翻译知识获取：专家知识+互联网数据？

刘洋

清华大学

知识获取

- 翻译知识获取是机器翻译的核心问题之一

翻译知识库

机器翻译

Machine translation

知识获取

- 广度
  - 语种
  - 领域
- 深度
  - 词法
  - 句法
  - 语义
  - 语用

阳光_一巴掌：逆天的中国好声音..我的E神..哈哈@KateWill @李言皓爱喝水 @梁光光早上非常困 //@lee_李小豪: 浙江卫视逆天了，湖南，你OUT了。//@欧美流行最前线: 首发阵容逆天了！！林夕老爷的词"因为相信，所以看得见"，陈胖子连台歌都唱得老纸想飙泪。

@最热视频精选: 陈奕迅、吉克隽逸、李代沫，浙江卫视台歌《梦想天空分外蓝》首发阵容逆天了！！林夕老爷的词"因为相信，所以看得见"，陈胖子连台歌都唱得老纸想飙泪。。http://t.cn/zWduFGY

9月4日16:30 来自iPad客户端 转发(1820) | 评论(114)

基于规则的方法

- 专家编纂的翻译知识语言层次深

翻译知识库

机器翻译

Machine translation

基于统计的方法

- 从数据中获取知识成本低，但语言层次浅

翻译知识库

机器翻译

Machine translation

发展趋势

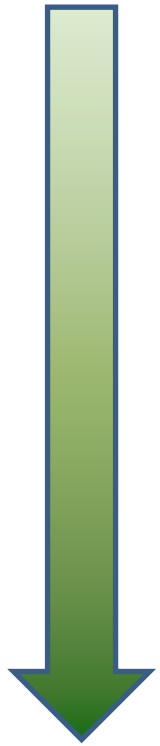
- 自动获取、语言层次不断加深

总统 -> President

奥巴马 总统 -> President Obama

X 总统 -> President X

(NP (X1:NR) (NN (总统))) -> President X1



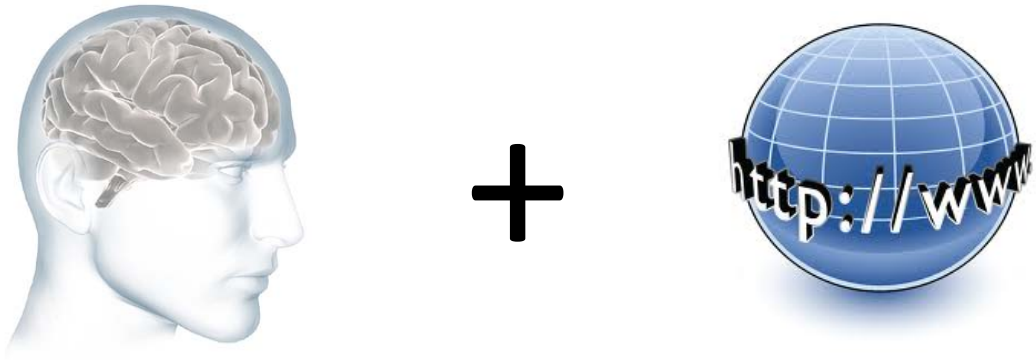
问题

如何从互联网上自动获取高覆盖、深层次翻译知识？

机器学习

- 有监督学习
  - 训练样本：标注数据
- 无监督学习
  - 训练样本：无标注数据
- 半监督学习
  - 训练样本：标注数据 + 无标注数据

半监督学习



专家制作少量高质量的深层次标注数据，利用无监督算法充分利用互联网上海量的无监督数据

国外进展

- Read the Web
  - Never-Ending Language Learning （美国卡内基梅隆大学）
  - Machine reading（美国华盛顿大学）
- Unsupervised Semantic Parsing（美国华盛顿大学）
- Parsing the Web

总结

- 知识获取是机器翻译的核心问题之一
- 如何从互联网上自动获取高覆盖率、深层次的翻译知识是机器翻译目前面临的一大挑战
- 半监督学习有可能将有限的专家知识与无限的互联网数据有效结合起来

谢谢！