



WordNet 简介

詹卫东

2003.6

zwd@pku.edu.cn

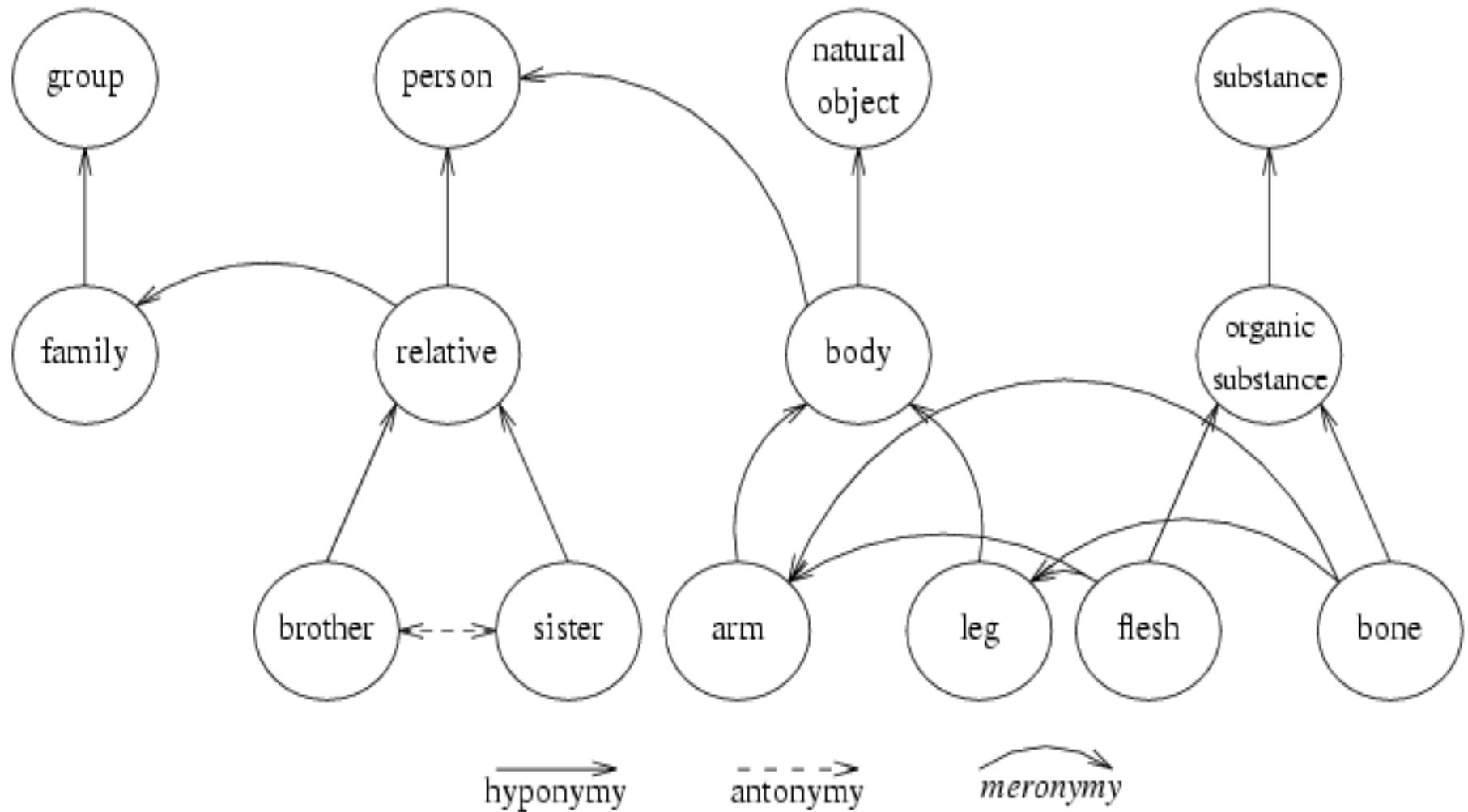
<http://ccl.pku.edu.cn/doubtfire/>



提纲

- 1 WordNet概述
- 2 WordNet中的名词
- 3 WordNet中的形容词
- 4 WordNet中的动词
- 5 WordNet词库与查询软件的设计与实施
- 6 WordNet的应用与发展
- 7 小结

1 WordNet概述





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中的核心概念(synonym set)

Illustrating the Concept of a Lexical Matrix:

F_1 and F_2 are synonyms; F_2 is polysemous

Lexical Matrix

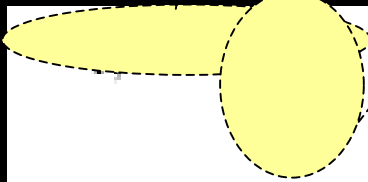
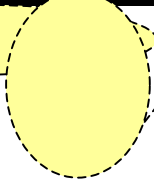
Sysnet:

{board plank}

{board committee}

Glossary

Semantic relations

Word Meanings	Word Forms				
	F_1	F_2	F_3	. . .	F_n
M_1			$E_{3,3}$. . .	$E_{m,n}$
M_2					
M_3					
\vdots					
M_m					

WordNet的关系指针及标记符号

名词		动词		形容词		副词	
反义关系 Antonym	!	反义关系 Antonym	!	反义关系 Antonym	!	反义关系 Antonym	!
下位关系 Hyponym	~	下位关系 Troponym	~	近义关系 Similar	&	导出形式 Derived from	\
上位关系 Hypemym	@	上位关系 Hypemym	@	关系型形容词 Relational Adj.	\		
部分关系 Meronym	#	蕴涵关系 Entailment	*	又见 Also See	^		
整体关系 Holonym	%	致使关系 Cause	>	属性 Attribute	=		
属性 Attribute	=	又见 Also See	^				



WordNet发展简史 (1985 ——)

■ 背景

- 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)

■ WordNet发展简史

- 1978: George A. Miller, *automated dictionary*
- 1985: G. Miller, *WordNet: A Dictionary Browser + Project*
- 1987年春: Philip N. Johnson-Laird, 名词分类
- 1987年夏: Christiane Fellbaum加盟WordNet, 动词分类
- Kitty Miller 负责形容词的描写



WordNet的规模与版本

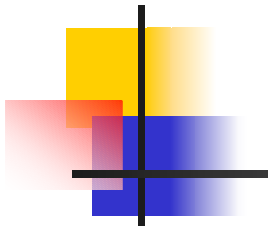
- 1989年4月，WordNet中有37409个同义词集合，没有注释
- 1991年7月，WordNet 1.0版，包含44983个同义词集合，13688个注释（30%）
- 1991年8月，WordNet 1.1版
- 1992年1月，WordNet包含49771个同义词集合，19382个注释（39%）
- 1992年4月，WordNet 1.2 版；1992年12月，1.3版
- 1993年1月，WordNet包含61023个同义词集合，36880个注释（60%）
- 1993年8月，WordNet 1.4版



WordNet的规模与版本（续）

- 1994年1月，WordNet中包含79542个同义词集合，58705个注释（74%）
- 1995年1月，WordNet包括了91050个同义词集合，同时包含了75389个注释（占同义词集合数量的83%）
- 1995年3月，WordNet 1.5版
- 1997年，WordNet 1.6版（支持Windows,Unix,Mac）
- 2001，WordNet 1.7版（只支持Unix）
- 2001，WordNet 1.7.1版（支持Windows和Unix）

Number of words, synsets, and senses



POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	109195	75804	134716
Verb	11088	13214	24169
Adjective	21460	18576	31184
Adverb	4607	3629	5748
Totals	146350	111223	195817

Polysemy information

WordNet
1.7.1的
规模

POS	Monosemous Words and Senses	Polysemous Words	Polysemous Senses
Noun	94685	14510	40002
Verb	5920	5168	18221
Adjective	15981	5479	15175
Adverb	3820	787	1900
Totals	120406	25944	75298

POS	Average Polysemy Including Monosemous Words	Average Polysemy Excluding Monosemous Words
Noun	1.23	2.75
Verb	2.17	3.52
Adjective	1.45	2.76
Adverb	1.24	2.41



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, hypernym, 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（这是框架语义学的方法）；
- 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中缺少不同词类词语间的关系（scholar – teacher -/- teach）；
- 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”



2 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）

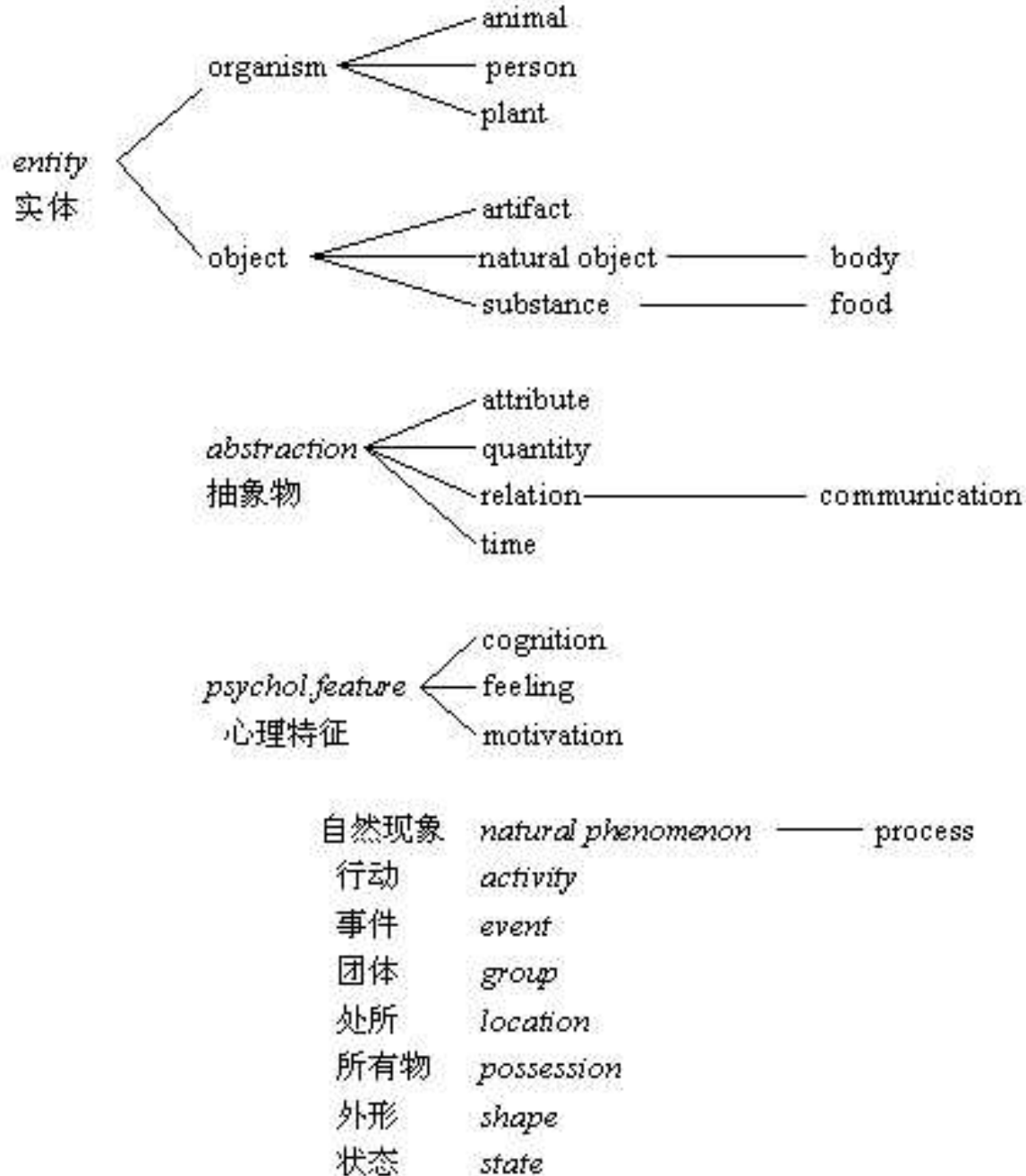
{act, activity} {food} {possession} {animal, fauna} {group, grouping}
{process} {artifact} ...

□ 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：

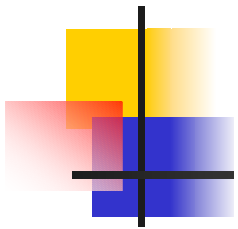
shetland pony @-> pony @-> horse @-> equid @-> odd-toed ungulate @->
placental mammal @-> mammal @-> vertebrate @-> chordate @-> animal
@-> organism @-> entity (12 levels)



名词的 分类树 (11棵)



/* 上图中 11 个斜体单词为基本类 */



Word	Polysemy
bronco	1
@-> mustang	1
@-> pony	
@-> horse	
@-> equine	0
@-> odd-toed ungulate	0
@-> placental mammal	0
@-> mammal	1
@-> vertebrate	1
@-> chordate	1
@-> animal	
@-> organism	
@-> entity	

多义性可以指示词语的熟悉度（Index of Familiarity）



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*

Collins A.M. and Quillian M.R., Retrieval Time From Semantic Memory, Journal of verbal learning and verbal behavior, 8, 240-248, 1969.

Smith, E. and Medin, D. (1981). Categories and Concepts. Cambridge, Mass.: Harvard University Press.



整体与部分关系(Meronymy)

- 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
 \Rightarrow the branch is a part of the forest



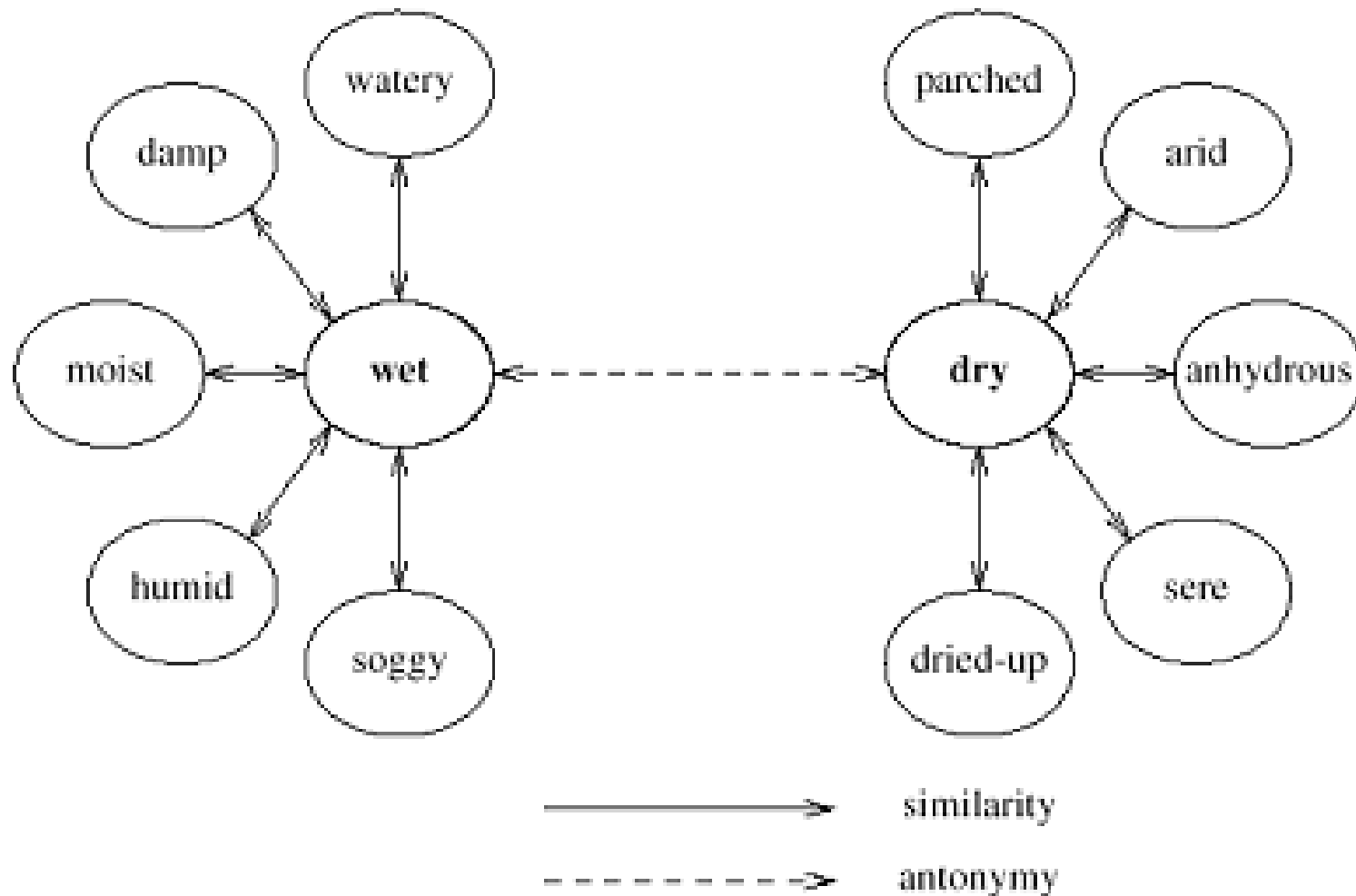
3 WordNet的形容词

- 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。

moist &-> wet !-> dry

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 (描写性形容词可以)
* *the extremely atomic bomb*
* *the very baseball game*
- 没有直接的反义词 (起分类作用, 而非否定作用)
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（兼属描写性和关系性）

- *old man* vs. *old house* 老教师 — ？ 旧教师 — 年轻教师
老房子 — 旧房子 — * 年轻房子
- *old friend* - *new friend* 老年朋友 — 青年朋友
old friend - *young friend* 老朋友 — 新朋友
- *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*



4 WordNet的动词

■ 英语动词的分类

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

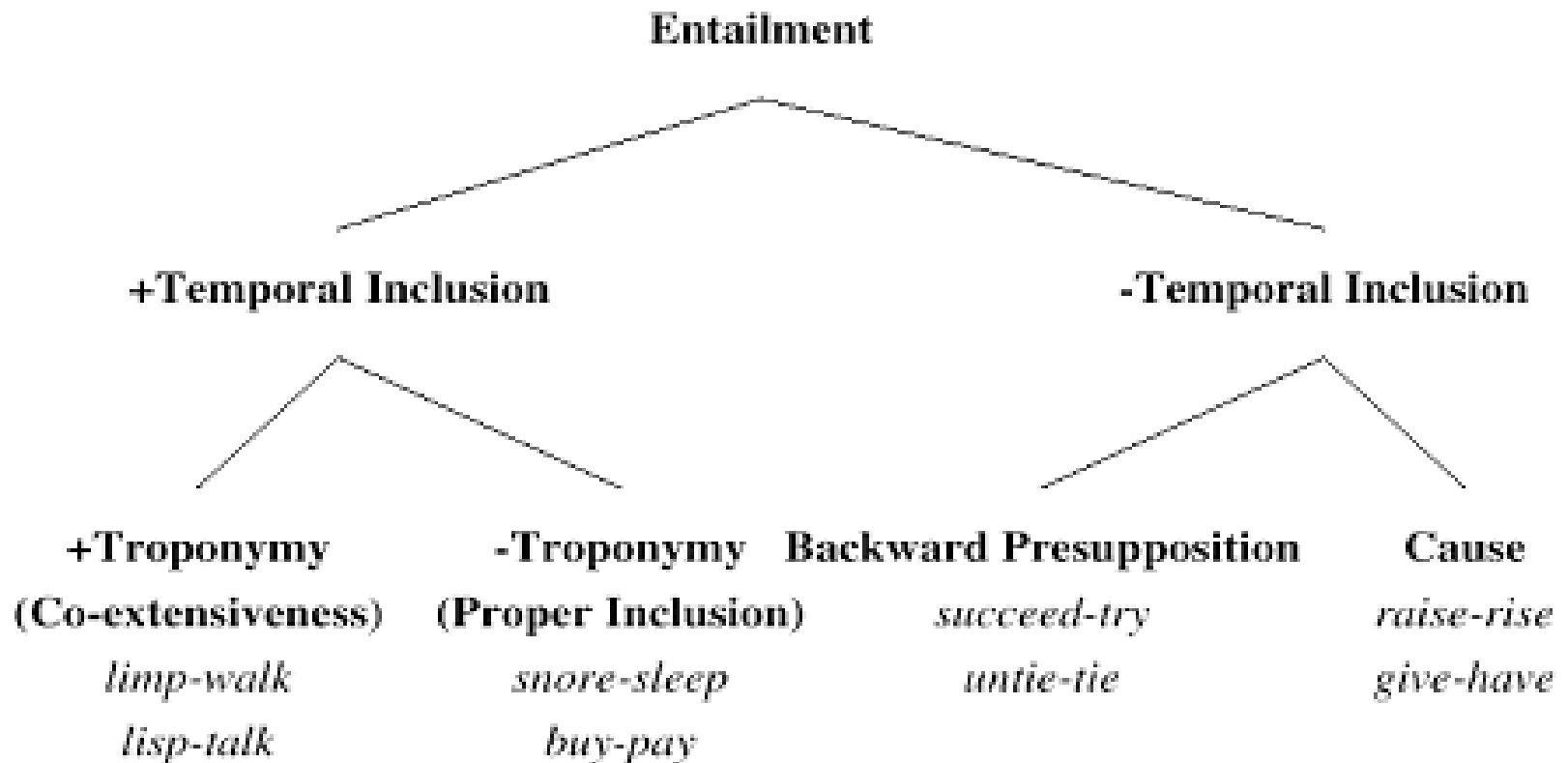
Pulman (1983): *be, do* = *activity, stative verb*

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



To V1 is to V2 in some particular manner

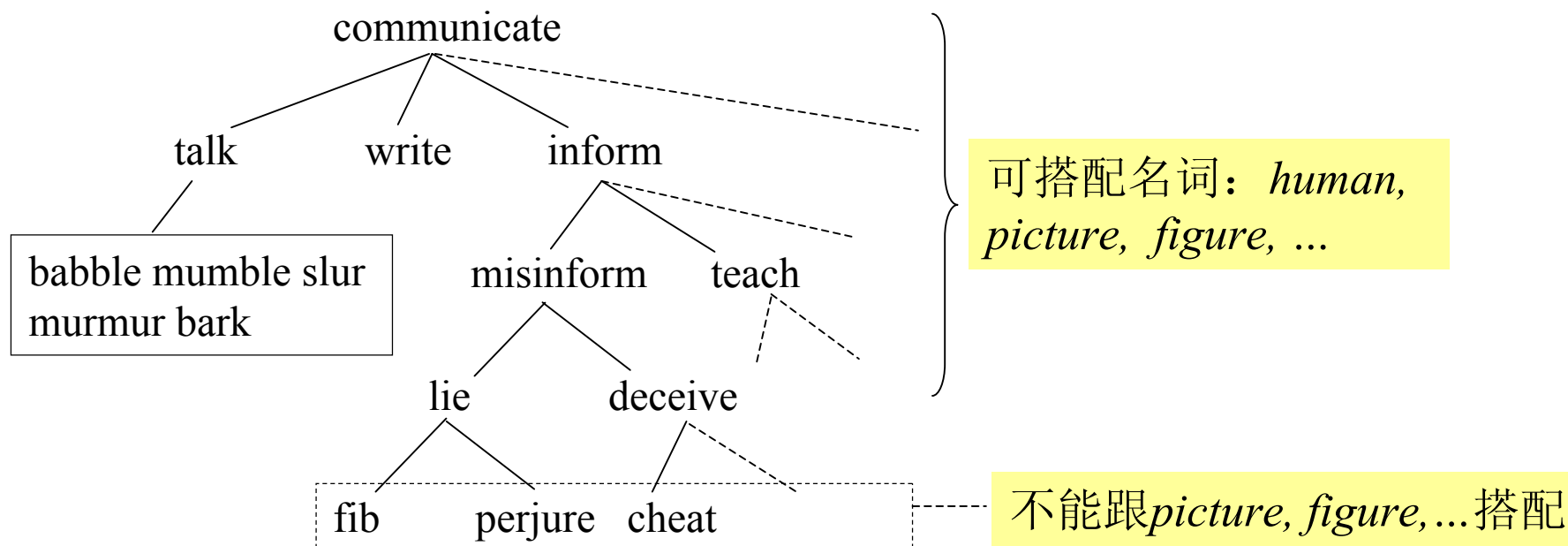


WordNet动词的反义关系

没有共同上位词，在
同一个场景中使用

- *give/take; buy/sell; lend/borrow; teach/learn*
- *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 @ → {travel, go, move,...}
- *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：（说、告诉） -> （隐瞒、欺骗） -> （欺诈、诈骗）



动词上下位层级在语法上的其他表现

- This vegetable **microwaves** (easily).
- This vegetable **cooks** * (quickly).

句子的现实性
(足量信息原则)

- This suitcase zips **shut** (in a flash).
- This suitcase **closes** * (easily).

- a **murdered** man — * a **killed** man
- an **altered** design — * a **changed** design
- a **divulged** secret — * a **told** secret



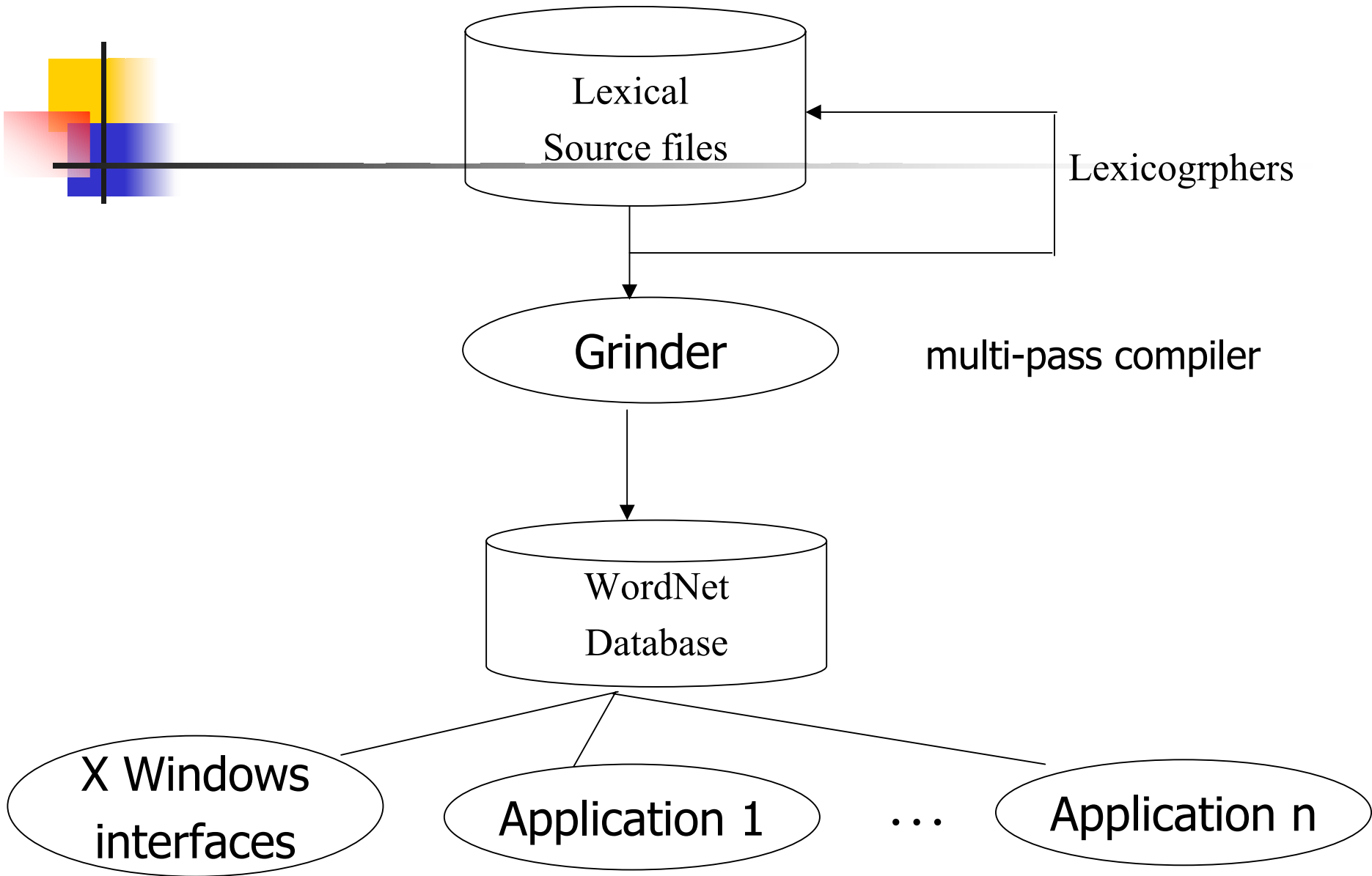
5 WordNet词库与查询软件的设计与实施

- 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（文本格式）；
 - 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WordNet词汇数据库的一套软件工具；
- 支持Unix, PC, Macintosh等多平台



WordNet软件工具的发展

- Bienkowski在1986年用LISP语言写了Grinder的第一个版本。
- Dan Teibel在1987年用C语言重写了Grinder程序。
- Antonio Romero在1989年又重写了Grinder程序，可以处理Glossary。
- Randee Teng从1991年开始负责管理该程序的所有这些版本。





WordNet文本数据库的格式

两类文件： X.dat X.idx (X=noun, adj, verb, adv)

SynSet_offset	lex_file_enum	ss_type	w_cnt	word	lex_id	p_cnt	ptr	frames	gloss
8位10进制数表示本概念的同义词集合在数据文件中的起始位置	2位10进制数表示文件编号（语义类编号）	词性标记	Synset中的词语个数	词语	1位16进制数表示一个词语的编号	3位10进制数表示关系指针个数	关系指针列表，包括指针符号，所对应概念的地址偏移，词性，关系类型等等	句型（仅对动词有意义）	注释及例句

WordNet文本数据库示例

上位关系

关系个数，000表示没有指向其他synset的指针

06922661 14 n 03 **date 0** **appointment 0** **engagement 0** 004 @
06877608 n 0000 ~ 06923070 n 0000 ~ 06923478 n 0000
%m 08202765 n **0000** | a meeting arranged in advance; "she
asked how to avoid kissing at the end of a date"

成员关系

分为00+00两部分，表示两个
synset之间的关系类型，0000表
示二者之间是语义关系

下位关系

SynSet_offset lex_filenum ss_type w_cnt word lex_id
[word lex_id...] p_cnt [ptr...] [frames...] | gloss

参见wn.db.5WN.html文件中的说明 #index file format #data file format

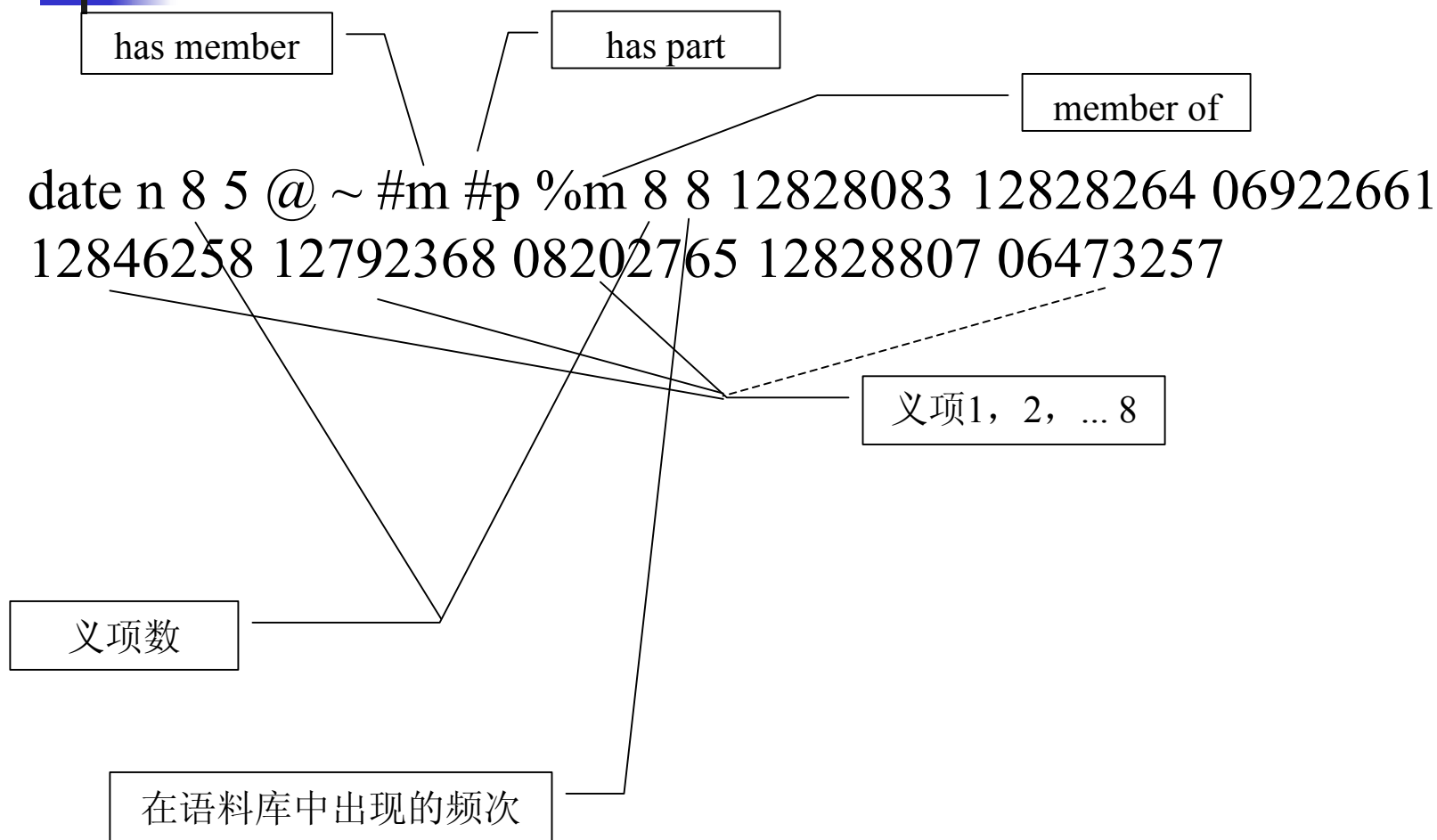


WordNet索引文件格式

lemma pos poly_cnt p_cnt [ptr_symbol...] sense_cnt tagsense_cnt synset_offset [synset_offset...]

lemma	pos	poly_cnt	p_cnt	ptr_symbol	sense_cnt	tagsense_cnt	synset_offset
词语	词性标记	10进制数 表示义项 个数	10进制 数表示 关系指 针个数	关系指针符 号	义项个数	频度	词语对 应的 synset在 wordnet 数据库 文件中的 偏移 地址

WordNet索引文件示例





6 WordNet的应用与发展

- 词义标注 (Word Sense Identification)
- 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
- 概念建模
-



利用WordNet进行文本校对

- G.Hirst & D. St-Onge, 1998, Lexical Chains as Representation of Context for the Detection and Correction of Malapropisms
- 原理
 - (1) 连贯的文本体现为Lexical Chains
 - (2) 根据WordNet中词语之间的各种关系可以计算文本中词语之间实际形成的Lexical Chains
 - (3) 如果有词语没有落在Lexical Chains中，则可能是错误
- 例子

Much of that data, he notes, is available toady/today electronically.



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, \dots , (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_i ,} * {or|and} NP_i $i \geq 1$

Output: for all NP_i , $i \geq 1$, HYPONYM(NP_i , NP_0)



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
- Global 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.



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、.....



参考文献和在线资料

- Christiane Fellbaum, ed., 1998, *WordNet : an electronic lexical database*, The MIT Press
- <http://www.cogsci.princeton.edu/~wn/>
- <http://www.hum.uva.nl/~ewn/>
- <http://ccl.pku.edu.cn/doubtfire/semantics/WordNet/C-wordnet/wordnet-c-index.htm>