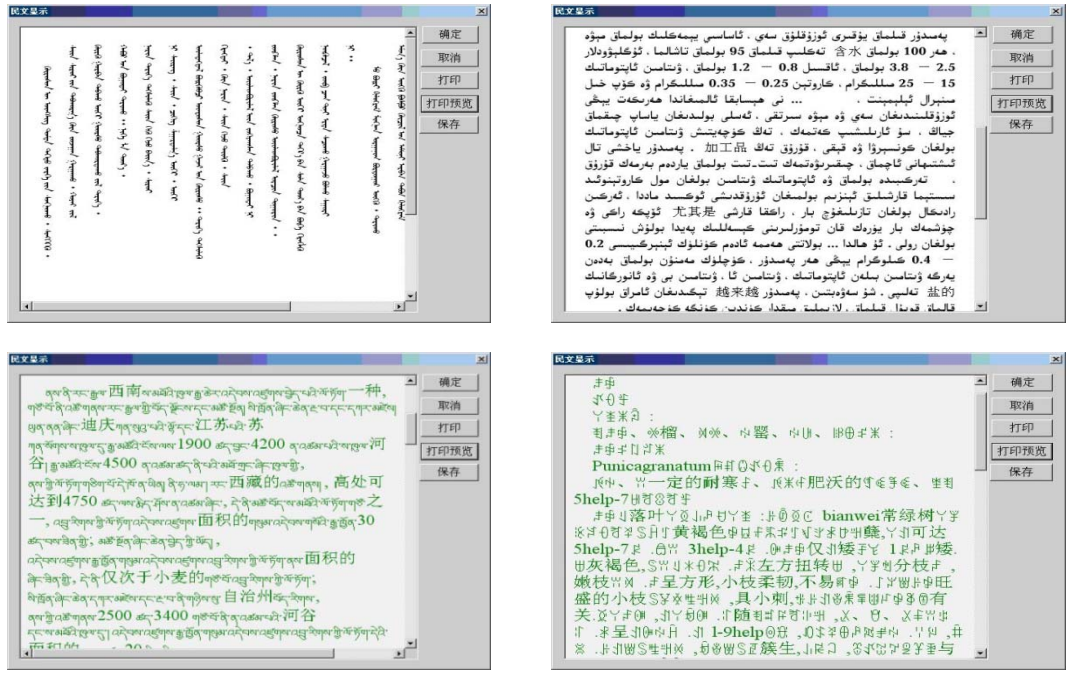
四、实验室工作简介

■多民族语言农业信息处理平台（单机版）

- •以机器翻译技术为基础的多民族语言农业信息处理平台
- •开发了小麦、玉米、马铃薯等多个汉民双语农业专家系统
- •在民族地区开展实验性应用，取得了明显的社会经济效益



多民族语言农业专家系统开发平台



蒙、维、彝、藏等民文的翻译与显示

四、实验室工作简介

多民族语言农业信息处理平台（网络版）

网络版多民族语言专家系统首页

内蒙古地区的四个专家系统

推理菜单对应的民文翻译与显示

推理结果对应的民文翻译与显示

五、总结：一些可借鉴的工作

- 在线翻译技术
  - Google翻译目前可提供63种主要语言之间的在线翻译。其主要采用统计翻译模型，通过海量统计数据来提高翻译精确度。
- 实时语音翻译技术
  - 2011年11月，Google推出了一款手机翻译软件。支持包括汉语在内的14个语种。用户几乎能实时听到他们的源语言被翻译成目标语言；而通话对方的语言也会被翻译成该用户的母语。
- 其他新思路
  - 南加州大学提出把英语视为一种初始语言，而需要翻译的外语视为类似一种加密后的高级文字，通过译码破译，把外语“破解”，变成英语。

请各位专家批评指正！
谢谢！