关于机器翻译的一些微观想法

朱靖波
东北大学自然语言处理实验室
HTTP://www.NLPLab.com

- 在基础理论研究基础上，加强应用驱动的研究选题，实现可用性研究
- 自然语言处理 vs 自然语言理解
- 各有专攻、不能大而全、没有特色
- 加强合作、优势互补、集团作战
- 只有合作将蛋糕做大，才有领域竞争力

新时代的中文信息处理技术

- 支持海量数据处理（互联网环境）
- 支持复杂数据处理（不规范数据）
- 高速处理能力（云计算、并行计算）
- 支持深度计算（语法语义分析技术）
- 良好可用性（应用驱动）

机器翻译

是前三十年研究热点

相信还是下一个三十年热点

相信也是产业发展的重要方向

机器翻译方法

- 基于规则
  - 人工书写规则、互联网环境下鲁棒性差
- 基于实例
  - 人工加工实例库、覆盖度问题
- 基于翻译记忆
  - 覆盖度和命中率问题，简单，适用于受限领域或特定环境
- 基于模板
  - 人工书写模板或自动构建模板，覆盖度问题
- 统计机器翻译
  - 鲁棒性较好
  - 可以全自动方式构建
  - 统计优化模型

如何自动构建一套 SMT系统

- 收集大规模双语句对库
- 针对每个句对，预处理后进行自动词对齐
- 完成SMT模型训练
  - 规则抽取
  - 规则打分
  - 特征权重优化 e.g., MERT
- 翻译新句子
  - 规则装载
  - 解码
  - 输入具有最高模型分数的候选译文

主流SMT模型

- 基于短语的模型
- 基于层次短语的模型
- 基于句法的模型
  - 树到串
  - 串到树
  - 树到树
- 但传统语言分析技术（NLP）没有在统计机器翻译领域得到预期的应用

源语句法分析能否帮助SMT?

- 从MT评测来看，性能比较强势有
  - 短语系统、层次短语系统和串到树模型
- 利用源语句法分析没有有效帮助改善SMT
  - 树到串、树到树模型
- 个人观点
  - 不是源语句法分析没有用
  - 而是可能我们没有用好源语句法分析

如何利用先验语言学只是帮助SMT?

- 目前SMT模型训练过程通常将先验语言学只是作为隐性假设
- 因为训练数据没有进行任何根据先验语言学知识的显性标注
- SMT模型只能根据数据分布完成模型训练，不太容易显性有效利用先验语言学知识来指导SMT
  - 1）SMT通常选择句法结构非常糟糕的译文作为输出
  - 2）句法分析性能最好的源语句法树是不是SMT系统所喜欢的，答案好像不是！但有待于进一步分析

SMT模型训练两个严重问题

- 规则打分通常采用MLE，没有考虑SMT的最终评价指标优化问题
- MERT训练线性对数模型只是利用句子的最终解码结果，没有考虑中间解码阶段
- 上述两个问题造成SMT模型训练策略鲁棒性存在问题。现在没有好的解决方案，由于计算复杂度问题

MT 评价指标

- BLEU几乎一统天下
- 但从应用角度来看，用户比较关心可读性
  - 例如通过翻译结果能够理解原文的程度
  - 对于翻译工作者来说，能否有效减少人工翻译代价
- 从应用角度来看
  - 搜索引擎、语音识别没有替代者
  - 机器翻译具有人工翻译替代者
    - 翻译质量成为应用的瓶颈

MT系统融合

- 相同模型融合
  - 不同特征空间
- 同类模型融合
  - 短语系统与层次短语系统
- 不同类模型融合
  - SMT+TM+EBMT+…
- 从应用角度
  - 如何优势互补有效改善MT性能，流水线？
  - 要求不能明显下降翻译效率

即时语音翻译技术

- 语音识别/合成+多国语机器翻译
- 需要云计算支持的语音识别和翻译平台
- 真实应用环境下，语音识别和翻译性能有局限性，但具有一定可用性
- 有趣和有前景的MT应用

翻译助手

- 机器翻译 => 辅助阅读
- 翻译公司的专业翻译人员不爱用MT系统
  - 除了TM系统以外
  - 理由：修改机器翻译译文还不如自己逐词翻译
  - 所有人光脚不穿鞋，这表示：
    - 卖鞋没有市场？
    - 还是待开拓的卖鞋市场？
    - 改善翻译质量是关键

Welcome to NiuTrans World

网址：http://www.NLPLab.com

机器翻译研究需要产学研结合驱动，翻译质量为王！

机器翻译技术虽还不成熟，但具有很好的可用性，应用前景广阔。

谢谢