自然语言处理中的问题思考

宗成庆

中国科学院自动化研究所
模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

1. 基本概念

自然语言处理是研究如何利用计算机技术对语言文本（句子、篇章或话语等）进行处理和加工的一门学科，研究内容包括对词法、句法、语义和语用等信息的识别、分类、提取、转换和生成等各种处理方法和实现技术。

- 自然语言理解
- 计算语言学
- 中文信息处理

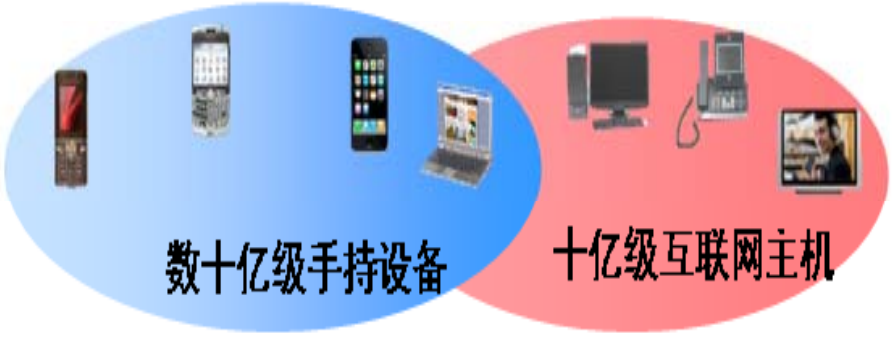
2. 应用目标

- 机器翻译(包括语音翻译)
- 信息抽取、问答系统、信息检索
- 自动文摘
- 文本自动分类、情感分类
- 人机交互

……

3. 社会需求

- 人类历史上以语言文字形式记载和流传的知识占知识总量的80%以上
- 2008年1月中国互联网络信息中心(CNNIC)发布的《第21次中国互联网络发展状况统计报告》表明，中国互联网上有87.8%的网页内容是文本表示的。



微博，博客，email等

4. 基本问题

- **词法分析：** 拉丁语系的形态分析、藏汉语系的分词问题
- **句法分析：** 普遍存在的歧义结构
- **语义理解：** 一词多义
- **常识表示：** 常识知识难以表示
- **推理机制：** 非常规逻辑推理

5. 基本方法

序列标注方法(分类方法)的广泛使用

序/B 列/E 标/B 注/E 方/B 法/E 的/S 广/B 泛/E 使/B 用/E

- 通常使用4个标签：B，I，E，S
- 利用上下文窗口内的特征：$\pm n$，字、标记
- 模型：CRF，SVM，ME，Bayes

5. 基本方法

类似地，

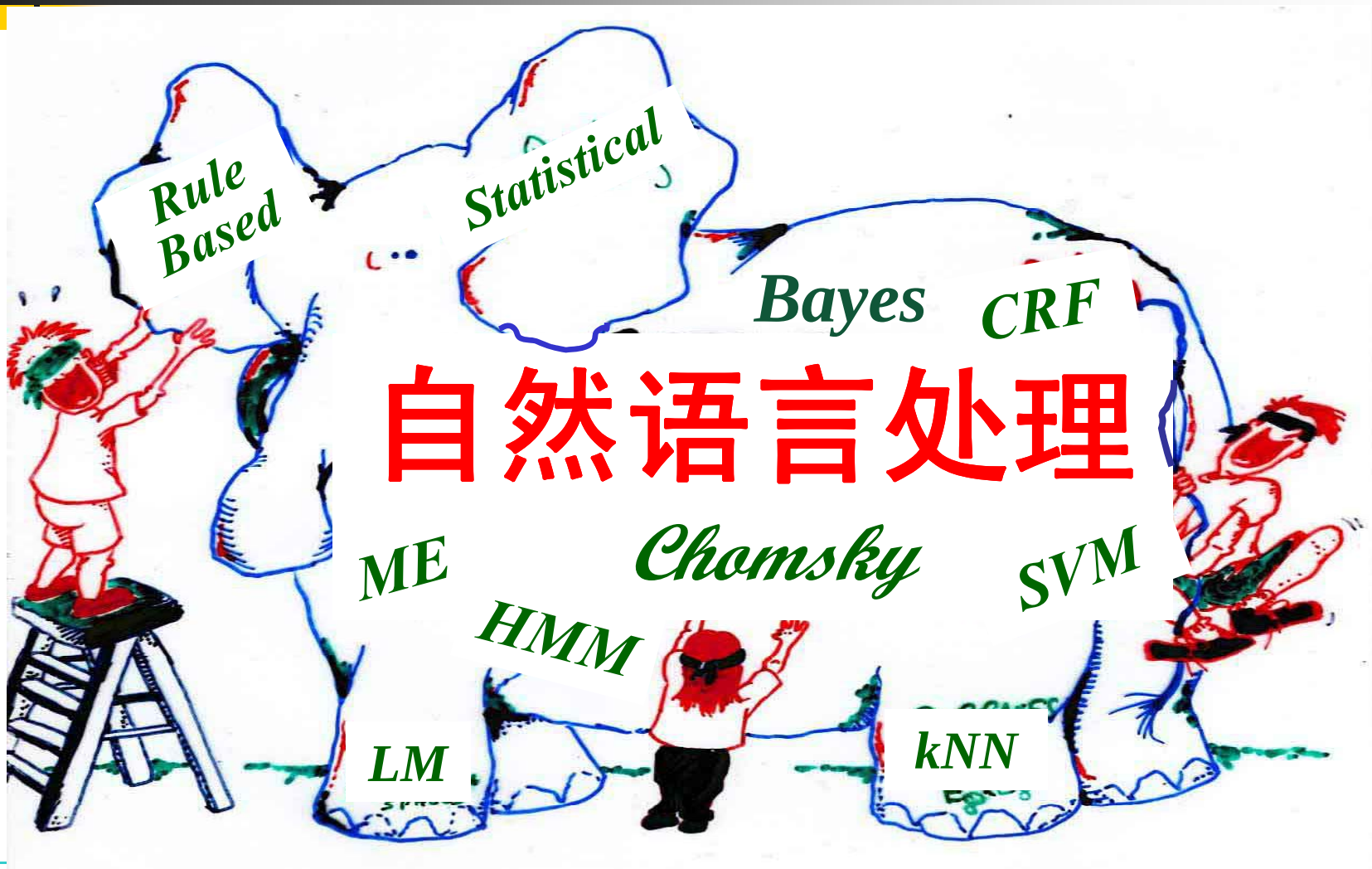
- 命名实体识别 (NER)
- 浅层句法分析 (Base NP, chunking)
- 依存句法分析 (Dependency parsing)
- 词义消歧 (Word sense disambiguation)
- 文本分类/情感分类 (sentiment classification)
- 抽取式自动文摘 (summarization)
- 邮件过滤 (e-mails filtering)

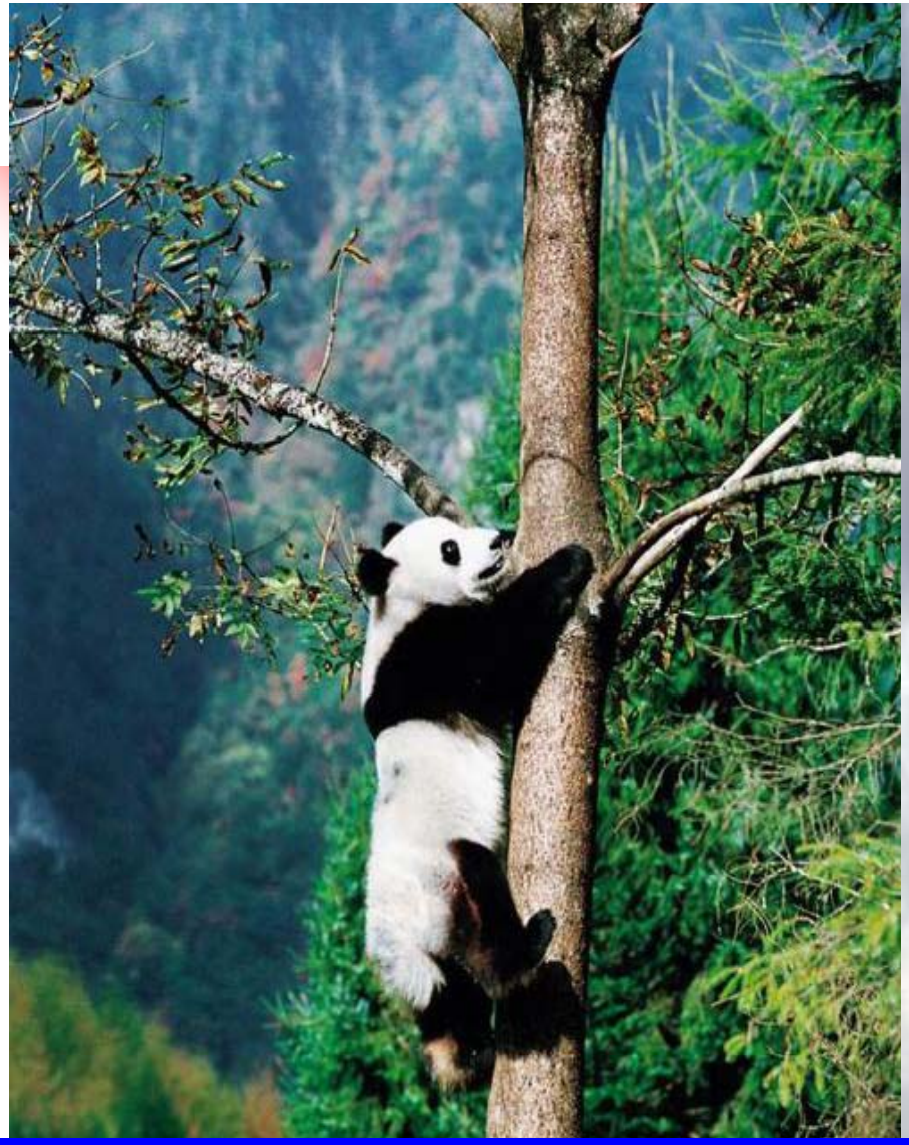
……

6. 面临的挑战

- 特征表示的局限性
  - 词袋(bag-of-word)模型
- 模型的局限性
  - 适应性差
- 常识的表示与利用
  - 建立知识库是一条可行之路吗？
- 语言认知模型的工作机理
  - 人脑进行语言理解时需要大规模事例吗？

7. 技术现状


Rule Based
Statistical
Bayes
CRF
自然语言处理
ME
Chomsky
SVM
HMM
LM
kNN



我们的方法仍是在“爬树”，还是“登月”？

统计学习＋big data＋云计算→问题的解？

Thanks

谢谢!