语义知识库简介

詹卫东

http://ccl.pku.edu.cn/doubtfire

http://ccl.pku.edu.cn/doubtfire/lc_list.asp?folder=semantics

提纲

1 WordNet

2 FrameNet

3 MindNet

4 CYC, ILD, EDR

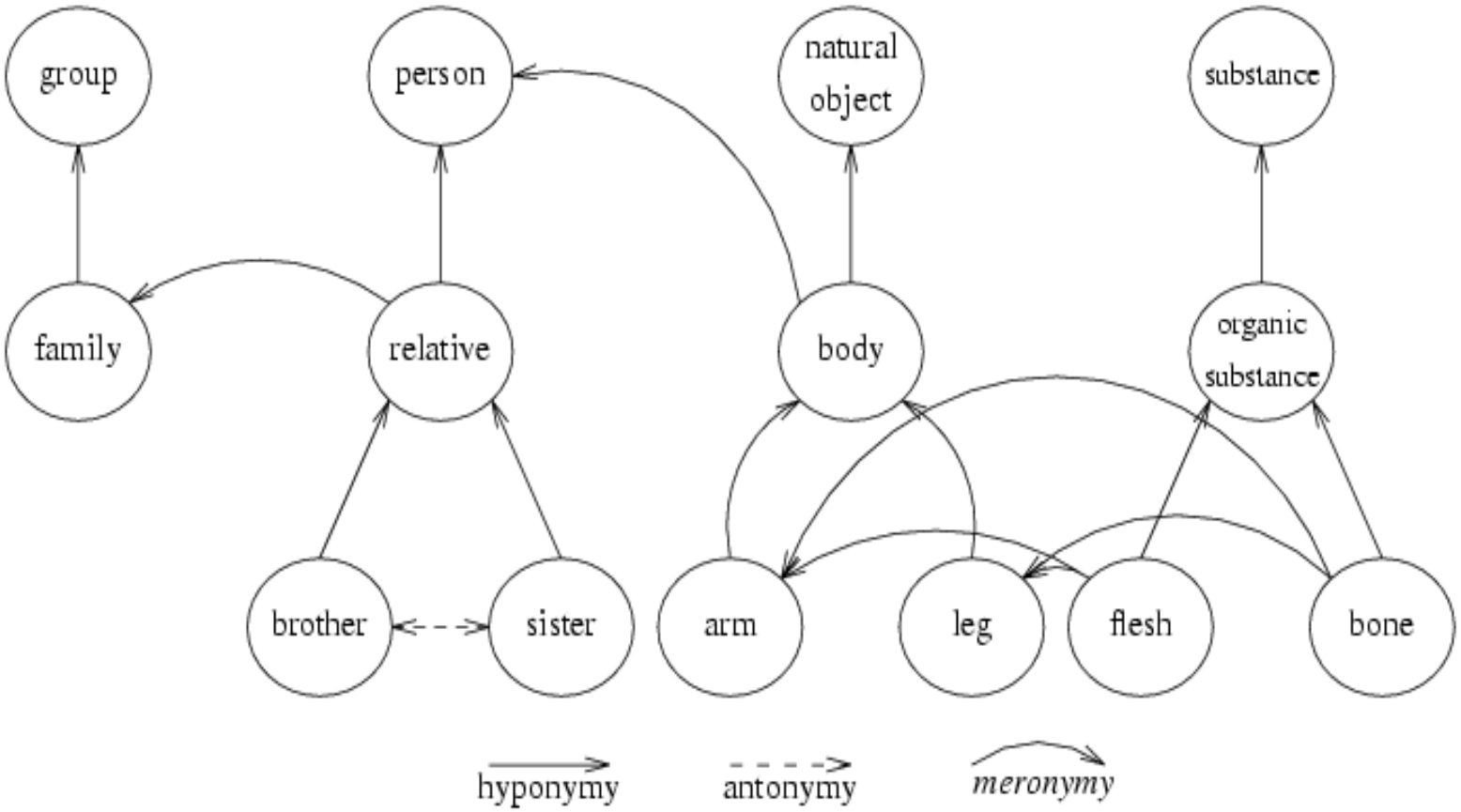
5 HowNet（知网）

6 905工程，汉语配价词典，CCD（中文概念辞典）

7 小结：对语义知识库的评价

1 WordNet

https://wordnet.princeton.edu/



WordNet概况

- 1985 ——
- George.A.Miller, Katherine J. Miller, Christiane Fellbaum, Randee Tengi, …
  Cognitive Science Laboratory, Princeton University
- Christiane Fellbaum, ed., 1998, WordNet : an electronic lexical database, The MIT Press
- http://www.cogsci.princeton.edu/~wn/
- http://www.hum.uva.nl/~ewn/

WordNet发展简史

- 70年代：基于义素分析的词汇语义学（componential lexical semantics）
- 80年代：基于关系的词汇语义学（relational lexical semantics）
- 1985: Miller, *WordNet: A Dictionary Browser*,
  - 可以使用同义词集合（synset）来代表词汇概念，形成词汇网络，即在词的形式和意义之间建立起映射关系（mapping）。
  - WordNet被设想为是一个词典浏览器，是一个机器可读词典的辅助工具。而这样一个机器词典不是按字母排序的，是基于意义组织起来的。
- 1987: Christiane Fellbaum加盟WordNet
- 1991年7月，WordNet 1.0版，包含44983个同义词集合
- WordNet 1.6，1.7.1， 2.1， 3.0，3.1 …

WordNet的心理语言学假设

- 可分离性假设（Separability hypothesis）：语言的词汇成分可以被离析出来并专门针对它加以研究。
- 可模式化假设（patterning hypothesis）：一个人不可能掌握他运用一种语言所需的所有词汇，除非他能够利用词义之间存在的系统的模式和关系。
- 广泛性假设（comprehensiveness hypothesis）：计算语言学如果希望能像人那样处理自然语言，就需要像人那样储存尽可能多的词汇知识。

WordNet词汇的来源

- 语料库
  - Brown语料库；
- 已有的一些词表
  - Laurence Urdang （1978）的《同义反义小词典》；
  - Urdang（1978）修订的《Rodale同义词词典》；
  - Robert Chapmand（1977）的第4版《罗杰斯同义词词林》；
  - 美国海军研究与发展中心的Fred Chang的词表，与WordNet原有词表只有15%的重合词语（1986）
  - Ralph Grishman和他在纽约大学的同事的一个词表，包含39143个词，这个词表实际上包含在著名的COMLEX词典中。WordNet当时词表与该词表重合率为74%（ 1993年）。

WordNet中有什么

- WordNet描述的对象

  compound（复合词）、phrasal verb（短语动词）、collocation（搭配词）、idiomatic phrase（成语）、word（单词），其中word是最基本的单位。

- 对象之间的语义关系

  同义反义关系（synonymy，antonymy）

  上下位关系（hyponymy，troponymy）

  部分整体关系（entailment，meronymy）

  ……

- 部分句法信息

  简单的动词基本句式信息（Verb Sentence Frames）

  e.g. beat (somebody ---s somebody)

WordNet中没有什么

- WordNet并不把词语分解成更小的有意义的单位（这是义素分析法的方法）；WordNet也不包含比词更大的组织单位（如脚本、框架之类的单位）；
- WordNet不是在文本和话语篇章水平上来描述词和概念的语义，因此WordNet中没有包含指示词语在特定的篇章话题领域的相关概念关系。例如，WordNet中没有将racquet（网球拍）、ball（球）、net（球网）等词语以一定方式联系到一起。
- WordNet中缺少关于词语的句法信息；
- WordNet中没有“IS-NOT-A-KIND-OF”这样的关系；
- WordNet中没有区分“IS-A-KIND-OF”和“IS-USED-AS-A-KIND-OF”两种关系，比如，“A thrush is a bird”是前一种关系，而“An adornment is a decoration”则是后一种关系。更典型的例子也许是“Chicken is a kind of bird”和“Chicken is a kind of food”
- ……

WordNet的名词

- 同义词集合（synset）与词汇层级（lexical hierarchy）

  {robin, redbreast} @-> {animal, animate_being} @-> {organism, life_form, living_thing},

- 25个基本类别（25 unique beginners）

  {act, activity} {food} {possession} {animal, fauna} {group, grouping} {process} {artifact} …

- 很少有超过10到12层的语义树，通常层次比较深的情况是由于专业词汇造成的，而不是日常语言中的用词。比如：

  shetland pony @-> pony @-> horse @-> equid @-> odd-toed ungulate @-> placental mammal @-> mammal @-> vertebrate @-> chordate @-> animal @-> organism @-> entity (12 levels)

词汇层级的心理学证据和语言学证据

- Collins & Quillian (1969)： distance in hierarchy
  A robin is a bird -- A robin is an animal
- Smith & Medin (1981)：typicality or prototypicality theory
  A robin is a bird -- A chicken is a bird
- √ *I gave him a good novel, but the book bored him*
  × *I gave him a good novel, but the catsup bored him*
- 动词的搭配选择限制也表明名词上下位关系的重要性。比如动词“drink”的直接宾语可以是 beverage （饮料）的任何一个下位词。这也暗示有关名词的上下位关系的知识应该以一种人们能够快速访问和搜索到的方式存贮

WordNet名词的整体与部分关系

- A是B的组成部分； beak / wing -> bird
- A是B的成员； tree -> forest
- A是B的构成材料。 aluminum -> plane

- {wheel} is a part of {vehicle}
  {sled} is a kind of {vehicle}
  {wheel} is NOT a part of {sled}

{wheeled_vehicle}

- the branch is a part of the tree
  the tree is a part of the forest
  $\not\Rightarrow$ the branch is a part of the forest

WordNet的形容词

- 描写性形容词（descriptive adjectives）

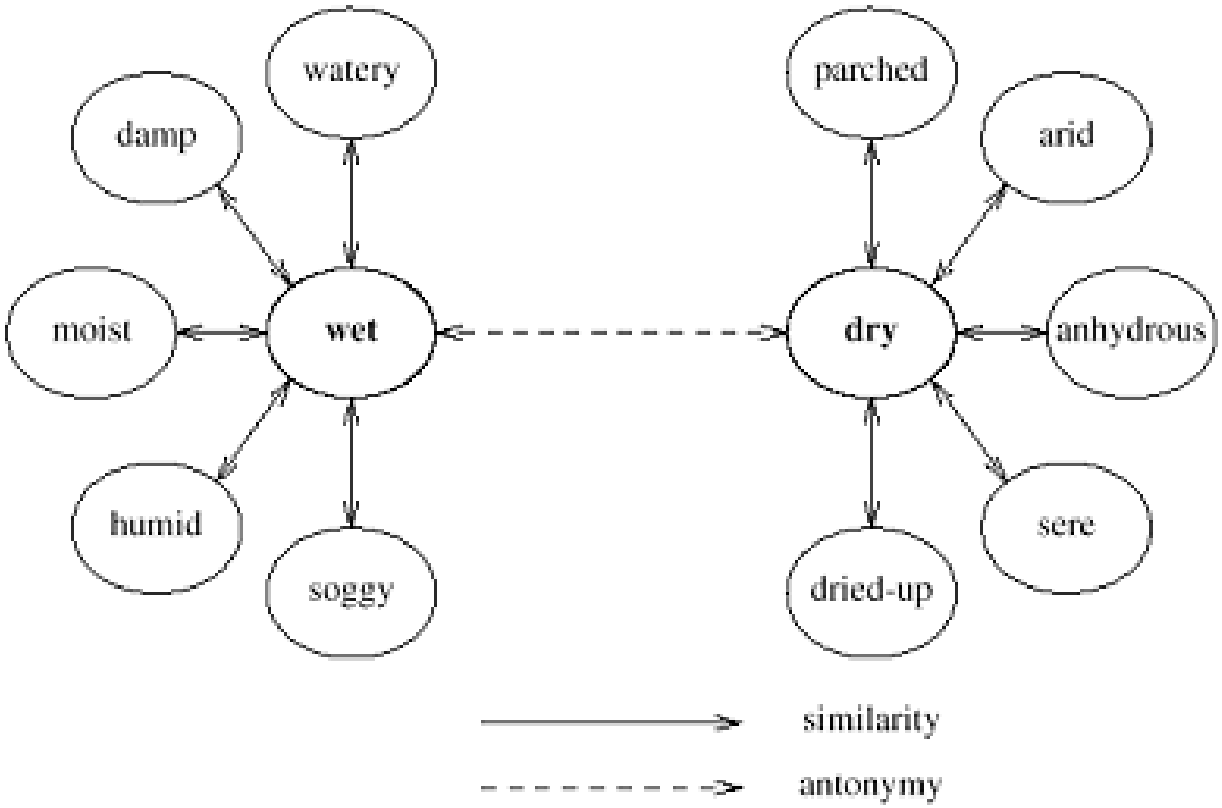
  e.g. big, beautiful, interesting, possible, married, ……

- 关系性形容词（relational adjectives）

  e.g. fraternal, electrical, sidereal, ……

  说明：关系形容词因其跟名词的关系而得名，如 *electrical engineer* 中的 *electrical* 实际跟名词*electricity* 相关。

描写性形容词的反义关系



关系性形容词的特征

- 只能出现在定语位置（*attributive position*）；
- 意义上跟一个名词非常相关；
  fraternal twins —— fraternal : brother
  dental hygiene —— dental : tooth
- 不受程度副词修饰
  ** the extremely atomic bomb*

- 没有直接的反义词
  non- : something else e.g. nonhuman, noncommercial
  extracellular vs. intracellular
  civil lawyer vs. criminal lawyer
  mechanical engineering vs. electrical engineering

形容词的多义性

- *old man vs. old house*

- *old friend - new friend*
  old friend - young friend

- *economic restructuring - the restructuring was economic*
  *economic slump - * the slump is economic*

- *the nervous person - the person's nervousness*
  *the nervous disorder - * the disorder's nervousness*

WordNet的动词

- 英语动词的分类

  Lyons *(1977)*： *act, move, get, become, be, make*

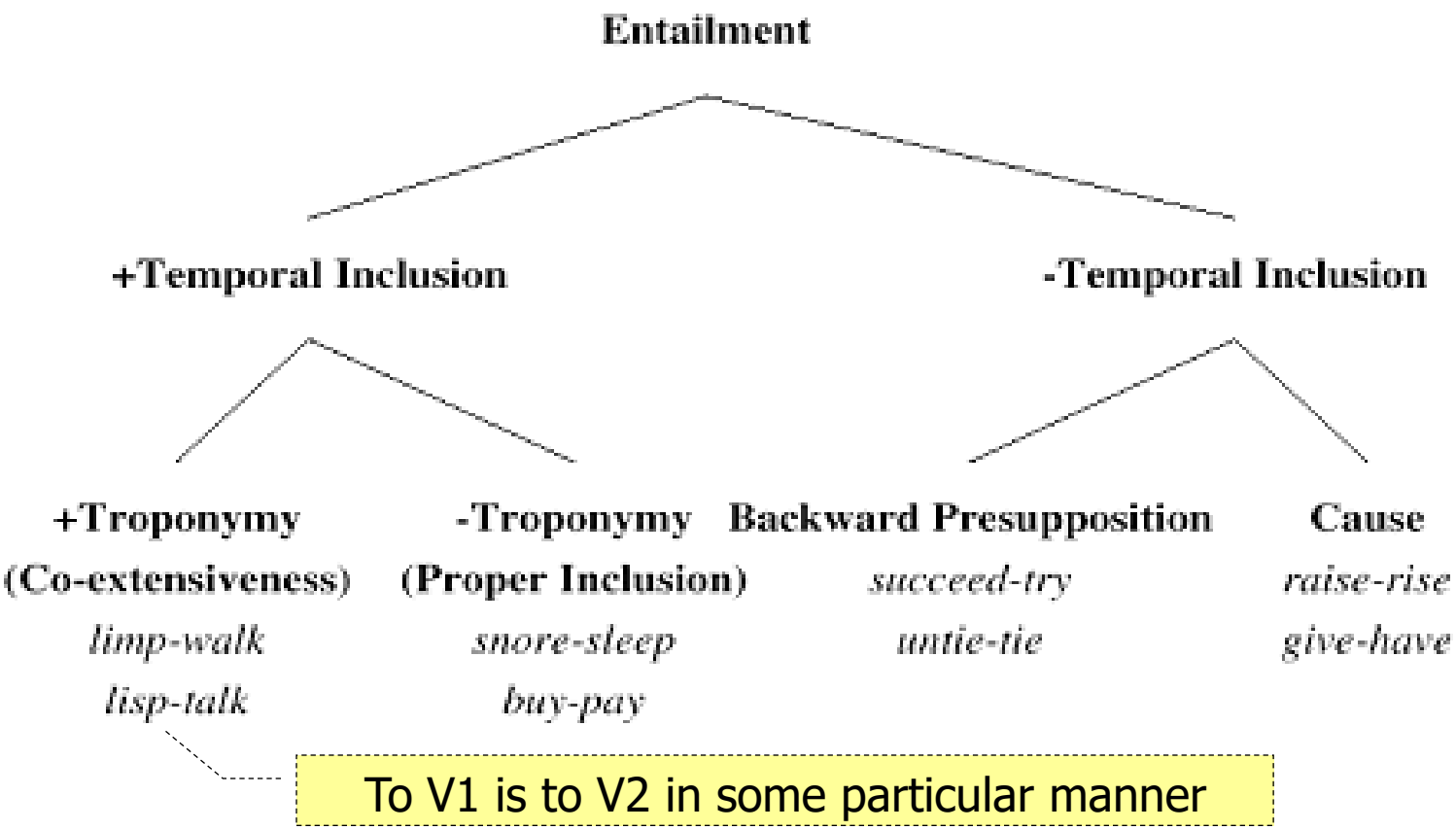
  Pulman *(1983): be, do* = *activity, stative verb*

  Jackendoff *(1983): event, state*

- WordNet动词的15个基本类(semantic domain)

Motion/动作	Perception/感知	Contact/联系	Communication/通信	Competition/竞争
Change/变化	Cognition/感知	Consumption/创造	Creation/创造	Emotion/情绪
Stative/状态	Possession/领有	Body/身体动作	Social/社会行为	Weather/天气动词

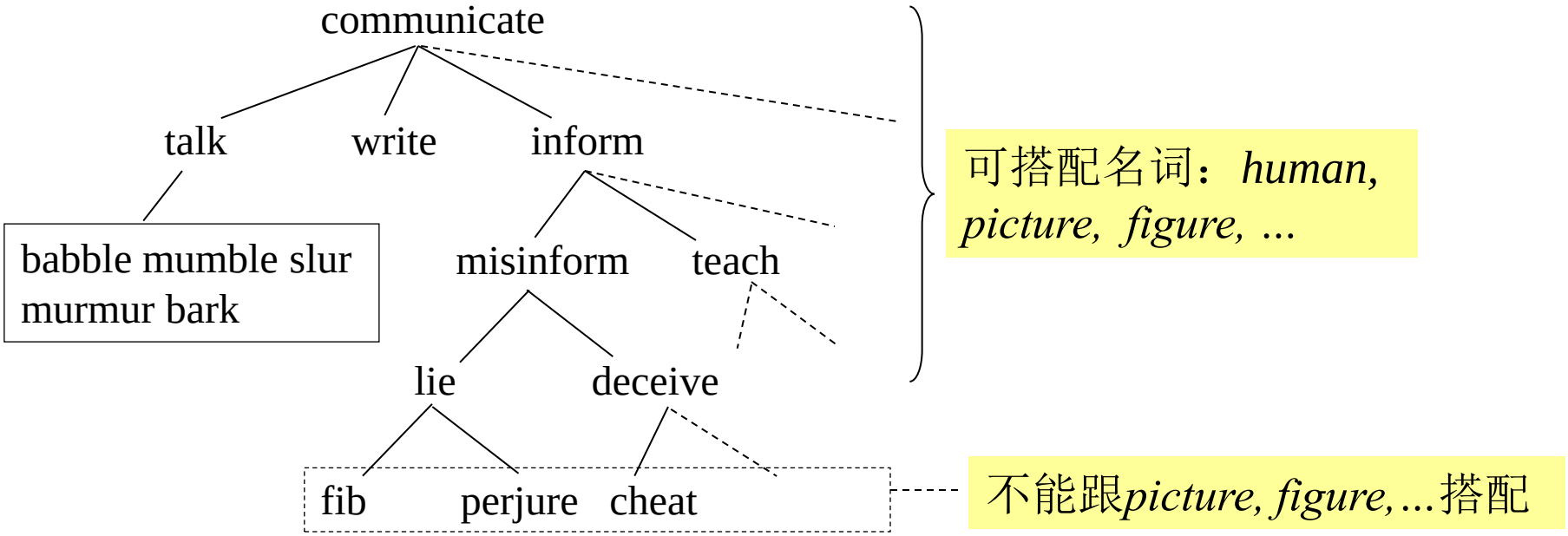
WordNet动词的蕴涵关系



WordNet动词的反义关系

- *give/take; buy/sell; lend/borrow; teach/learn* 没有共同上位词
- *live/die; exclude/include; differ/equal; wake/sleep* 状态动词
- *lengthen/shorten; strengthen/weaken; prettify/uglify* 变化动词
- *tie/untie; appear/disappear* 有标记与无标记的对立
- *rise/fall; walk/run* 有共同上位词
- *fail/succeed* → *try;* *forget/remember* → *know* 蕴涵关系
- *damage/repair* → *damage ; remove/replace* → *remove*

动词上下位层级与动名搭配关系



汉语： 告诉 -> （隐瞒、欺骗） -> （欺诈、诈骗）

WordNet的实施

- 两个相对独立的任务:
  - 人工编写WordNet源文件——这些文件的内容是WordNet词库的实体;
  - 开发一系列计算机程序，这些程序可以处理源文件，并最终产生出可以在用户面前呈现的词典内容。
- WordNet系统包含四部分:
  - 1）WordNet词典编纂人员的源文件;
  - 2）将这些源文件转成WordNet词汇数据库的软件;
  - 3）WordNet词汇数据库;
  - 4）用于访问这些数据库的一套软件工具;
- 支持Unix, PC（Windows）, Macintosh等多平台

WordNet关系的自动发现

- *Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.*

  Pattern: NP_0 such as NP_1 {, NP_2,…, (and|or) NP_i} $i \geq 1$
  Output: for all NP_i, $i \geq 1$, HYPONYM(NP_i, NP_0)

- *Most European countries, especially France, England, and Spain, …*

  Pattern: NP_0 {,} especially {NP,} * {or|and} NP_i $i \geq 1$
  Output: for all NP_i, $i \geq 1$, HYPONYM(NP_i, NP_0)

WordNet的应用

- 词义标注
- 基于词义分类的统计模型
- 基于概念的文本检索
- 文本校对
- 知识处理 —— 推理
- ……

Euro-WordNet，Global WordNet Association

- Euro-WordNet
  - 1996 —— 1999
  - Department of Computational Linguistics, University of Amsterdam
  - Dutch, Italian, Spanish, German, French, Czech and Estonian
  - Swedish, Norway, Danish, Greek, Portuguese, Basque, Catalan, Romanian, Lithuan, Russian, Bulgarian, Slovenic.
  - Inter-Lingual-Index based on the Princeton wordnet
- Gloabal WordNet Association
  - A free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world.

2 FrameNet

- Charles J. Fillmore, Dan Jurafsky,…
  UC Berkeley, Univ. of Colorado, Oxford University Press, …
- 1997－
- Fillmore,C.J.1982, *Frame semantics*, In *Linguistics in the morning calm*, The Linguistic Society of Korea ed. Hanshin Publishing Co. Seoul, 111-137. 中译文可在网上查到：
  http://ccl.pku.edu.cn/doubtfire/semantics/FrameNet/framesem.htm
- http://www.icsi.berkeley.edu/~framenet/
- http://163.136.182.112/framesql/notes/index.html
- http://163.136.182.112/framesql/menu/menu.html

Frame Semantics（框架语义学）

- 从格（case）到框架（frame）
  - 试图在“框架”理论内对词义做出系统的描述和解释；
  - 将考察词义之间的联系的单位从句拓展到篇章；
  - 对动词的论旨角色（格）进行了细化；

从框架的角度看词义

- The Commercial Transaction Frame（商业行为框架）

  买者（buyer），卖者（seller），商品（goods），钱（money），成本（cost），价格（price），找零（change），……

- shore - coast

  岸边 海边

  We will soon reach the coast.
  We will soon reach the shore.

  land - ground

  陆地 地面

  The bird spends its lift on the land.
  The bird spends its lift on the ground.

- good pencil, good coffee, good mother, good pilot, good stick

句内词义关系与跨句词义关系

- He pushed against the door. The room was empty.

- He pushed against the door.
  THE DOOR OPENED.
  HE LOOKED INSIDE.
  HE SAW THAT the room was empty.

从格角色到框架元素（Frame Element）

- Agent, Patient, Instrument, …

- Driver, Rider, Vehicle, Path, Distance, …
  - **[$_{D}$** *Van Cheele***]** *was* ***driving*** **[$_{R}$** *his guest* **] [$_{P}$** *back to the station***].**
  - **[$_{D+R}$** *We***]** ***drive*** **[$_{P}$** *home***] [$_{P}$** *along miles of empty freeway***].**
  - **[$_{D+R}$** *We***]** ***drove*** **[$_{Dist}$** *the short distance***] [$_{P}$** *along the A149***].**

FrameNet工程

- FrameNet Project I (1997 – 2000)
  - 以框架语义学理论为基础，构建一个词库；
  - 经过人工语义标注的句子集合（大规模语料库） —— 用于对词项的句法/语义组配性质（combinatory possibility）进行示例；
  - 从上述标注语料中自动抽取词项的组配框架信息；
  - FrameSQL网上查询界面
  - 主要是动词、形容词、和由动词派生的名词
- FrameNet Project II (2000 – 2003)
  - 继续构建词库：（1）增加框架和词项；
    （2）从governing词到dependent词；
  - 在NLP应用系统(WSD,MT,IE等)中测试FrameNet Database的性能

FrameNet数据库中的框架设计

frame(TRANSPORTATION) frame_elements(MOVER(s),MEANS,PATH) scene(MOVER(s) move along PATH by MEANS)
frame(DRIVING) inherit(TRANSPORTATION) frame_elements(DRIVER(=MOVER),VEHICLE(=MEANS),RIDER(s)(=MOVER(s)),CARGO(=MOVER(s))) scenes(DRIVER starts VEHICLE, DRIVER controls VEHICLE,DRIVER stops VEHICLE)
frame(RIDING_1) inherit(TRANSPORTATION) frame_elements(RIDER(S) (=MOVER(S)),VEHICLE(=MEANS)) scenes(RIDER enters VEHICLE,VEHICLE carries RIDER along PATH,RIDER leaves VEHICLEE)

FrameNet数据库中的框架设计（续）

Frame: Causation

Lexemes: cause.v, cause.n, make.v

Frame Elements

FE（框架元素名称）	Tag（标记）	Example
Cause	Cause	The wind caused the tree to sway.
Affected	Affected	The wind caused the tree to sway.
Effect	Effect	The rain caused flooding

FrameNet数据库中的框架设计（续）

- **General Description**

  A Cause, animate or inanimate, causes an Effect. Those frames that inherit the Causation frame convey the idea that some event is responsible for the occurrence of another event (or state). In the inheriting frame, typically an FE like Agent or Causer is proposed for an individual or force associated with the causing event, but at bottom we assume event causation.

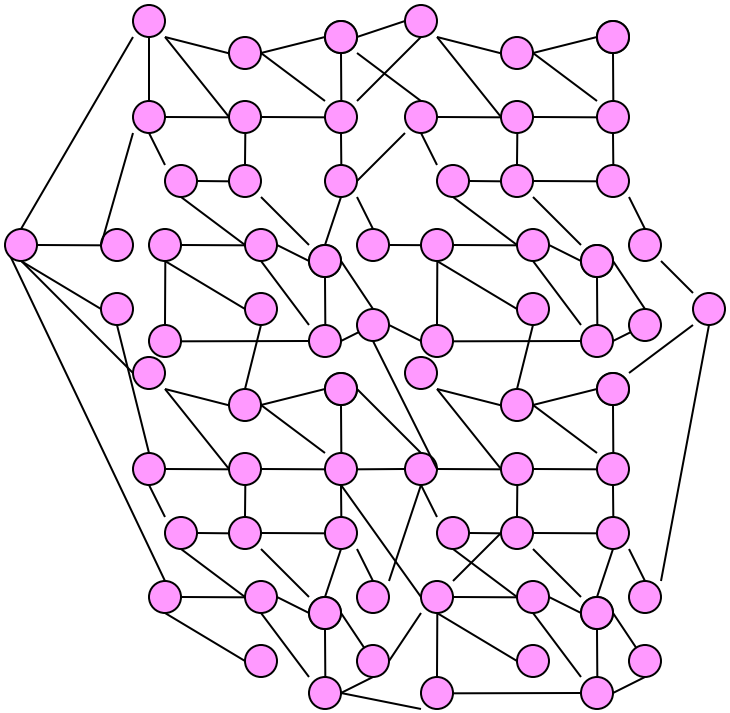
- FE: Cause

  An animate or inanimate entity, a force, or event that produces an effect. Volitionality（意志性） is not a necessary characteristic of Causes.

  - [John] made me give up smoking.
  - [The wind] made the door rattle.
  - [The accident] caused them to be more careful the next time.

3 MindNet

- 美国Microsoft公司
- 大规模语义知识库
- 从《微软百科全书》(Microsoft Encarta)等词典中全自动提取
- 词语之间由关系进行链接
- 目前已有700万关系链接，并且仍在增长



MindNet中的词义关系

Attribute属性	Goal目标	Possessor领有者
Cause原因	Hypernym上位	Purpose意图
Co-Agent联合施事	Location场所	Size大小
Color颜色	Manner方式	Source源点
Deep_Object深层宾语	Material材料	Subclass子类
Deep_Subject深层主语	Means方法	Synonym同义
Domain领域	Modifier修饰语	Time时间
Equivalent同位	Part部分	User使用者

MindNet关系的自动提取

- car: *a vehicle with 3 or usually 4 wheels and driven by a motor, esp. one for carrying people*

```
car
 |
 |___Hyp>---- vehicle
 |
 |___Part>---- wheel
 |
 |___<Tobj ---- drive
 |                  |
 |                  |___Means>--- motor
 |
 |___Purp>----- carry
                    |
                    |___Tobj>--- people
```

MindNet关系的扩展

- car　　car - Hyp -> vehicle
  truck　　vehicle <- Hyp - truck

  car - Hyp -> vehicle <- Hyp - truck

- watch　　watch - Hyp -> observe
  telescope　　observe - Means -> telescope

  watch - Hyp -> observe - Means -> telescope

4 CYC、ILD、EDR

- CYC 美国CYC公司 (1984 -)

  http://www.cyc.com/index.html

- ILD 英国剑桥综合语言知识库 (1993 - 1996)

  http://www.hcrc.ed.ac.uk/Site/ILD.html

  http://www.ltg.ed.ac.uk/projects/ild/

  http://icl.pku.edu.cn/doubtfire/semantics/CUP_ILD/ild-index.htm

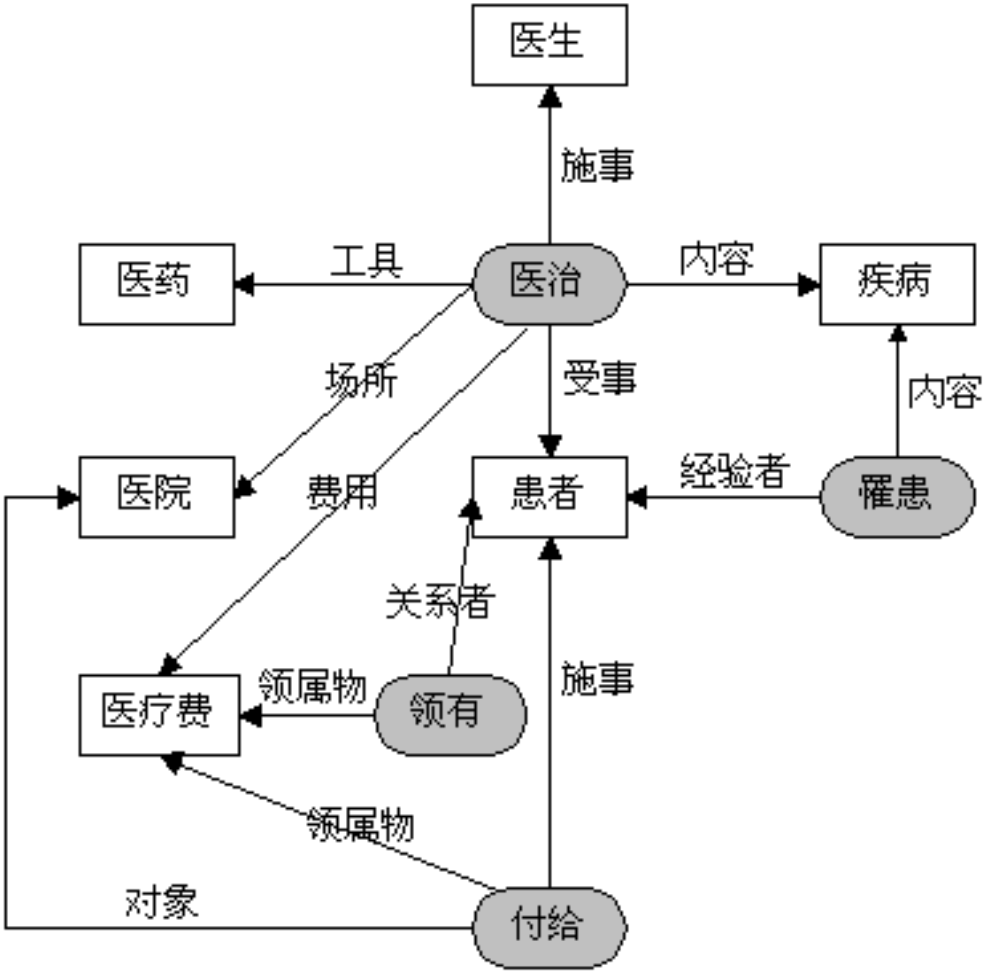
- EDR 日本电子词典研究所 (1986 -)

  http://www.iijnet.or.jp/edr/

5 HowNet（知网）

- 1988 — 1998 —　　董振东　董强
- 知网（英文名称How-Net）是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库
- 人工构建，自底向上归纳义原（知网标记集）
- 董振东,1998,《语义关系的表达和知识系统的建造》,载《语言文字应用》1998年第3期。
- http://www.keenage.com/

HowNet的目标：通向“真正”的理解



HowNet定义的语义关系

- (a) 上下位关系
- (b) 同义关系
- (c) 反义关系
- (d) 对义关系
- (e) 部件-整体关系
- (f) 属性-宿主关系
- (g) 材料-成品关系
- (h) 角色-事件关系

施事/经验者/关系主体 - 事件关系
受事/内容/领属物 - 事件关系
工具 - 事件关系
场所 - 事件关系
……

HowNet义原的上下位关系示例

```
- entity|实体
 ├ thing|万物
  …  ├ physical|物质
     …  ├ animate|生物
        …  ├ AnimalHuman|动物
           …  ├ human|人
              │  └ humanized|拟人
              └ animal|兽
                 ├ beast|走兽
   …
```

HowNet词项基本形式

NO.=030010

W_C=拐杖

G_C=N

E_C=

W_E=walking stick

G_E=N

E_E=

DEF=tool|用具,#walk|走,#disable|残疾

记录号

中文词

中文词性

中文用例

定义/释义

HowNet名词示例

雇主： DEF=human|人,*employ|雇用

雇员： DEF=human|人,$employ|雇用

熨斗： DEF=tool|用具,*AlterForm|变形状,#level|平

假期： DEF=time|时间,@rest|休息,@WhileAway|消闲

旅馆： DEF=InstitutePlace|场所,@reside|住下,#tour|旅游

救生艇：DEF=ship|船,*rescue|救助

心脏：DEF=part|部件,%AnimalHuman|动物,heart|心

CPU：DEF=part|部件,%computer|电脑, heart|心

* 表示施事、经验者、或关系主体等角色；$ 表示受事、内容、领属物等角色；
# 表示相关关系；@ 表示场所、时间等角色；%表示部分整体关系

HowNet动词示例

NO.=015492

W_C=打

G_C=V

E_C=~毛衣，~毛裤，~双毛袜子，~草鞋，~一条围巾，~麻绳，~条辫子

W_E=knit

G_E=V

E_E=

DEF=weave|辫编

救灾：DEF=rescue|救助,StateIni=unfortunate|不幸

扭亏为盈：DEF=alter|改变,StateIni=InDebt|亏损,StateFin=earn|赚

HowNet动名语义关系描述

V event|事件

V1 static|静态

V2 act|行动

ActGeneral|泛动 {agent,content}

start|开始 {agent,content}

do|做 {agent,content,manner}

try|尝试 {agent,content}

endeavour|卖力 {agent,content}

VieFor|快干 {agent,content}

RashlyAct|蛮干 {agent,content}

venture|冒险 {agent,content}

……

HowNet信息结构库

- 信息结构
  - 餐馆 ：可以吃饭的场所
  - 走私集团：一个从事犯罪活动的团体，特征是转移物品
- 句法分布式
  - 餐馆：N1 + N2　　走私集团：V + N
- 句法结构式
  - 餐馆：N1 ← N2　　走私集团：V ← N
- 信息结构模式
  - 餐馆：{(物质,食物) [受事] <-- <事件,行动,吃>} <-- [处所] (组织/场所)
  - 走私集团：(事件,行动) <-- [施事] (人/拟人)

HowNet信息结构库（续）

分类号

1.3.3.30

句法结构式

SYN_S=N <-- N

信息结构模式

SEM_S=(万物) [领属物] <-- (万物)

Query1: 什么？

Answer1: N1 + N2

Query & Answer: 表示该信息结构模式传达的真正信息并由此可产生的问与答

Query2: 什么样的？

Answer2: “有” N1 “的” N2

例子：花园-洋房，星-空，艳阳-天，草-原，草-地，沙-地，花-园，林荫-道，林荫-路，林荫-大道，水翼-船，气垫-船，

功勋-演员，功勋-运动员，技术-人员，专业-技术人员，专业-人才，专业-人员，技-师，技-工，技术-员，手艺-人，

HowNet的规模

HowNet双语知识库

中文词项 53335
英文词项 57392
中文词条 65953
英文词条 75356
总记录数 116533

HowNet信息结构库

信息结构模式：271个
句法分布式：49个
句法结构式：58个
实例：11,000词语
总字数：中文60,000字

6 905工程、汉语配价词典、CCD

- 905工程 （1990.5 — 1995）
  - 国防科工委、电子部、CCID，清华、北大、……
  - 陈力为、袁琦，1995，《中文信息处理应用平台工程》，电子工业出版社
- 汉语配价语义词典 （1997 — 1998）
  - 中科院计算所、北大计算语言所
  - 服务于汉英机器翻译的语义词典
- CCD中文概念辞书 （2000 —
  - 北大计算语言所
  - WordNet-style Chinese Dictionary

汉语配价语义词典

词语	词性	语义信息			
		语义类	配价数	论元角色选择限制	
				主体	客体
大衣	n	服饰	0		
父亲	n	人	1	[语义类:人]	
后胎	n	构件	1	[汉字:*车]	
高兴	a	境况	1	[语义类:人]	
热情	a	品格	2	[语义类:人]	[语义类:人 | 事]
走	v	自移	1	[语义类:人|动物]	
洗	v	促变	2	[语义类:人]	[语义类:人工物]

7 小结：对语义知识库的评价

- 语义知识的类型
  属性：值　　条件 → 动作
- 语义知识的侧重
  聚合关系　　组合关系
  词句水平　　跨句水平
- 语义知识的应用
  句法分析、词义消歧、……
  机器翻译、信息检索、……
- 语义知识的获取方式
  人工，自动，人机交互
- 一致性、兼容性、可扩展性、规模、……

小结

- 语义描述的两个方面
  - 关系
  - 约束

进一步阅读文献

- 陈力为、袁琦，1995，《中文信息处理应用平台工程》，电子工业出版社
- Beth Levin & Malka Rappaport Hovav, 1996, *Lexical Semantics and Syntactic Structure*, In Shalon Lappin, ed., *The Handbook of Contemporary Semantic Theory*, Oxford: Blackwell, Chapter 18.
- Christiane Fellbaum, ed., 1998, *WordNet : an electronic lexical database*, The MIT Press
- http://icl.pku.edu.cn/doubtfire/l_list.asp?folder=semantics
- http://icl.pku.edu.cn/doubtfire/Semantics/973_Beida/index.htm

复习思考题

1. 了解WordNet，FrameNet，MindNet的更多资料，加以比较和评论；
2. 对汉语中跟声响有关的词（比如：喊、叫、哀嚎、……）进行调查，发掘相关语义知识，并组织到一个知识表示系统中；
3. 对汉语描写人的情绪的形容词（比如：高兴、悲伤、颓废……）进行调查，任务同上。
4. 举例说明可以通过词义上下位层级关系来系统描述（解释）的语言现象；