中文信息学会成立三十周年学术会议

机器翻译技术的进展与展望

刘群、王海峰、王惠临、宗成庆、赵铁军
史晓东、朱靖波、陈家俊、张民

2011-12

内容提要

	研究	应用
国内外现状分析	**进步与不足** **机器翻译研究范式的变化**	应用范围的延伸 互联网翻译 计算机辅助翻译 口语翻译
急待解决的核心问题	深层次语言知识的运用 复杂形态语言翻译建模 资源缺乏语言机器翻译 机器翻译的自动评价	翻译质量还不够高 翻译结果还不够可信 系统使用还不够方便 语种和领域的支持还不够多
发展前景与趋势展望	基于复杂特征的翻译建模 基于结构的语言建模 混合机器翻译知识的主动学习 引入结构的机器翻译自动评价	在应用深度上延伸 在应用广度上扩展

国内外现状分析：研究

进步与不足

机器翻译研究范式的变化

已经取得的进步

- 近十几年来，机器翻译取得了巨大的进步
- 统计机器翻译取得巨大成功，从基于词的模型发展到了基于短语的模型和基于句法的模型
- 机器翻译的统计方法和规则方法走向融合
- 机器翻译系统开发效率大为提高：数年→数周
- 应用范围大大拓展：Google翻译支持几十种语言
- 翻译质量也有了明显上升，已经成为日常工具

依然面临的困难

- 翻译质量仍然不够理想（婴儿期）
- 需要大规模训练语料库：数据稀疏问题
- 需要与应用场合相近的语料：领域适应性
- 语言形态的复杂性还没有好的处理办法
- 语言之间差异性很大时翻译质量不理想

国内外现状分析：研究

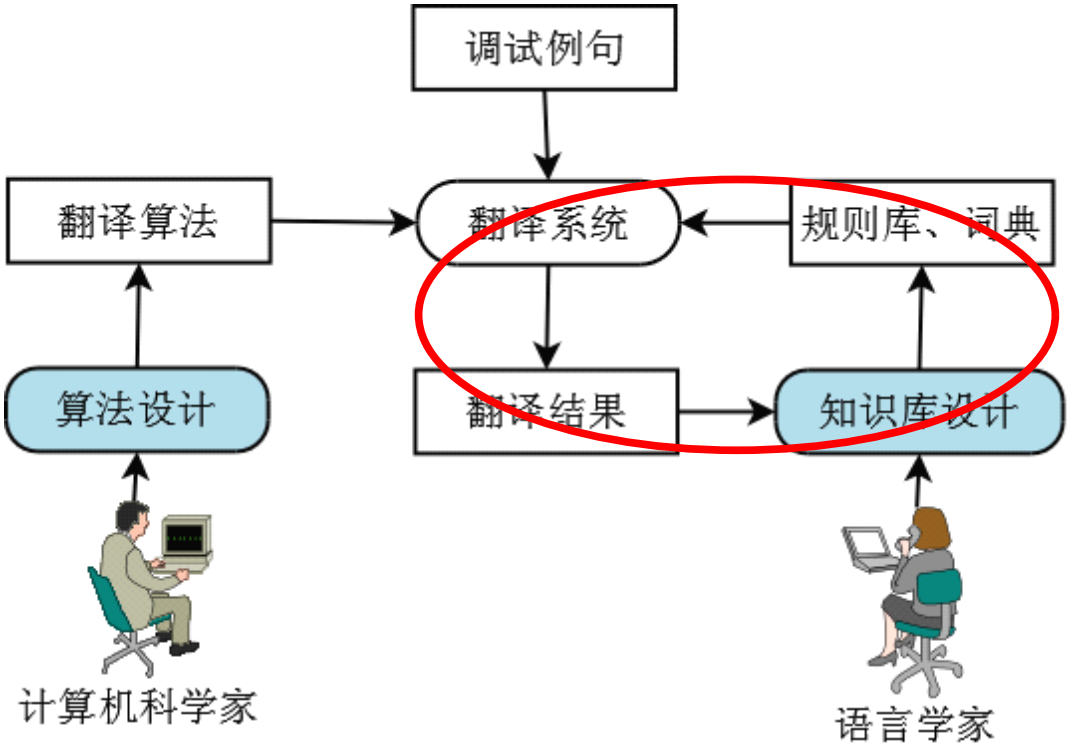
进步与不足

机器翻译研究范式的变化

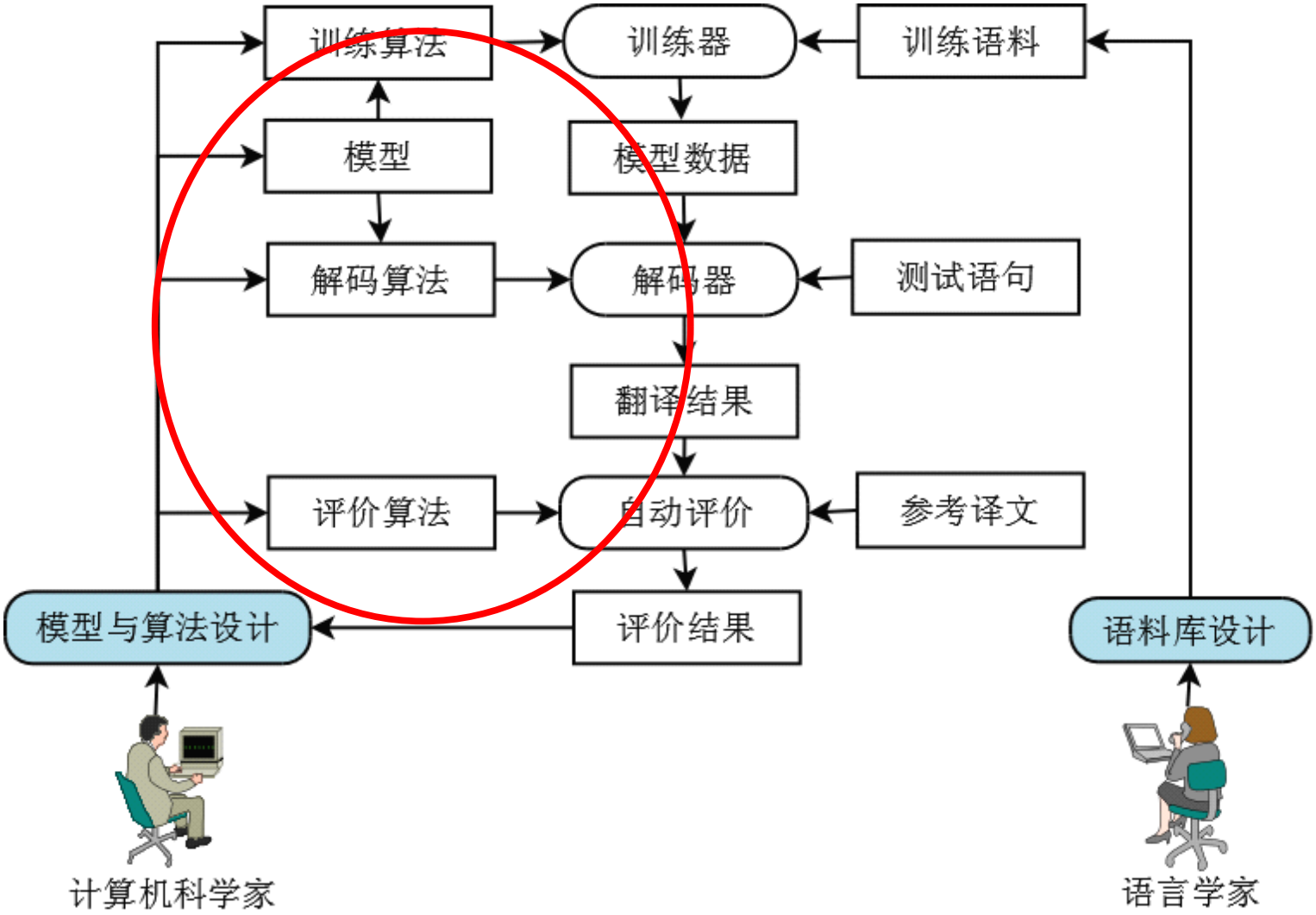
RBMT vs. SMT

	RBMT	SMT
知识表示	规则	模型参数
知识获取	人工构造	自动学习

基于规则的机器翻译研究范式



统计机器翻译研究范式


训练算法
训练器
训练语料
模型
模型数据
解码算法
解码器
测试语句
翻译结果
评价算法
自动评价
参考译文
模型与算法设计
评价结果
语料库设计
计算机科学家
语言学家

研究范式的变化

算法与规则的分离

分析、转换、生成的分离

模型与搜索的分离

语言专家为中心→计算机专家为中心

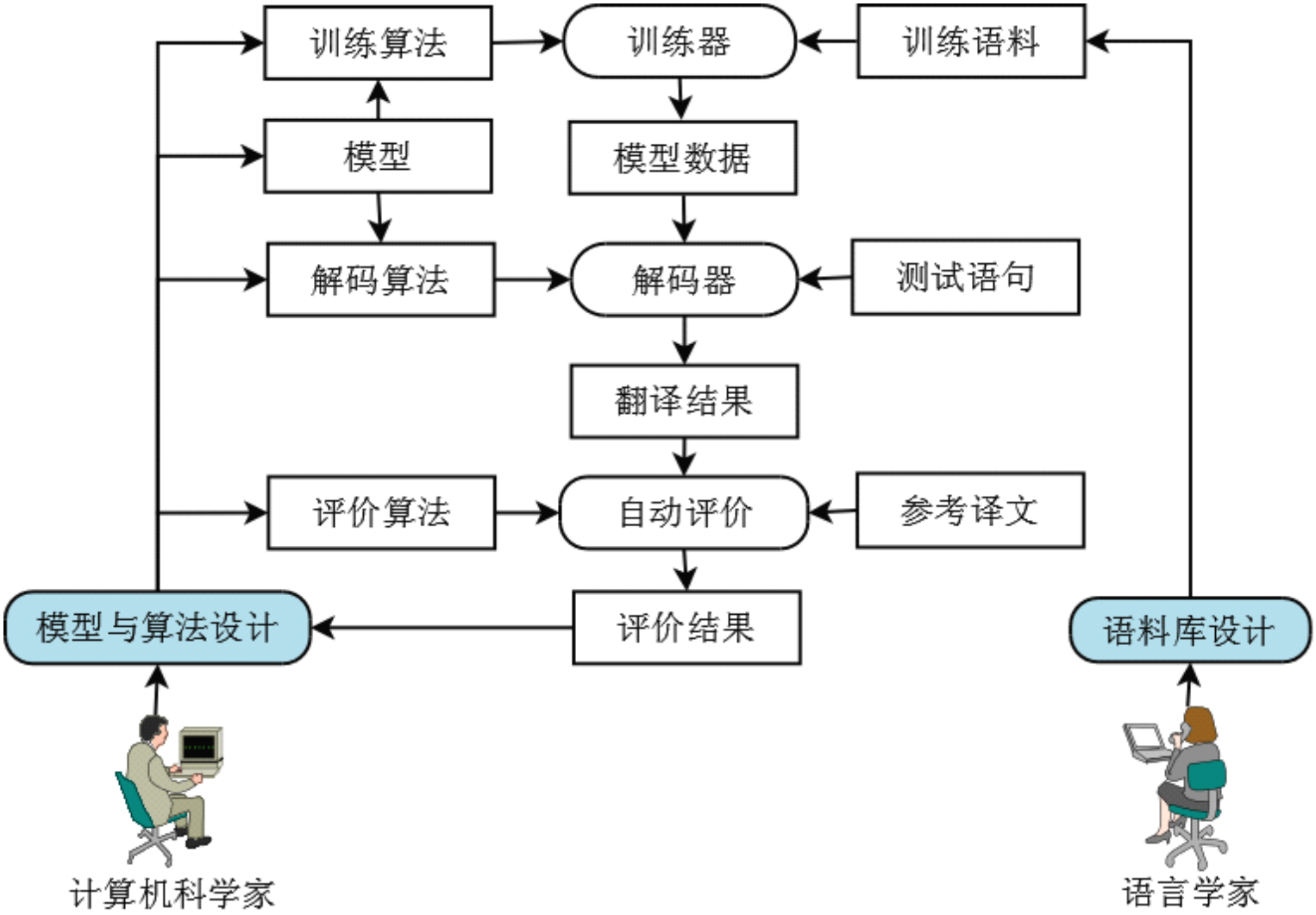
RBMT和SMT的融合

- 知识表示
  - 统计模型已经成为必须
    - 语言的不确定性是天然存在的，规则永远不足以刻划语言的不确定性
    - 统计为处理语言的不确定性提供了不一定合理但有效的解决办法
    - 统计模型大大扩展了搜索空间
  - 规则表示形式已经在统计模型中普遍采用
- 知识获取
  - 训练数据充分时：统计学习非常有效
  - 训练数据不足时：统计方法的优势无法体现
- 发展趋势是殊途同归，作用互补，二者缺一不可，无需再争论谁更重要，事物的发展是螺旋式的，无论统计方法或规则方法都没有可能独自“复兴”。

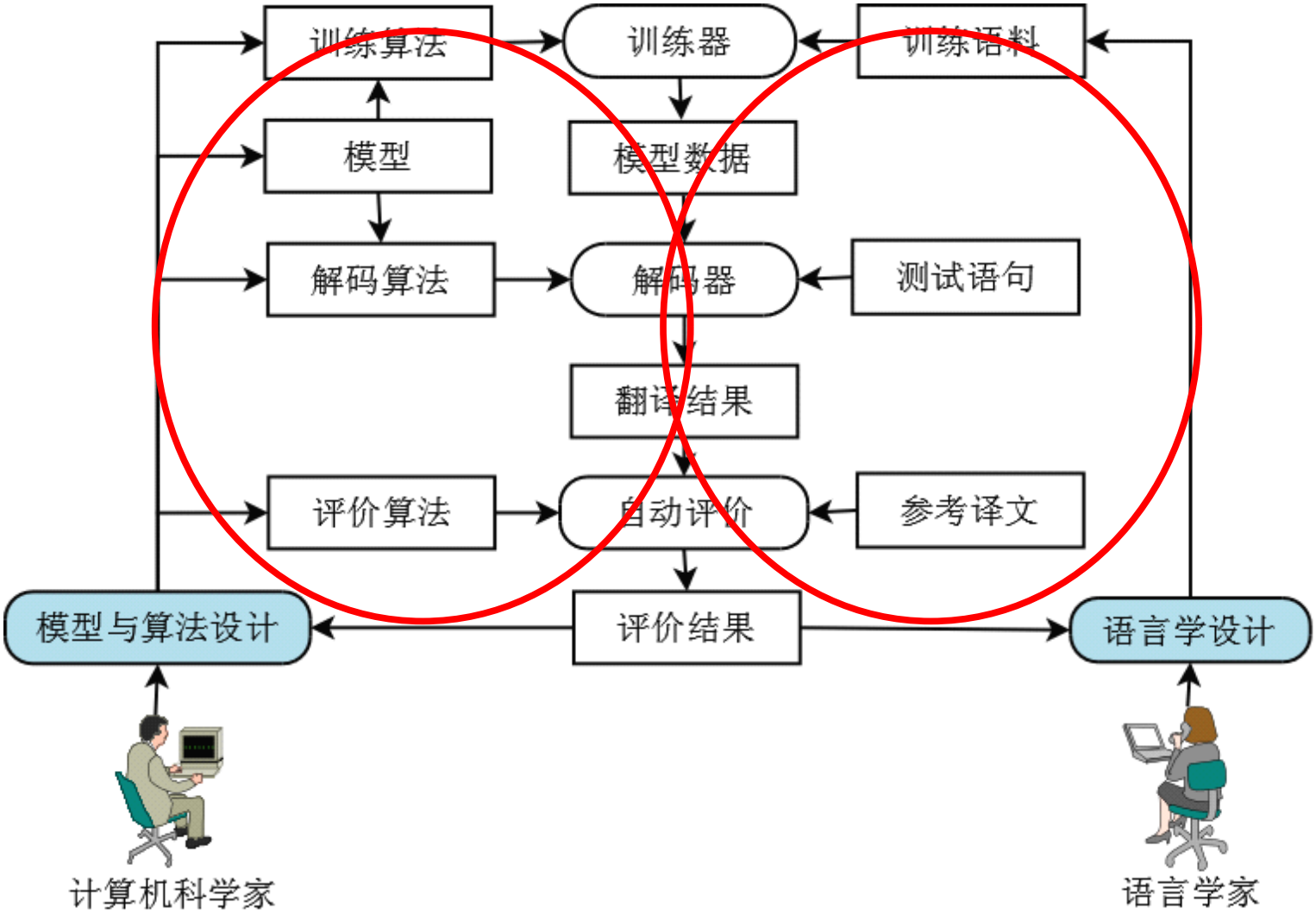
语言知识 & 机器学习

- 规则方法与统计方法的融合是必然的趋势
- 改进的方法不是片面强调某一方面，而是两方面都要加强
  - 需要引入更深层的语言知识
  - 需要采用更强大的机器学习方法
- 从研究范式上看，有必要再引入语言学设计的迭代循环

规则与统计结合的新研究范式



规则与统计结合的新研究范式



内容提要

	研究	应用
国内外现状分析	进步与不足 机器翻译研究范式的变化	应用范围的延伸 互联网翻译 计算机辅助翻译 口语翻译
急待解决的核心问题	**深层次语言知识的运用** **复杂形态语言翻译建模** **资源缺乏语言机器翻译** **机器翻译的自动评价**	翻译质量还不够高 翻译结果还不够可信 系统使用还不够方便 语种和领域的支持还不够多
发展前景与趋势展望	基于复杂特征的翻译建模 基于结构的语言建模 混合机器翻译知识的主动学习 引入结构的机器翻译自动评价	在应用深度上延伸 在应用广度上扩展

急待解决的核心问题：研究

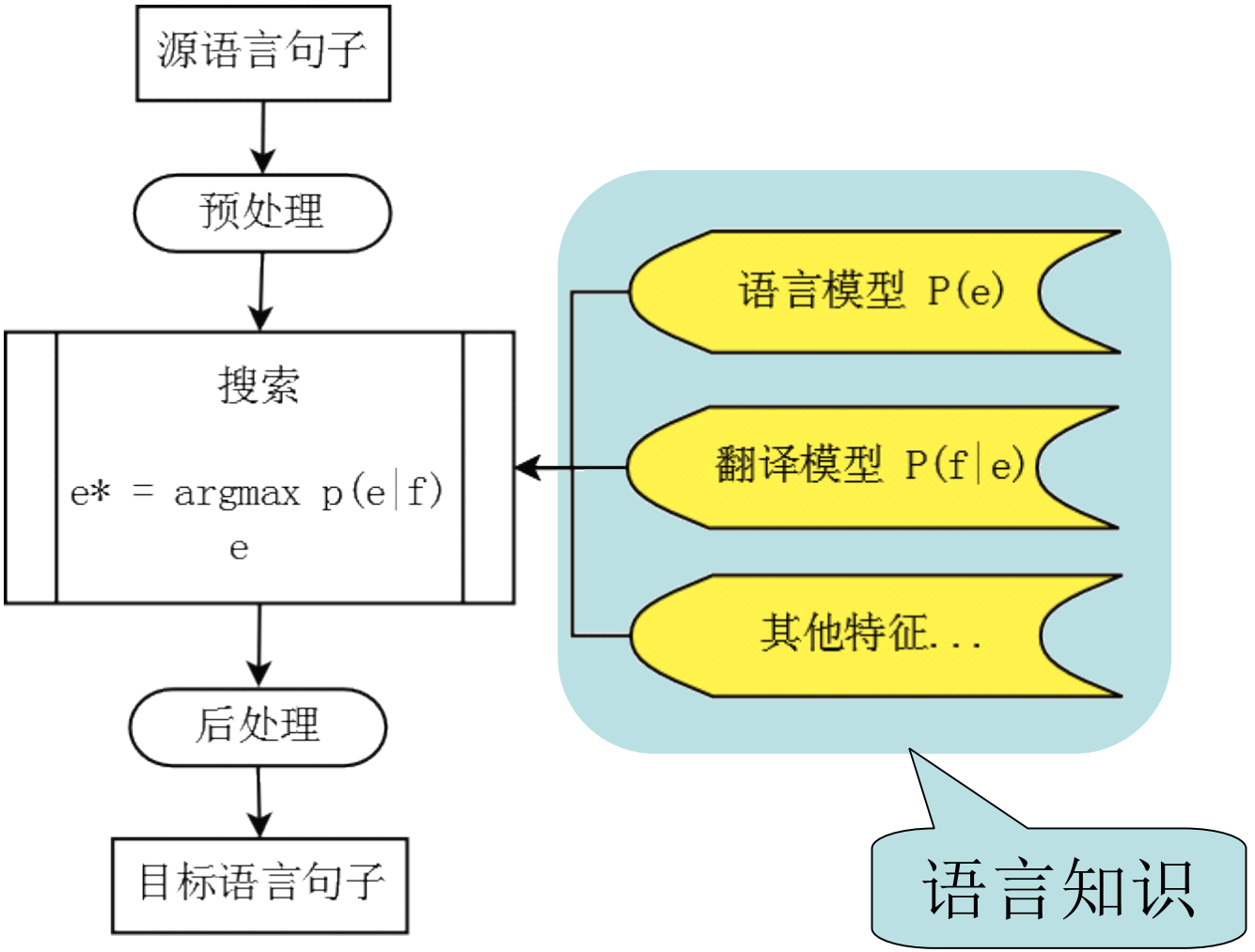
深层次语言知识的运用

复杂形态语言翻译建模

资源缺乏语言机器翻译

机器翻译自动评价

机器翻译的基本流程


源语言句子
预处理
搜索
e* = argmax p(e|f)
e
后处理
目标语言句子
语言模型 P(e)
翻译模型 P(f|e)
其他特征...
语言知识

语言模型 & 翻译模型

- 语言模型
  - 目前最成功的语言模型仍然是n-gram模型
  - 依存语言模型作为n-gram模型的补充可以起到一定的效果，但理论上仍不完善，作用也很有限
  - 语言模型的研究远不充分
- 翻译模型
  - 目前成功的模型使用的语言知识仅限于句法树，更复杂的句法知识、语义知识和篇章知识都没有引入

对机器翻译有用的语言知识

- 语音层面：韵律知识
- 词语层面：屈折变化、以字构词、构词模式、汉语重叠、离合、缩合
- 句法层面：句法结构、句法约束、位移
- 语义层面：知识本体、语义偏向、语义角色
- 篇章层面：指代消解、话题结构
- 语用层面：情境语义、情感

一些例子

- 进出口、男女生、二十三四个
- 我摔断了腿 → I broke my leg
  他摔断了腿 → he broke his leg
- 洗了个热水澡
- 我昨天上街，(*)看见一个人，(*)穿着军大衣，(*)喝得醉醺醺的，(*)嘴里(*)还唱着歌，(*)调都(*)找不着(*)了。
- 儿子买了辆自行车，没骑两天，车就坏了。

深层语言知识应用的难点

- 实际语料库中出现较少，对BLEU值贡献太小
- 缺乏语言资源
- 现有统计模型无法支持
- ……

急待解决的核心问题：研究

深层次语言知识的运用

复杂形态语言翻译建模

资源缺乏语言机器翻译

机器翻译自动评价

复杂形态语言机器翻译问题

- 粘着语的形态变化都非常复杂，蒙古语和维吾尔语的动词的形态变化理论上都可以达到1000种以上
- 部分屈折语也有很复杂的形态变化，某些屈折语的动词形态变化可达上百种
- 现在的统计机器翻译中，通常不对形态变化做处理，每种不同形态的词语都当成不同的符号，这对于形态变化简单的汉语和英语是比较方便的，但对于复杂形态语言则会带来严重问题
- 对于复杂形态语言在统计机器翻译中的处理，目前还没有理想的处理方法

复杂形态语言机器翻译问题

- 复杂形态语言机器翻译的主要问题
  - 数据稀疏：由于词的形态变化太多，导致很多词语的变化形式在训练语料库中没有出现过
  - 语法差异：形态丰富语言中，通过形态变化表示的语法信息，在形态简单语言中通常通过虚词或者词序变化来表示，二者差异非常大，这种差异目前的统计翻译模型很难准确刻划
  - 形态生成：形态简单语言翻译到形态复杂语言，所需的形态信息往往不足，需要补充

急待解决的核心问题：研究

深层次语言知识的运用

复杂形态语言翻译建模

资源缺乏语言机器翻译

机器翻译自动评价

资源缺乏语言的机器翻译

- 统计方法需要大规模平行语料库支持，而世界上绝大部分语种之间并不存在大规模平行语料库
- 如果存在另外一种语言，跟两种语言都有较大规模平行语料库，可以采用基于中介语的机器翻译方法，但效果会有所降低

资源缺乏语言的机器翻译

- 在完全没有双语语料库的情况下，如何构造一个机器翻译系统最为高效？
  - RBMT：人工构造规则和词典
    - 需要有经验专家
    - 词典和规则库总体规模较小
  - SMT：人工构造双语语料库
    - 只需要普通翻译人员
    - 语料库规模要求非常大
  - 两种方法结合？如何做？
- 基于非并行的大规模单语语料的机器翻译

资源缺乏语言的机器翻译

- 资源缺乏的另一个方面是领域资源缺乏
- 有些语言对之间虽然存在一些双语语料库，但领域过于狭窄和单一，根据这些语料构造的机器翻译系统几乎无法应用于其他领域，领域的自适应也是一个很难解决的问题
- 目前的机器翻译模型本身还不刻划模型分布随着领域变化发生迁移的现象

急待解决的核心问题：研究

深层次语言知识的运用

复杂形态语言翻译建模

资源缺乏语言机器翻译

机器翻译自动评价

机器翻译自动评价

- 目前的机器翻译自动评价，主要还是采用基于n-gram匹配的指标，如BLEU值
- 虽然出现了一些新的评价方法，但优势并不明显，没有得到广泛接受
- 在美国Gale项目中，已经采用了基于人工编辑距离的评价方法，人力成本很高，并不具有在研究工作中普遍推广使用的价值
- 随着一些机器翻译水平的逐步提高，机器翻译自动评价技术越来越难以满足机器翻译研究的需要，但到目前为止还没有可接受的替代方案

内容提要

	研究	应用
国内外现状分析	进步与不足 机器翻译研究范式的变化	应用范围的延伸 互联网翻译 计算机辅助翻译 口语翻译
急待解决的核心问题	深层次语言知识的运用 复杂形态语言翻译建模 资源缺乏语言机器翻译 机器翻译的自动评价	翻译质量还不够高 翻译结果还不够可信 系统使用还不够方便 语种和领域的支持还不够多
发展前景与趋势展望	**基于复杂特征的翻译建模** **基于结构的语言建模** **混合机器翻译知识的主动学习** **引入结构的机器翻译自动评价**	在应用深度上延伸 在应用广度上扩展

发展前景与趋势展望：研究

基于复杂特征的翻译建模

基于结构的语言建模

混合机器翻译知识的主动学习

引入结构的机器翻译自动评价

基于复杂特征的翻译建模

- 语法理论的研究中，复杂特征已被证明是一种强有力的描述手段，基于单一标记的语法理论描述能力非常有限。
- 目前所有的统计机器翻译模型，都还是基于单一标记的，如句法树。统计翻译模型无法建立在基于复杂特征描述的语言学理论上，语言学特征只能作为机器学习算法中海量特征中的个别特征，作用非常有限。

基于复杂特征的翻译建模

- 在复杂特征上建立翻译模型的好处：
  - 可以方便刻划复杂的形态变化
  - 可以方便地表示语言成分之间的句法约束
    - 特征可以在句法成分之间继承和传递，这是语言中非常常见的现象
  - 可以较方便的处理语言中的位移现象
- 在复杂特征上建立翻译模型的难处：
  - 需要用到更加复杂数学工具和机器学习方法
  - 缺乏具有复杂特征标记的语料库

发展前景与趋势展望：研究

基于复杂特征的翻译建模

基于结构的语言建模

混合机器翻译知识的主动学习

引入结构的机器翻译自动评价

基于结构的语言模型

- 目前机器翻译中最有效的语言模型还是n-gram模型，由于n-gram模型完全不考虑结构，显然是不合理的。
- n-gram语言模型最大的问题是无法使生成的目标语言尽量符合语法的约束
- 目前研究人员已经开始更多关注这一领域，但还没有出现被普遍接受的更有效的方法
- 沈李斌工作表面依存语言模型对n-gram模型可以起到较好的补充，但单独使用效果并不好，理论上也不完善

基于结构的语言模型

- 理想的结构化语言模型：
  - 基于某种结构形式，如短语结构树、依存树，或者任何可以反映句子中长距离约束的语言结构形式
  - 应该可以利用大规模单语语料库进行训练，而不是只能依赖于少量的句法树库之类的标注语料库进行训练
  - 应该可以跟基于句法的解码算法有效结合，使用时的时间复杂度应该在可接受范围内

发展前景与趋势展望：研究

基于复杂特征的翻译建模

基于结构的语言建模

混合机器翻译知识的主动学习

引入结构的机器翻译自动评价

混合机器翻译知识的主动学习

- 在没有双语语料库的情况下，如何花费最小的人力成本有效构造一个机器翻译系统？
- 双语语料库设计：成本极高，应设法设计源语言句子集合用于翻译成双语语料库，使得花费最小代价获得最好的系统性能
- 主动学习算法：通过算法寻找最有必要进行人工标注（翻译）的语料库进行标注

混合机器翻译知识的主动学习

- 问题：
  - 人工标注的语料库是否会带来统计分布的偏置？
  - 引入部分人工编写规则是否会比完全人工翻译语料库更加有效？
  - 人工编写的规则如何与统计学习获得的模型知识融合？

发展前景与趋势展望：研究

基于复杂特征的翻译建模

基于结构的语言建模

混合机器翻译知识的主动学习

引入结构的机器翻译自动评价

引入结构的机器翻译自动评价

- 跟语言模型一样，机器翻译的自动评价也应引入结构信息
- 这种结构信息不应该过多依赖于语言学知识
- 自动评价方法如果不能得到改进，将严重影响机器翻译研究的进展

内容提要

	研究	应用
国内外现状分析	进步与不足 机器翻译研究范式的变化	**应用范围的延伸** **互联网翻译** **计算机辅助翻译** **口语翻译**
急待解决的核心问题	深层次语言知识的运用 复杂形态语言翻译建模 资源缺乏语言机器翻译 机器翻译的自动评价	翻译质量还不够高 翻译结果还不够可信 系统使用还不够方便 语种和领域的支持还不够多
发展前景与趋势展望	基于复杂特征的翻译建模 基于结构的语言建模 混合机器翻译知识的主动学习 引入结构的机器翻译自动评价	在应用深度上延伸 在应用广度上扩展

国内外现状分析：应用

应用范围的延伸

互联网翻译

计算机辅助翻译

口语翻译

国内外现状分析：应用

- 应用范围的延伸
  - Google支持60多种语言
  - 跨国恋爱
  - 商品交易
- 互联网翻译已经成为网民最常用的工具之一
  - Google
  - 必应
  - 百度
  - 有道

国内外现状分析：应用

- 计算机辅助翻译市场上取得巨大进步
  - 已经形成产业
  - 成为专业翻译人员不可或缺的工具
  - 在高校中设立专门的计算机辅助翻译专业：香港中文大学、北京大学
  - 典型系统：TRADOS、格微软件
- 口语翻译走出实验室

内容提要

	研究	应用
国内外现状分析	进步与不足 机器翻译研究范式的变化	应用范围的延伸 互联网翻译 计算机辅助翻译 口语翻译
急待解决的核心问题	深层次语言知识的运用 复杂形态语言翻译建模 资源缺乏语言机器翻译 机器翻译的自动评价	**翻译质量还不够高** **翻译结果还不够可信** **系统使用还不够方便** **语种和领域的支持还不够多**
发展前景与趋势展望	基于复杂特征的翻译建模 基于结构的语言建模 混合机器翻译知识的主动学习 引入结构的机器翻译自动评价	在应用深度上延伸 在应用广度上扩展

急待解决的核心问题：应用

翻译质量还不够高

翻译结果还不够可信

系统使用还不够方便

语种和领域的支持还不够多

翻译质量还不够高

- 翻译质量远没有达到理想的水平
- 汉语和英语、汉语和各少数民族语言的机器翻译与国际上英语、法语、阿拉伯语之间的机器翻译相比还有较大差距
- 翻译质量问题严重影响了机器翻译的应用范围

翻译结果还不够可信

- 在翻译质量不够理想的情况下，很多用户希望能知道，哪些机器翻译结果是可信的，哪些是不可信的，这一点还很难做到

系统使用还不够方便

- 由于翻译质量还不够理想，要让机器翻译被用户接受，需要创造真正体贴用户的机器翻译使用模式
- 计算机辅助翻译：专业用户很难接受在机器翻译的结果上进行修改
- 跨语言检索：虽然Google已经提供，但用户仍然不习惯使用
- 口语翻译：使用仍然不是非常方便

语种和领域的支持还不够多

- Google提供了60多种语言的翻译支持，仍然难以覆盖很多的语种，如我国的一些少数民族语言，而且很多非英语语言之间的翻译质量也很糟糕
- 国内的多语种机器翻译研究还相当欠缺
- 机器翻译的领域覆盖面还很差，在相当多的专业领域和应用场景中，机器翻译还不能提供有效的服务

内容提要

	研究	应用
国内外现状分析	进步与不足 机器翻译研究范式的变化	应用范围的延伸 互联网翻译 计算机辅助翻译 口语翻译
急待解决的核心问题	深层次语言知识的运用 复杂形态语言翻译建模 资源缺乏语言机器翻译 机器翻译的自动评价	翻译质量还不够高 翻译结果还不够可信 系统使用还不够方便 语种和领域的支持还不够多
发展前景与趋势展望	基于复杂特征的翻译建模 基于结构的语言建模 混合机器翻译知识的主动学习 引入结构的机器翻译自动评价	**在应用深度上延伸** **在应用广度上扩展**

发展前景与趋势展望：应用

在应用深度上延伸

在应用广度上扩展

在应用深度上延伸

- 要深入研究用户的需求和使用习惯，让机器翻译真正满足用户的需求
- 社会计算、移动计算、云计算的浪潮，为机器翻译提供了前所未有的机遇
- 社会计算的启示：用户贡献数据
- 苹果Siri的启示：Siri的成功主要不在语音识别性能的提高，而在与手机各项功能的深度集成
- 让用户感觉不到的服务才是好的服务！

互联网语言特点

- 传统语言：书面语、口语
- 网络语言：词汇、语法、转义

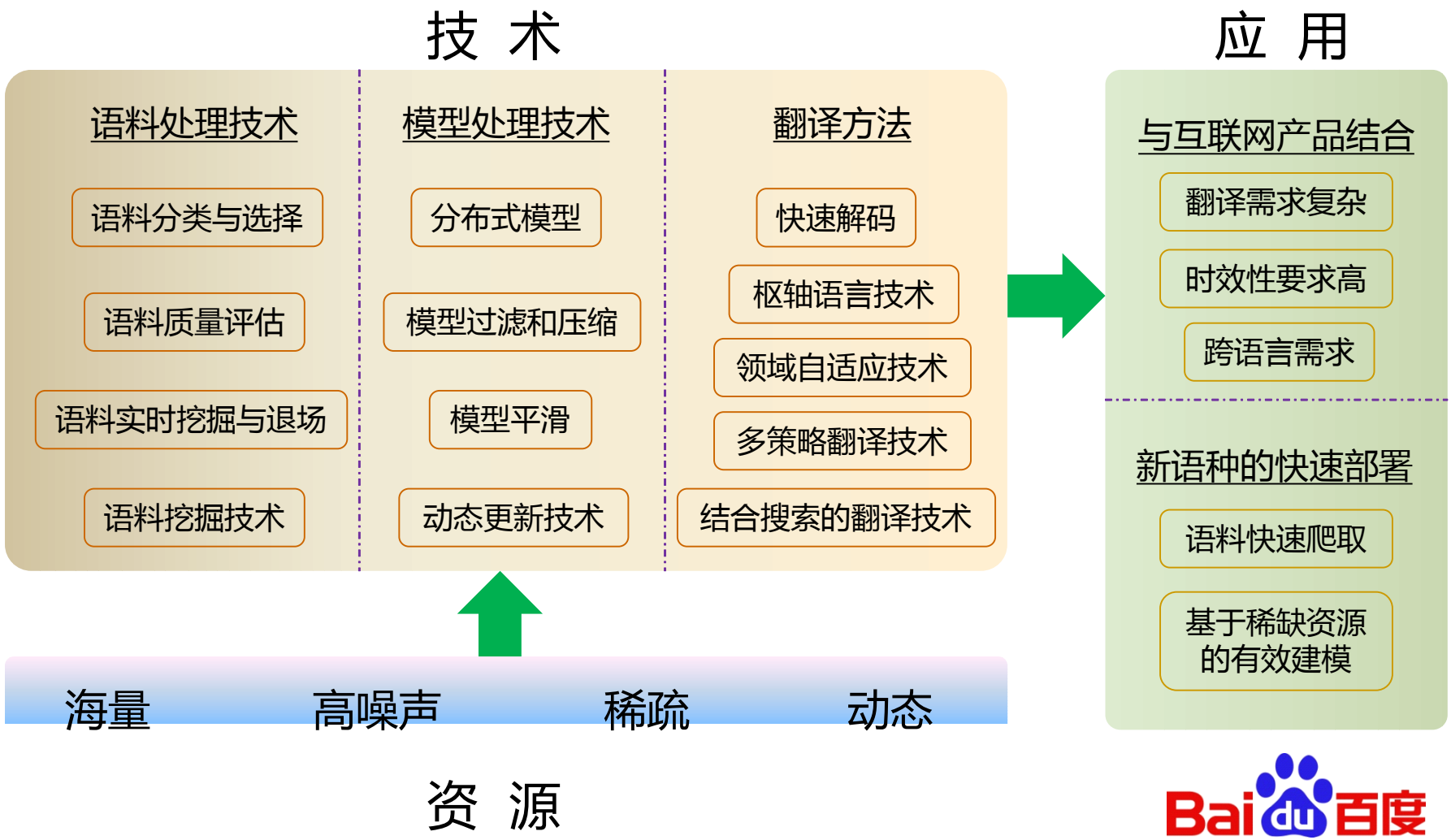
- 语言间的基本不对称：词汇、语法
- 语序：句子、段落
- 内容：内容增删、转义
- 编辑：如原文在正文，译文在HTML tag 中

- 网络语言变迁速度
- 新词、新表达方式
- 热点转移
- 时效性词退场

- 特定领域
- 特定语境
- 特定人群

Baidu 百度

机器翻译关键技术


技 术
语料处理技术
语料分类与选择
语料质量评估
语料实时挖掘与退场
语料挖掘技术
模型处理技术
分布式模型
模型过滤和压缩
模型平滑
动态更新技术
翻译方法
快速解码
枢轴语言技术
领域自适应技术
多策略翻译技术
结合搜索的翻译技术
应 用
与互联网产品结合
翻译需求复杂
时效性要求高
跨语言需求
新语种的快速部署
语料快速爬取
基于稀缺资源的有效建模
海量
高噪声
稀疏
动态
资 源

iPhone 4S 给我们的启发

- 口语仍然是人们相互交流最直接、最方便、最简单的方式
- 计算能力越来越强大的移动终端和无处不在的移动通讯为实现 Anywhere、Anytime、Any language 的无障碍通讯提供了可能
- iPhone 4S 的推出值得我们重新思考和定位口语翻译技术实用化的方式

iPhone 4S 给我们的启发

- 用户的使用模式
  - 传统的 speech-to-speech/ 结合网络对话的多模态人机交互？
  - 口语翻译与其它通讯方式结合的可能性

— 手机短信翻译

— 网络聊天翻译

— 拍照（关键词、短语）翻译

……

口语翻译实用化面临的挑战

- 任意时间、任意地点、任意说话人的口语语音识别问题
  - 噪声、准确率、鲁棒性、口音 ……
- 口语理解和翻译问题
  - 语音识别结果中的噪声
  - 口语本身的非规范性
  - 对话翻译的新方法
  - 多语言互译问题 ……
- 系统响应速度问题
  - 保证对话的实时性

在应用广度上扩展

- 机器翻译应该支持更多的语种
  - 国内少数民族语言
  - 周边国家主要语言
  - 国际经济贸易交流主要语言
- 机器翻译可以扩展到更多的应用领域
  - 旅游
  - 科技：论文（日本为例）、专利
  - 经贸：本地化、商务谈判、网上购物
  - 文化：国家重视（以Kevin Knight的工作为例）

总结

- 机器翻译近十余年来发生了深刻的变化
- 研究方面
  - 机器翻译研究范式发生了明显的迁移，研究水平有了大幅度提高
  - 离全自动高质量的理想还有很大差距
- 应用方面
  - 机器翻译应用的广度和深度都达到了前所未有的水平
  - 未来的应用前景将更加美好！