关注语言本身难题，推进民族语言翻译

姜文斌 吕雅娟

中国科学院计算技术研究所
自然语言处理研究组

第八届全国机器翻译研讨会
西安，2012年9月

两个维度

- 资源短缺问题
- 语言本身难题
  - 拼音语言的拼写校对问题（蒙语，维语）
  - 分词和分句问题（藏语）
  - 黏着语的形态分析问题（蒙语，维语）

大纲

- 拼写校对与翻译
- 分句/分词与翻译
- 形态分析与翻译

拼写校对

- 拼音语言普遍存在拼写错误问题
  - 在保证发音基本一致基础上的灵活拼写
  - 不影响人类阅读，但使得机器无法识别
  - 为后续NLP任务如机器翻译带来未登录词问题
- 看似简单问题，未得到足够重视
  - 查词典，根据编辑距离寻找最佳候选
  - 寻找多候选，借助简单的统计模型排歧

拼写校对在MT

- 视而不见
  - 假定没有拼写错误问题

词语对齐，规则抽取，翻译解码各阶段的未登录词问题

- 使用所有可能的候选
  - 生成词语正确拼写的候选列表
  - 借助Lattice翻译模型

仅在翻译解码阶段有容错效果

可能的方案

- 语言学规则加统计排歧的拼写校对
  - 语言学规则枚举可能的正确形态候选
  - 基于词干/词缀、音节或词的统计排歧

改进词语对齐、规则抽取和翻译解码阶段

- 拼写校对系统输出候选词图结构

进一步改进翻译解码阶段

大纲

- 拼写校对与翻译
- 分句/分词与翻译
- 形态分析与翻译

分句/分词

- 藏语同时存在分句和分词的问题

- 问题本身已得到较好的解决
  - 藏语分词（F%）：＞96
  - 藏语分句（F%）：~98
  - 领域适应任何，对于机器翻译是否够用?

分句/分词在MT

- 分句/分词是机器翻译各流程的基础

- 单一结果 / 词图结果
  - 分句通常使用单一结果
  - 词语对齐和规则抽取通常使用单一分词结果
  - 翻译解码可能使用词图分词结果

可能的方案

- 分词系统输出高质量且稀疏的词图
  - 基于词图的词语对齐?
  - 基于词图的翻译规则抽取和概率估计?

- 分句系统如何做适应机器翻译的改进?

大纲

- 拼写校对与翻译
- 分句/分词与翻译
- 形态分析与翻译

形态分析

- 黏着语词汇形态变化丰富，须形态分析
  - 通过词干和词缀的缀接构成更大的表意单位
  - 缀接可能性多，数据稀疏问题严重

- 问题本身解决尚好，但仍存在严重问题
  - 维语（P%）：~93，蒙语（P%）：~95
  - 形态分析和拼写校对通常纠结在一起

分析器实际情景下的性能可能较差

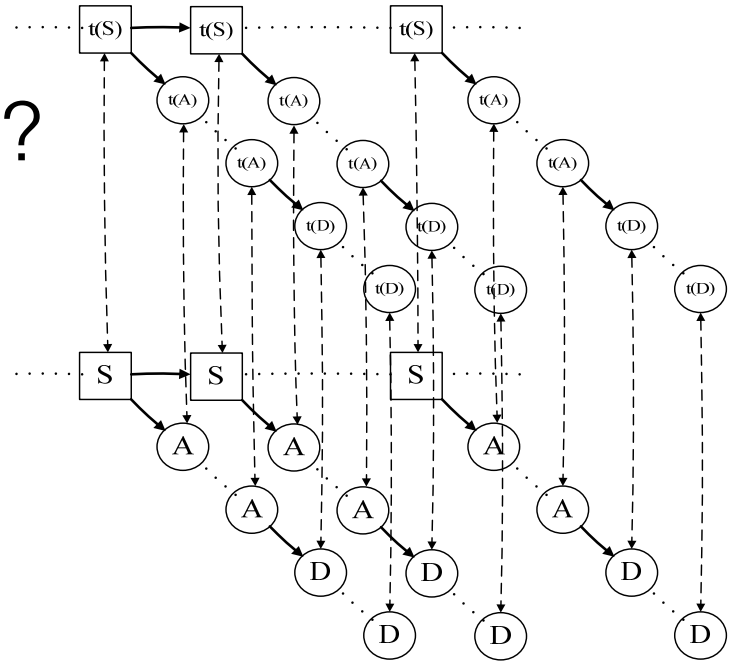
形态分析在MT

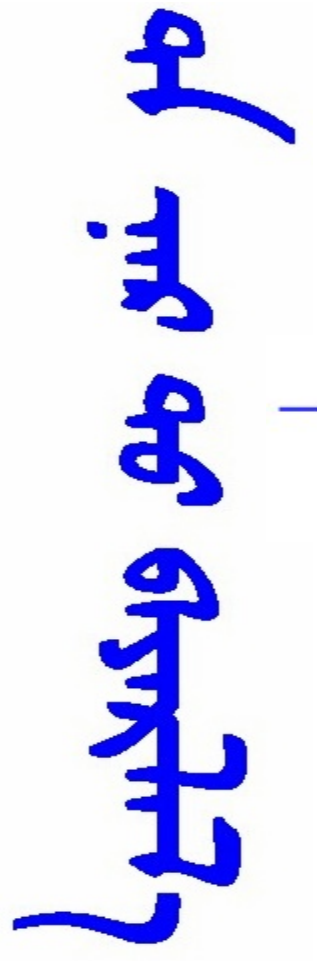
- 不考虑形态分析
  - 将词干和词缀构成的复杂词汇仍视为普通词
  - 语料规模越大，该做法越能被接受
- 借助形态分析
  - 词和词干/词缀多种粒度进行对齐融合
  - 词和词干/词缀多种粒度生成翻译词图
  - 区别词干和词缀信息进行翻译规则选择

用过去问题的思想考虑新的问题?

可能的方案

- 词干和词缀的缀接更像是句法构造过程
- 基于图状结构的词语对齐和规则抽取?
- 基于图状结构的翻译建模?





كۆپچىلىككە رەھمەت

谢谢大家！