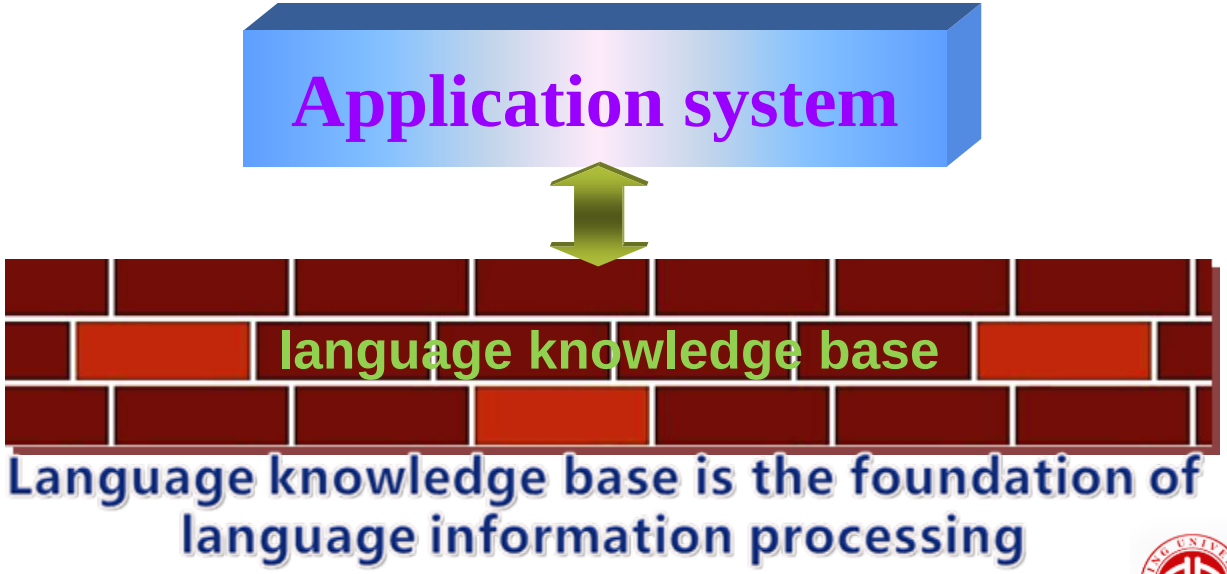
语言知识资源建设的思考

王厚峰(wanghf@pku.edu.cn)
北京大学计算语言学研究所
北京大学计算语言学教育部重点实验室

April, 2014

语言知识资源库的重要性

- **语言知识资源是NLP的基础**
  - 基于规则的方法需要知识(库)支撑(规则本身也是知识)
  - 基于统计的方法需要语料库作为训练数据—建模
- **带标语料库作为黄金标准用于评测**
  - 评测推动技术的发展



语言知识资源库的重要性(续)

- 语言知识有助于语言分析
  - 确定搭配关系（知识：谓词论元关系），例子：
    - 西藏农牧学院努力探寻适合西藏特点的“高产、优质、高效”的农牧业发展路子，取得一系列适合西藏气候与地理特征的科研成果。
  - 分析隐喻（选择性限制，知识：谓词论元关系），例子：
    - 他根据民间传说编织成一篇美丽的童话
    - 不能让错误思想和言行继续自由泛滥
- 语言资源库（语料库）用于建模
  - 分词/标词性/句法结构/语义角色
  - 统计机器翻译
  - …

语言知识资源的特点

- **语言知识库**
  - 对语言知识的一种结构化表示
    - 抽象化特点
    - 结构化特点
    - 密集型表示
  - 典型代表
    - 句法为主（北京大学《现代汉语语法信息词典》）
    - 概念关系：**HowNet, WordNet, CCD, CYC**，同义词词林(扩展板)
    - 谓词论元：**FrameNet**，**VerbNet**
- **语言资源库（语料库）**
  - 带标语料库
    - 知识的实例化表示
  - 不同类别：
    - 多语言（**MT**）
    - 多层次（词、句子、篇章）

知识库的构建

- **构建步骤和方法**
  - 制定规范
    - 样例、说明
    - 培训
  - 知识的获取与工具（构建）
    - 专家（或经专家培训的人）填写
    - 众包（非专家的群体智慧）
    - 自动/半自动提取
  - 知识库的管理与更新
    - 管理、维护、更新
  - 语料库的构建与上述方法类似

规范的适用性（1）

- 知识资源库可能的应用场景是什么（不能包打天下）
  - 能为可能的应用增加哪些信息
  - 如何确保增加期望的信息？
- 知识资源库的适用范围如何界定
  - 表示什么与不表示什么？
  - 如何使知识库有一定的扩展性和灵活性？
    - 语料库的目标与要求
    - 知识库的构建的目标与要求
      - WordNet & HowNet 主要适用于什么领域？效果如何？
  - 不够与多余并存

规范的合理性(2)

- 简单与复杂之间的平衡
  - 简单便于实施，信息不够丰富；
- 体系对语言现象的覆盖性
  - 所有现象都可描述吗？
  - 遗漏的大约有多少（比例）？
- 体系对各种现象的区分性
  - 词的界定问题（仍然存在争议）
  - 谓词论元的角色：“受事”与“对象”的区分

规范的可操作性(3)

- 规范的可操作性
  - 如何确保规范是可操作的
    - 需要考察全部的语言现象
    - 如何减少对人的依赖？
  - 如何确保资源建设的实施中前后是一致的
    - 交叉检查（工作量大）
  - 例子：
    - 词义问题
      - 究竟如何区分词义，“打破”，封锁、怪圈，记录，玻璃，玻璃打破了
  - 太多情况存在不确定性

“先生”的同义词集

Concept number	Synset	Csynset	Hypernym 上位	Hyponym 下位	Definition	Cdefinition
07632177	teacher instructor	教师 教员 老师 先生 导师 臭老九 …	07235322	07086332071623040720946507243767072796590729762207341176074010980741425107425180074940250752093807533674075514040755158107561151076326240763 2736	a person whose occupation is teaching	以教学为职业的人

Sense-1 of “先生”

“先生”的另2个义项

概念编号	Synset	Csynset	上位概念	下位概念	Definition	Cdefinition
07331418	husband hubby married_man	丈夫 先生 夫君 夫婿 爱人 老公 郎君 驸马 驸马爷	07602853	07109482071959680725726073280 08	a married man; a woman’s partner in marriage	已婚男子; 婚姻中女性一方的伴侣

Sense-2 & sense-3 of “先生”

概念编号	Synset	Csynset	上位概念	下位概念	Definition	Cdefinition
07414666	Mister Mr.	先生 师傅 同志 大哥 老兄 老弟	07391044		a form of address for a man	对男子的一种称呼

知识的抽象与具体

- 以谓词论元为例
  - 例子：吃
  - 框架结构：抽象表示施事和受事，出现“过泛化”
    - 施事抽象为“人/动物”，受事：抽象为“食物”
    - 问题：
      - 什么是食物？
        - » 羊吃草，“草”是“羊”的食物，不是“老虎”的食物
        - » 老虎吃羊，“羊”不吃“羊”
        - » 黄鼠狼吃鸡，鸡吃小虫，反过来不成立
  - 框架结构：具体化（实例化）：组合数太大
  - 问题：抽象到什么层次？

名词的物性结构

- **名词：抽象对象+具体对象**
  - 生成词库理论：
    - 强调意义的组合型，意义具有动态性和生成性
    - 源于亚里士多德的“四因说”（Aristotel’s four causes）：质料因、形式因、目的因和动力因
      - 构成角色：描写对象与其组成部分之间的关系。包括材料（material）、重量（weight）、 部分和组成成分
      - 形式角色： 描写对象在更大的认知域内区别于其它对象的属性。包括方位（orientation）、大小（magnitude）、形状（shape）和维度（dimensionality）等
      - 功用角色：描写对象的用途（purpose）和功能（function）
      - 施成角色：描写对象是怎样形成或产生的，如创造、因果关系
    - 例子：
      - 名词“小说”：构成（故事）；形式（书）；功用（读）；施成（写）
      - 名词“书”：构成（纸、文字）；形式（印刷品）；功用（读）；施成（写、印）

名词的物性结构（续）

- 名词物性究竟应该设置多少角色
  - 袁毓林教授增设到9个角色：
    - 形式（formal，简写为FAL）：名词分类属性、语义类型和本体层级特征
    - 构成（constitutive，简写为CON）：名词所指事物的结构属性，包括：构成状态、组成成分、在更大的范围内构成或组成哪些物体、跟其他事物的关系
    - 单位（unite，简写为UNI）：名词所指事物的计量单位（斤，尺）
    - 评价（evaluation，简写为EVA）：对名词所指事物的主观评价、情感色彩（声音：洪亮，动听）
    - 施成（agentive，简写为AGE）：名词所指的事物是怎样形成的
    - 功用（telic，简写为TEL）：名词所指的事物的用途和功能
    - 行为（action，简写为ACT）：名词所指的事物的惯常性的动作、行为、活动（如，“水”，流，滴，翻滚）
    - 处置（handle，简写为HAN）：人或其他事物对名词所指的事物的惯常性的动作、行为、影响（如，“水”，倒，洒，泼）
  - 如何避免角色的交叉（区分）

语篇标注

- **0-指代标注**
  - 什么是 0-指代？
  - 二个例子：
  - 例-1：美国宣布（）部分取消（）对朝鲜长达近半个世纪的经济制裁
  - 例-2：据调查，当时的情况是：旅客苗德和上厕所的时间，正是２９５次列车临进长治北站和乘务员交接班之际，２号车厢列车员在未听懂旅客询问的情况下，(1)急于锁闭厕所，(2)没有正确解答，(3)导致了误解。
- **句间关系的基本单位（小句？标点句）**
  - 上面的例-2
  - 例-3：在茫茫大海上航行了近5个月的中国科考船“雪龙”号，31日结束南印度洋海域的搜寻任务，启程回国，[]沿途将坚持瞭望，预计4月中旬抵达上海。紧张的科考任务让科考队员身心疲惫，但大家自21日执行搜寻任务以来，24小时值守，搜索疑似海域1.17万平方海里。（新华网）

语篇标注

- 句间关系标注

第1层：CLASS	第2层：TYPE	第3层：SUBTYPE
联合关系 (multi-nuclear)	并列(conjunction) [CONJ]	1等立(coordinate) [COOR]
		2时序(temporal) [TEMP]
		3选择(alternative) [ALT]
		4递进(progression) [PROG]
		5顺承(succession) [SUCC]
主从关系 (single-nuclear)	对比(comparison) [COMP]	6转折(contrast) [CONT]
		7让步(concession) [CONC]
	推论(contingency) [CON]	8因果(cause) [CAUS]
		9结果(result) [RESU]
		10目的(purpose) [PURP]
		11假设(hypothetical) [HYP]
		12条件(condition) [COND]
	扩展(expansion) [EXP]	13解证(explanation) [EXPL]
		14分述(list) [LIST]
		15总括(generalization) [GENE]

语篇标注实例

- {1}虽然样子上，【TOP, 1, 2】{2}感觉不怎么有特色，【CONT, 1-2, 3-8】{3}但是它在手感上给用户带来的舒适度却是非常不错的，【LIST, 3, 4-8】{4}它与掌心贴和的比较紧密，【CAUS, 4, 5】{5}不会出现脱手的现象；【COOR, 4-5, 6-7】{6}鼠标移动起来相当灵敏，【RESU, 6-7】{7}没有出现乱跳；【COOR, 6-7, 8】{8}中间滚轮的段落感也不错哦。
- 其中：
  - TOP:话题；CONT:转折；LIST:分述；CAUS:原因；COOR:等立；RESU:结果。

最主要的问题

- **知识资源库的使用状况**
  - 现有的知识资源库多大程度上推动了NLP的发展？
  - 现有的知识资源库在使用中存在什么问题？
  - 如果没有知识资源库，NLP还有什么办法？
- **知识资源库构建的问题**
  - 我们的目的是什么
  - 如何确保可操作性
  - 如何确保与应用契合？
- **当前最需要的知识资源库有哪些，为什么？**


北京大学
PEKING UNIVERSITY
Thank You !