



中文深层语义表达资源描述及其构建技术

穗志方
北京大学

计算语言学教育部重点实验室



北京大学



交流内容

- 中文深层语义分析及资源建设 总体介绍
(穗志方)
- 代表性工作介绍
 - 中文句义分析技术研究 (博士生 : 夏乔林)
 - 中文词义消歧技术研究 (硕士生 : 罗福莉)
 - 中英文语言与知识融合网络 (硕士生 : 奥德玛)



北京大学



中文深层理解——问题与挑战

sina 新闻中心 社会万象

新闻 ▼ 王义博



什么语言最难学？精通32门语言翻译秒回：中文



微博



微信



QQ空间



679

评论



2017年03月31日23:35 未来网

- 中文是世界上最多人使用的语言之一。
- 联合国教科文组织公布的世界十大难学语言中，汉语名列榜首！



北京大学

中文语义计算——问题与挑战

以下翻译来自“百度翻译在线”

1、形同义不同

王长青正在考学生。

Wang Changqing is taking a test. (谁在考试？)

这次考试王长青考第一。

The exam was first. (王长青考得怎么样？)

2、隐含的时态

你们先聊，我去吃饭了 You talk, I went to dinner (我吃过饭了吗？)

3、汉语动结式

她特别会演哭戏，把导演都哭哭了。She would be crying, to see the director. (谁哭了？)

4、语义鸿沟

学钢琴 Learn piano (深层理解：学习的内容是弹钢琴而不是钢琴 learn to play the piano)

敲两个字 Knock two words (深层理解：type the keyboard to input two words)

5、不规则现象 (1+1≠2)

糊涂得可以 Be confused (褒义 or 贬义？)

胖不到哪里去 Fat is not where to go (胖还是不胖？)

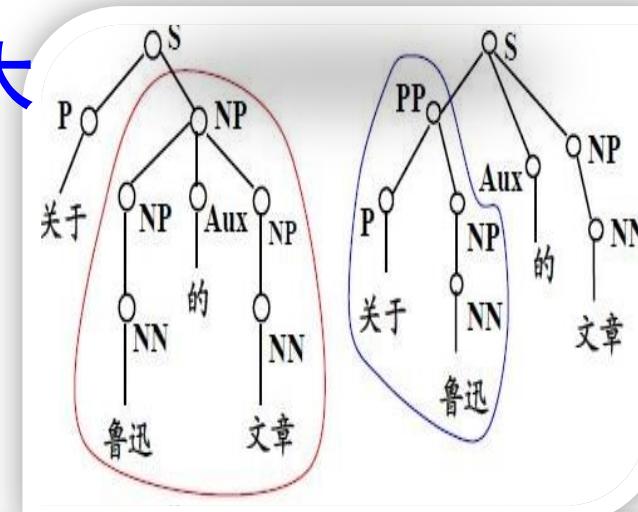
计算机需要实现中文深度理解，才能解决这些问题！



北京大学

中文深层理解——问题与挑战

- ✓ 自然语言理解（NLU）是人工智能领域比人机对弈更难的难题
- ✓ 与西文相比，中文以意合为主，缺乏形态标记
- ✓ 对中文的深层理解又是NLU中的一大难题
- ✓ 对中文深层理解的成功解决，将成为人工智能领域的又一里程碑
- ✓ 需要语言学、认知科学、计算机科学等多学科交叉，联合攻关



北京大学



中文深层理解

- 互联网语言信息处理（问答系统、自动摘要、情感分析、机器翻译、个性化推荐）智能化发展的关键
- 从表层的语言匹配到深层的内容理解
- 实现中文深层分析技术
 - Shallow & Partial (分词、词性标注、句法分析、浅层语义分析) → Deep & Complete (深层语义分析)
- 怎样定义中文深层理解的内涵与任务？
- 怎样从语言资源层面为中文深层理解任务提供基础设施？



北京大学



深度学习与中文语义计算

- 深度学习(deep learning)技术在计算机视觉、语音识别等领域取得突破性进展
- 在NLP任务中，深度学习能带来什么样的进展？
- NLP任务的特点：结构学习(structure prediction)
 - NLP任务通常表现为语言结构的发现和挖掘，结构包括：句法结构、语义结构、信息结构
 - 形式上可以是：序列结构、树形结构、图结构
- 如何构建基于深度学习的结构化学习模型？
 - 分词：面向序列结构解析的神经网络模型
 - 句法分析：面向树形结构解析的神经网络模型
 - 语义分析：面向图结构解析的神经网络模型



北京大学



大数据与NLP

- 大数据是否总能提供更可靠的结果？
- NLP基础数据的两个现状
 - (大而粗)文本数据规模越来越大 (万亿词量级)
 - (小而精)标注语料规模依然很小(百万词、千万词级)
语料标注通常由人工完成、代价昂贵
- 主流方法：有指导机器学习方法
 - 严重依赖标注文本数据



北京大学



大数据与NLP

- 两种类型的NLP任务

- 存在端到端(end-to-end)大数据的NLP任务

- 机器翻译
 - 进行端到端建模
 - 可以利用大数据的效果提升处理效果

- 不存在端到端大数据的NLP任务

- 分词、句法分析、语义分析
 - 性能改善除了模型的合理性，还严重受制于标注数据的规模



北京大学



中文深层理解任务

- 中文深层理解任务不存在端到端的大数据
- 基于中文深层理解语言资源体系，研发中文深层语义技术，实现多层次细粒度（*deep and complete*）分析



北京大学



研究内容

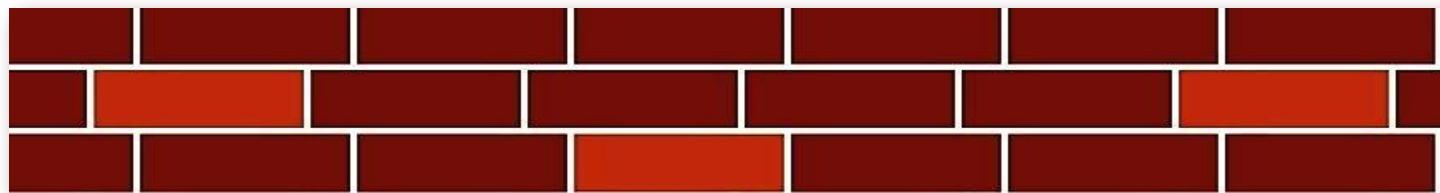
- 怎样定义中文深层理解的内涵与任务
?
- 怎样从语言资源层面为中文深层理解任务提供基础设施？



北京大学

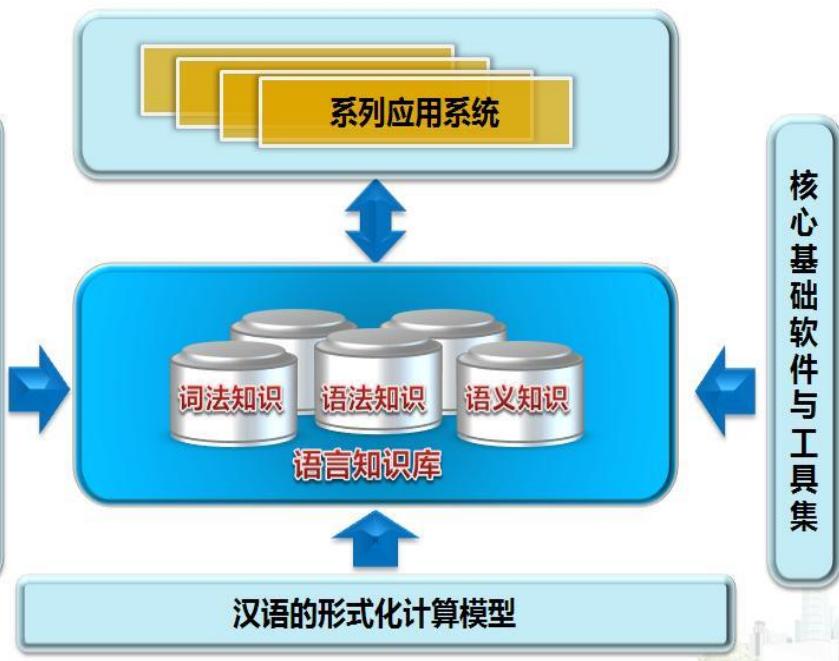
建立面向中文深层理解的 大规模高质量语言知识资源体系

语言知识库是支撑语言信息处理发展的基础设施



CLKB的整体架构

语言知识库规范与国家标准



北京大学



面向中文深层理解的语言知识资源体系



面向中文深层理解的最终目标，在CLKB的基础上力图实现从语法到语义、从语义到认知的逐步扩展



北京大学



研究思路

- 充分利用多学科（语言学、计算机科学、认知科学）融合的优势，基于“语言、认知与计算”基础理论研究成果，探索语言深度理解（尤其是中文深度理解）的内涵
- 借鉴论元结构理论、事件语义学、生成词库论、构式语法理论，突破语义角色标注等浅层语义分析的瓶颈，制订中文深层理解的描述规范
- 从计算机深度计算和语言工程的角度，对文本蕴含的语义信息进行分层次细粒度的深入挖掘。



北京大学



中文深层理解资源描述体系

中文深层语义表达CDSR

(Chinese Deep Semantic Representation)

描述手段1： 知识库（语型）	描述手段2： 标注语料 (语例)	描述对象： 文本语言表达	到认知层面的映射
连接词知识	句际关系标注	篇章	事件、状态关联，情境
谓词的论旨结构	句义标注	句子	事件、状态
词义（义项， 义类）	词义标注	词	词汇化概念 10万
语素义类、结 构规则	构词分析	语素	原子概念 0.5万



北京大学



中文深层理解资源描述体系—整体架构

中文深层语义表达描述策略

分层标注，将句子、语篇蕴含的语义信息逐层剥离出来

I 标注语料库（语例）：

- (1) **命题意义层**：句子的基本意义，是句中主要动词（或形容词）和与其共现的名词性成分之间的关系。
- (2) **逻辑补足层**：是句子基本客观意义之上的主观义，主要由句中的助动词表达，是助动词和句中主要动词（或形容词）之间的关系。
- (3) **实体关系层**：名词短语内部名词与名词之间的语义关系，对应实体之间的关系；
- (4) **事件关系层**：句中出现多个动词（或形容词）表达多个事件时，事件之间发生不同类型的关联。
- (5) **句际关系层**：句子与句子之间的逻辑关系。
- (6) **构式语义层**：句中（或其部分）无法由其构成成分通过简单加合方式得出其整体意义。

II 语义知识库（语型）：语素义知识库、词义知识库、构式义知识库、大学

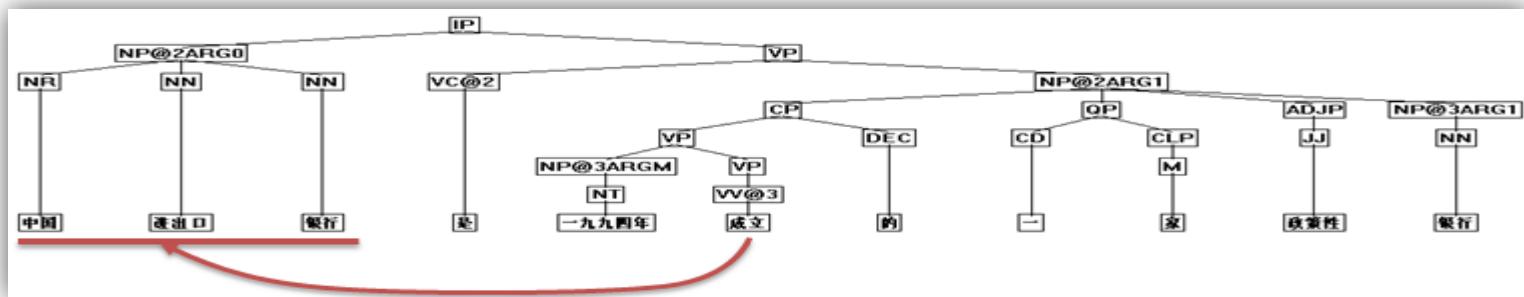


命题义标注 (1)



- 提出了不同于句法结构分析的汉语句义标注策略
- 目前现存多种语法描述体系，各有其优点和不足之处。我们希望对语义信息的标注不受制于某一种句法描述体系。因此，不在句法树上直接标注语义信息。而采用语义为主、句法为辅的方式凸显语义信息。

PropBank



Ours

图 1 “关系事件”概念框架

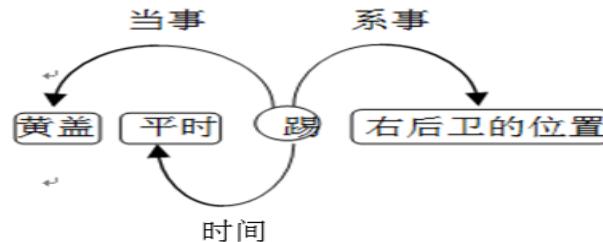
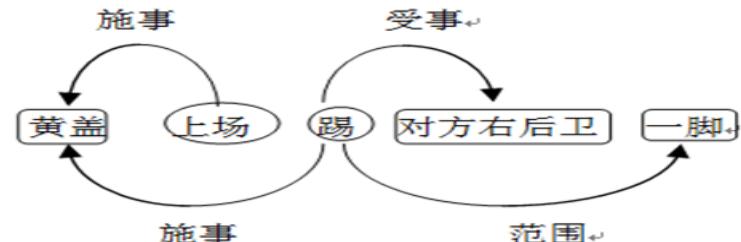


图 2 “动作事件”概念框架





命题义标注 (2)

- 针对汉语中各种典型的句法结构，挖掘表层语言表达蕴含的深层语义。实现针对汉语特点的多样性标注策略
- 述宾结构语义分析
 - 赶 论文 —— 赶 [写] 论文
 - 学 钢琴 —— 学 [弹] 钢琴
 - 催 稿子 —— 催 [人 交] 稿子
 - 参与 意见 —— 参与 [讨论 发表] 意见
 - 敲 两个字 —— 敲 [键盘 输入] 两个字
 - 请 问 —— 请 [允许 我] 问
- 动结式VP语义分析
 - 他 把 玻璃 打 碎 了
 - [%施事 他 %] 把 [%受事 玻璃 %] [# 打 #] 碎 了
 - 他 把 [%当事 玻璃 %] 打 [# 碎 #] 了
 - 事件因果关系：打（他，玻璃） ==> 碎（玻璃）
- 比较结构语义分析
 - 差比结构：[%比较主体 他 %] 的 [%+比较项目 篮球 %] 比 [%比较对象 我 %] [%比较项目 打 %] 得 [%比较结果 好 %]。
 - 同比结构：[%比较主体 张三 %] 跟 [%比较对象 李四 %] [%比较结果 一样 %] [%比较项目 淘气 %]。





命题义标注 (3)

“的”字结构作谓语

- [%受事 这本书 %] [%施事 我 %] [%时间 昨天 %] [# 买 #] 的。

复指成分充当论元

- [%内容 那件事 %] [%施事 你 %] 最好 [%&内容 把它 %] [# 忘 #] 掉。
- [%&结果 本书 %] 是 [%目的 为语言的初学者 %] [# 编写 #] 的 [%结果 一本教材 %]。

不连续成分充当论元

- [%VP内容 “告诉我，” %] [%施事 他 %] [# 问道 #] , [%+VP内容 “他们什么时候枪毙我?” %]

存现结构

- [%处所 桌子一角] [# 放着 #] [%存现物 一本词典 %]。



北京大学



逻辑补足义标注 (1)

1、否定义指对基本命题的否定。否定算子主要是副词“不、不必、没、没有、未、未曾”等。

2、时态义：“时”用来称呼具体的时态，包含将来时和过去时。“体”是用来描写动作行为进行状况的，包含进行体和完成体。

时态/体	标记	典型算子
将来时(future tense)	tense_fut	将/d、即将/d、再/d
过去时(past tense)	tense_past	刚/d、刚刚/d
进行体(progressive aspect)	tense_prog	着/d、呢/d、正/d、在/d、正在/d
完成体(perfect aspect)	tense_perf	了/u、了/y、过/u

- 你们聊，我去吃饭 <了> **tense_fut**。 (将来时)
- 明天我们就在 上海 <了> **aspect_fut_perf**。 (将来完成体)



北京大学

逻辑补足义标注 (2)

3、模态也称情态，是指说话人针对句子客观表述系统的语义进行加工而表达的主观意念，模态语义来源于说话人，指向整个客观表述系统。

	模态	标记	典型算子
可能	必然(certainty)	mod_certainty	一定/d、必然/d、必定/d
	或然(possibility)	mod_certainty	可能/vu、也许/d、或许/d
能愿	意愿(intention)	mod_intention	肯/vu、原意/vu、想/vu
	能力(ability)	mod_ability	能/vu、会/vu
允许	强调(emphasis)	mod_emphasis	的确/d、就/d
	许可(permission)	mod_permission	可以/vu、能/vu
	要求 (requirement)	mod_requirement	应/d、必须/d
	领悟 (comment_1)	mod_comment_1	原来/d、怪不得/d
评注	料定 (comment_2)	mod_comment_2	果然/d、果真/d
	庆幸 (comment_3)	mod_comment_3	幸亏/d、幸好/d
	意外 (comment_4)	mod_comment_4	居然/d、竟然/d
	情绪 (comment_5)	mod_comment_5	胆敢/d、本来/d
	建议(advice)	mod_advice	最好/d
	评判(judgement)	mod_judgement	值得/v
	反诘(rhetorical)	mod_rhetorical	何必/d、何不/d
	捩转(torsion)	mod_torsion	反而/d、反倒/d





逻辑补足义标注 (3)

4、程度义指对所修饰的主要谓词的限制程度，程度算子按限制程度分为增强和削弱两类。

程度	标记	典型算子
增强程度 (high degree)	dgr_high	很/d、非常/d、特别/d
削弱程度 (low degree)	dgr_low	稍微/d、稍许/d 等

5、语气表示说话人对某一行为或事情的看法和态度，是思想感情运动状态支配下语句的声音形式。按表达的语气分为四类；

(功能) 语气	标记	算子
陈述语气	intonation_indicative	的/u
疑问语气	intonation_interrogative	吗/y
祈使语气	intonation_imperative	吧/y
感叹语气	intonation_exclamation	啊/y



北京大学



构式义标注

(北京大学詹卫东)

Language is a mixture of regularity and idiosyncrasy (语言是规则与不规则的混合体) —— Ronald Langacker, 1987, Foundations of Cognitive Grammar, p.411

面向互联网语言的特点，探索现代汉语构式的语义表征。

- 从“常规”的短语结构到“超常规”的构式
- 短语结构 (phrase structure) $1+1=2$
- 构式 (construction) $1+1 \neq 2$
构式 (construction) 指整体意义无法从其组成部分简单加合出来的语言单位。
- 语言中的不规则现象：
- “不是办法的办法”、“胖不到哪里去”、“好你个李云龙”、“男人中的男人”、“左一个报告，右一个请示”、“糊涂得可以”、“你以为你是谁”.....
- 常规的分析方法难以分析其语义



北京大学



构式义标注

(北京大学詹卫东)

针对特定构式的认知机制进行分析，尝试建立从表层语言到深层认知的映射。

构式	
主观认识：Y 有 [特别] X 的特征 Y 是 [特别] 的 X	有一种 X 叫 Y
成功 有毒药 的 特征	甲 有一种毒药 叫 成功
范冰冰有（特别）从容 的 特征	有一种 从容 叫 范冰冰
平静 有（更大）力量 的 特征	有一种 力量 叫 平静
放手 有（特别）爱 的 特征	有一种 爱 叫 放手
永不放弃 有（特别）爱 的 特征	有一种 爱 叫 永不放弃
表象是X，真相（事实）是Y	乙 有一种 误差 叫 数据造假
表象是“误差”，真相是“数据造假”	有一种 倒下 叫 站起
有一种 倒下 叫 站起	表象是“倒下”，真相是“站起”

认知图式		
	X	Y
甲	有一种 X 叫 Y	图式1： 集合 (类)
乙	有一种 误差 叫 数据造假 有一种 倒下 叫 站起	元素 (成员/特例) 远看 (表象) 近看 (实质)

有一种苹果叫富士
有一种毒药叫成功
有一种误差叫数据造假



北京大学



实体关系义标注

序号	语义关系类型	说明	例子
1	POS 整体-部分	释义：“属于”+N1+de+N2 注：N1的外延大 例：北大中文系：属于北大的中文系	机组人员、北大中文系、汪峰女儿、国家财政、居民收入、鲁迅著作、私人电视台、中国银行
2	PAR 部分—整体	释义：“由”+N1+“构成”+的+N2/“包括”+N1+的+N2 注：N2的外延大 例：电脑网络：由电脑构成的网络	电脑网络、花园洋房、食品网络
3	PRO 属性	释义：“是”+N1+的+N2 例：农民群众：是农民的群众	农民群众、农民专家 股份制企业、高个子男孩、重点学校
4	LOC 处所	释义：“位于”+N1+的+N2/“在”+N1+V+的+N2 例：雅典奥运会：在雅典举办的奥运会	雅典奥运会、中国电影、中国银行、河南矿难、江苏霍乱、印尼火山
5	TIM 时间	释义：“在”+N1+（V）+的+N2 例：梅雨季节：在梅雨的季节	梅雨季节、去年春节 唐代诗人、清代家具 晚间新闻、冬季运动、清明时节
6



多层级多分面的中文语义深度刻画

- 命题义标注：

- 考学生 —— [#考#] [%受事 学生%]
- 考语文 —— [#考#] [%内容 语文%]
- 考第一 —— [#考#] [%结果 第一%]
- 考高中 —— [#考#] [%目的 高中%]

- 述宾结构语义分析：

- 学钢琴——[#学#] [%目的 {弹} 钢琴%]

- 动结式VP语义分析：

- 他 把 玻璃 打 碎 了
- [%施事 他%] 把 [%受事 玻璃%] [#打#] 碎 了
- 他 把 [%当事 玻璃%] 打 [#碎#] 了
- 事件因果关系：打(他，玻璃) ==> 碎(玻璃)

- 模态义标注：

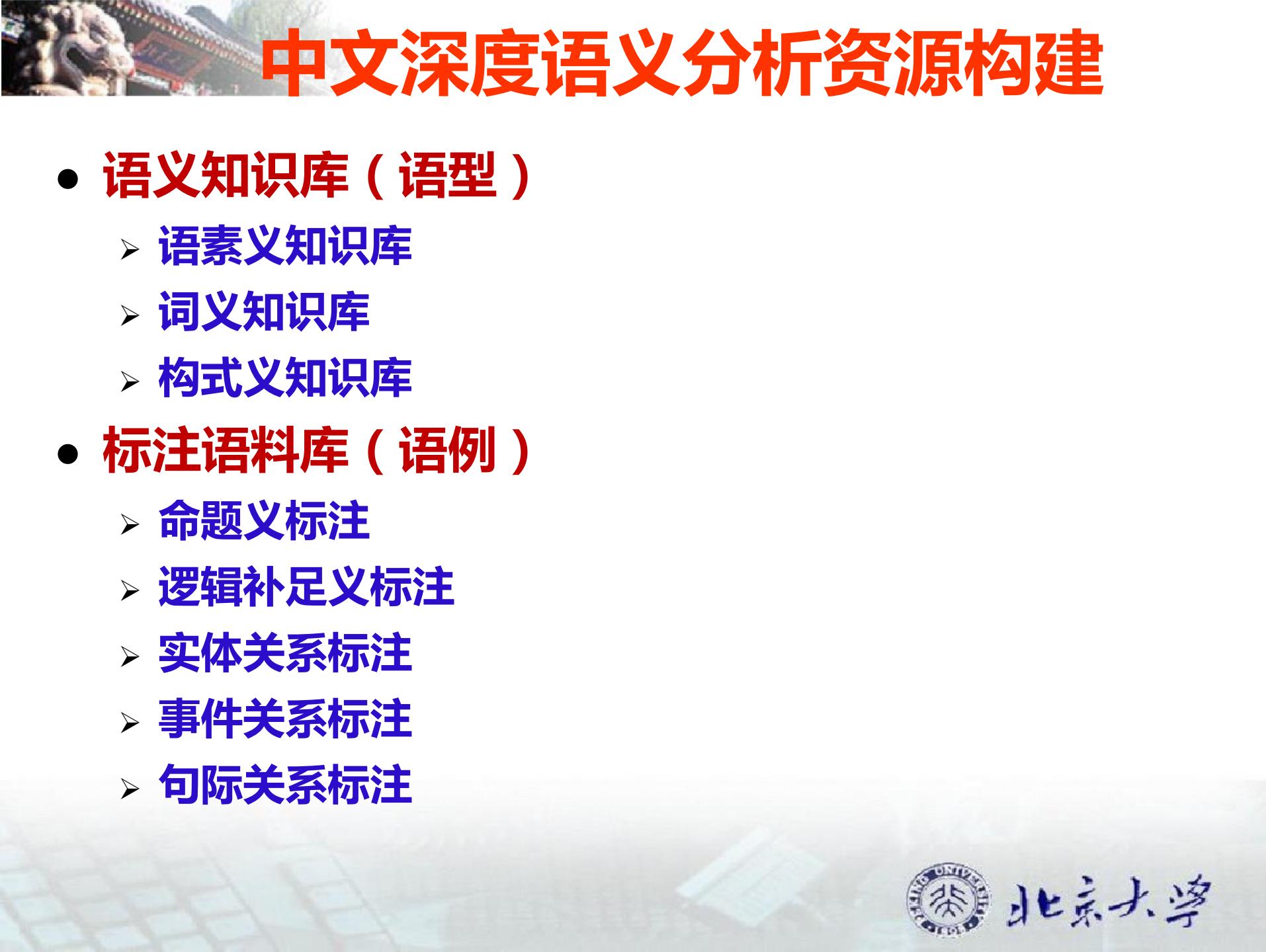
- 你们聊，我去吃饭 <了> tense_fut。(将来时)
- 明天我们就在上海 <了> aspect_fut_perf。(将来完成体)

- 构式义分析：

- 用一吨，少一吨；
- 泡一次，淡一次；
- “A—X，B—X”的释义模板：越A越B(因果倚变义)



北京大学



中文深度语义分析资源构建

- **语义知识库（语型）**

- 语素义知识库
- 词义知识库
- 构式义知识库

- **标注语料库（语例）**

- 命题义标注
- 逻辑补足义标注
- 实体关系标注
- 事件关系标注
- 句际关系标注



北京大学



中文深度语义分析资源——语素义知识库

中文语素义知识库：

- 基于汉语构词的特点，以同义语素集作为最小意义单元来构建中文原子概念知识库
- 基于现汉第5版0.85万单字的2.08万个语素义
- 自底向上构建了2398、1633、723名、动、形同义语素集
- 建立基于汉字的中文原子概念本体：包括同义语素集0.5万个，并形成了原子概念的语义分类体系和属性描述。
- 应用：通过汉语语法构词和语义构词理论，服务于词义知识表示、自动本体构建、新词新义的预测与识别、词典编撰、隐喻研究等



北京大学



中文深度语义分析资源——词义知识库

多源异构的语义词典构建策略

- 聚合关系与组合关系
- 基于标注语料库及多知识源（北大语义词典、现汉、HowNet等）
- 词典与语料库的互动构建
- 语言知识库与世界知识库的连接



北京大学

中文深度语义分析资源——词义知识库

● 聚合关系：1.2万动词的语义分类体系

The screenshot shows the Peking University Concept Editor system interface. The top menu bar includes '北京大学概念编辑浏览系统' (Peking University Concept Editor), '文件系统' (File System), '概念编辑' (Concept Editing), '概念浏览' (Concept Browsing), and '系统帮助' (System Help). Below the menu is a toolbar with buttons labeled NF, OF, CF, ET, NB, NS, NR, DR, DF, DS, CS, XS, PB, PS, FC, FE, EL, EW, AT, and AU.

The main window displays a hierarchical tree structure under '0002 {生物活动}' (0002 {Biological Activities}). The tree includes categories like '0001 {个体活动}' (0001 {Individual Activities}), '0002 {群体活动}' (0002 {Group Activities}), '0003 {生产}' (0003 {Production}), and '0010 {耕作}' (0010 {Farming}). A specific node, '0001 栽_植_种_培植_栽培_栽植_栽种_种植' (0001 Planting_Sowing_Planting_Nursery_Planting_Planting_Planting_Planting), is selected and highlighted with a red border. This node has several sub-nodes listed below it, such as '0002 换茬_抢种', '0003 浸种_选种_引种', etc.

On the right side of the interface, there is a list of terms corresponding to the selected node:

- 栽@null
- 植@null
- 种@null
- 培植@null
- 栽培@null
- 栽植@null
- 栽种@null
- 种植@null

At the bottom of the interface, the status bar displays the following information: 'Concept_index : 3576 @ 5264 Concept_Level : 9 Concept_Position : 000100030001000200020003001000040001 Feature : 0 Synset : 8'.



中文深度语义分析资源——词义知识库

● 组合关系：示例

动词 ^o	CSD 各项 ^o		释义 ^o	标注例句 ^o	基本语义结构 ^o	语义角色 1 ^o		语义角色 2 ^o		语义角色 3 ^o				
	义项 ^o	义项 ^o				角色 ^o 名称 ^o	充任词语 ^o	语义类 ^o	角色 ^o 名称 ^o	充任词语 ^o	语义类 ^o	角色 ^o 名称 ^o		
吃 ^o			把食物等放到嘴里 经过咀嚼咽下去 (包括吸、喝)。 ^o	[%施事 我们 %][# 吃 #] 的 是 [%受事 烤 牛肉 和 土豆 %] 。 ^o	[施事][pred][受事] ^o	[施事] ^o	我们 鸟 警察 史 密斯先生 孩子 骆 驼 ^o	人、兽、职 业、姓名、 ^o	[受事] ^o	午饭 烤牛肉 和 土豆 最普 通的饭菜 药 白饭 馒 头 虾和小鱼 小鹅 瘦肉 草 十八碗 晚饭 一些水 果 面包 早 餐 菜 巧克 力 肥肉 鱼 ^o	食物、药物、 兽、鸟、鱼、 事件、 ^o	^o	^o	^o
			依靠某种事物来生 活。 ^o	云南省委 书记 李 纪恒： 决不 允许 [%施 事 媒体 %][# 吃 #][%受事 共产党 的 饭 %]， 砸 共产党 的 锅 。 ^o	[施事][pred][受事] ^o	[施事] ^o	人 我 媒体 ^o	人、机构 ^o	[受事] ^o	共产党的饭 书 山 ^o	食物、创作 物、地表物 ^o	^o	^o	^o
			消灭(多用于军事、 承受；禁受。 ^o	[%施事 老百姓 %][%原因 因为 舆论 和 谣言 %] 恨不得 [# 吃 #] 了 [%受事 我 们 %] 。 ^o	[施事][原因][pred][受 事] ^o	[施事] ^o	老百姓 ^o	人 ^o	[原因] ^o	舆论和谣言 ^o	信息 ^o	[受事] ^o	我们 ^o	人 ^o
				整整 一个 星期 我 根本 没有 吸烟 。 [%时 间 在 此 期间 %]， [%施事 我 妻子 %][# 吃 #] 尽 了 [%受事 苦头 %] 。 ^o	[时间][施事][pred][受 事] ^o	[时间] ^o	在此期间 ^o	相对时间 ^o	[施事] ^o	我妻子 ^o	关系 ^o	[受事] ^o	苦头 ^o	^o
			受；挨。 ^o	[%施事 武松 %][# 吃 #][%受事 那 一 惊 %]， 酒 都 变做 冷汗 出 了 。 ^o	[施事][pred][受事] ^o	[施事] ^o	武松 ^o	姓名 ^o	[受事] ^o	那一惊 ^o	^o	^o	^o	



研究现状——语言知识库

词汇级：

- 普林斯顿大学的英语词汇语义知识库WordNet
- 中科院董振东先生的知网HowNet
- 加州大学伯克利分校的框架知识库FrameNet
- 北京大学的中文概念词典CCD
- 北京大学现代汉语语义词典

句子级：

- 宾夕法尼亚大学的命题库PropBank
- 北京大学中文网库

篇章级：

- 宾夕法尼亚大学的语篇库PDTB
- 南加州大学的RST篇章树库



北京大学



研究现状——世界知识库



- Dbpedia : 从维基百科中抽取结构化数据，建立结构化知识库
- YaGO : 融合Dbpedia,WordNet,GeoNames，建立常识事件知识库
- Freebase:在线免费开源结构化知识库，包含6800万实体，10亿关系实例
- Knowledge Graph : Google于2010年启动，致力于构建相互关联的实体及其属性的巨大知识图表。基于FreeBase，包含超过5亿实体，35亿知识条目。



北京大学



存在的问题

- 语言知识库和世界知识库相互独立，缺乏关联，在内容理解时，无法将语言知识与世界知识有效地融合起来。

无法满足中文深层理解的需要

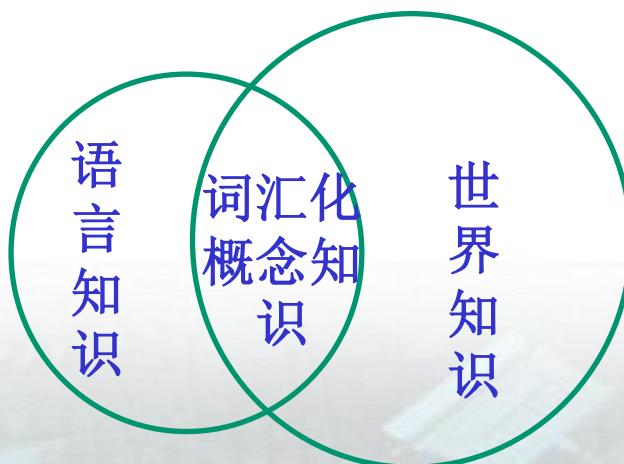


北京大学



语言知识与世界知识的映射与链接

- 将语法与语义相结合、语言与认知相结合，建立面向深度分析和内容理解的知识描述体系
 - 通过名词的物性角色体系和动词、形容词的谓词论元结构体系，描述名词、动词代表的概念、关系之间的相互关联，从而构建一个网状互联、相互照应的知识描述体系。
 - 语义知识集中描述汉语中词汇化的概念知识，以此构建汉语语义概念知识的核心，进而构建更广范围的世界知识。



北京大学



总结

- 怎样定义中文深层理解的内涵与任务？
- 怎样从语言资源层面为中文深层理解任务提供基础设施？
- 基于中文深层理解语言资源体系，研发中文深层语义分析技术，实现多层次细粒度（deep and complete）分析



北京大学



代表性工作介绍

- 中文句义分析技术研究（博士生：夏乔林）
- 中文词义消歧技术研究（硕士生：罗福莉）
- 中英文语言与知识融合网络（硕士生：奥德玛）



北京大学



谢谢！



北京大学