第三章 词典

- 词典概述
- Hownet
- Wordnet
- NLTK的Wordnet调用

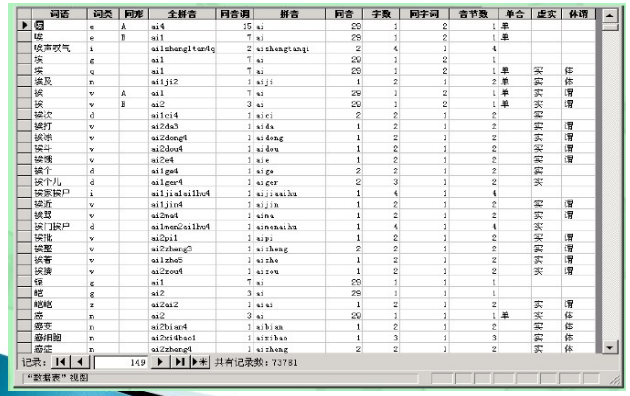
3.1 词典概述

- 词典学lexicology
  -Theory and description of lexical information
- 计算词典学computational lexicology
  - formal modellingof lexical information
- 词典编纂学lexicography
  -Construction of dictionaries (databases, handbooks)
- 计算词典编纂学computational lexicography
  - construction and production of dictionaries using electronic publishing

机读词典与人读词典

- 人读词典（Human Readable Dictionary）
  - 格式不规范
  - 数据完整性和一致性不好
  - -非结构化
- 机读词典（Machine Readable Dictionary）
  - 格式规范
    - 数据完整性和一致性较好
  - 结构化

汉语语法信息词典·总库



词语	词类	同形	全拼音	同音调	拼音	同音	字数	同字词	音节数	单合	虚实	体谓
爱	e	A	ai4	15	ai	29	1	2	1	单		
哎	e	B	ai1	7	ai	29	1	2	1	单		
哎声叹气	i		ai1shengltan4q	2	aishengtanqi	2	4	1	4			
挨	g		ai1	7	ai	29	1	2	1			
挨	q		ai1	7	ai	29	1	2	1	单	实	体
挨及	n		ai1ji2	1	aiji	1	2	1	2	单	实	体
挨	v	A	ai1	7	ai	29	1	2	1	单	实	谓
挨	v	B	ai2	3	ai	29	1	2	1	单	实	谓
挨次	d		ai1ci4	1	aici	2	2	1	2		实	
挨打	v		ai2da3	1	aida	1	2	1	2		实	谓
挨冻	v		ai2dong4	1	aidong	1	2	1	2		实	谓
挨饿	v		ai2e4	1	aie	1	2	1	2		实	谓
挨个	d		ai1ge4	1	aige	2	2	1	2		实	
挨个儿	d		ai1ger4	1	aiger	2	3	1	2		实	
挨家挨户	i		ai1jia1ai1hu4	1	aijiaaihu	1	4	1	4			
挨近	v		ai1jin4	1	aijin	1	2	1	2		实	谓
挨骂	v		ai2ma4	1	aima	1	2	1	2		实	谓
挨门挨户	d		ai1men2ai1hu4	1	aimenaihu	1	4	1	4		实	
挨批	v		ai2pi1	1	aipi	1	2	1	2		实	谓
挨整	v		ai2zheng3	1	aizheng	2	2	1	2		实	谓
挨着	v		ai1zhe5	1	aizhe	1	2	1	2		实	谓
挨揍	v		ai2zou4	1	aizou	1	2	1	2		实	谓
锿	g		ai1	7	ai	29	1	1	1			
皑	g		ai2	3	ai	29	1	1	1			
皑皑	z		ai2ai2	1	aiai	1	2	1	2		实	谓
癌	n		ai2	3	ai	29	1	1	1	单	实	体
癌变	n		ai2bian4	1	aibian	1	2	1	2		实	体
癌细胞	n		ai2xi4bao1	1	aixibao	1	3	1	3		实	体
癌症	n		ai2zheng4	1	aizheng	2	2	1	2		实	体

记录：14 149 共有记录数：73781

"数据表"视图

新华社词语数据库

全库分为中文和外文两个大类，主要包括中文新闻库、经济信息库、证券库、人物库、组织机构库、专题资料库等中文数据库，还包括XinhuaNews Bulletin 、Who's Who in China等英文数据库。共有28个库100多个子库，数据量达80多亿汉字，并以日均150万汉字的速度增长

新华社词语数据库·国际组织

- "２０００年问题"联合委员会/joint year 2000 council/ International
- "４·１９"运动/movement april19/ Colombia
- "阿尔法６６"/"alpha 66"/ Cuba
- "俄罗斯地区"社会联盟/regions of russiagroup/ Russia
- "法中－２０００年"协会/france-china association for the year 2000/ France
- "繁荣"党/prosperity/ Russia
- "光明的日本"国会议员联盟/parliamentary union for a bright japan/ Japan
- "基地"组织/al qaeda/ Saudi Arabia
- 《财富》杂志/fortune/ USA
- 《朝日新闻》/asahishimbun/ Japan
- 国际献血组织联合会/international federation of blood donor organizations/ International
- 国际宪法学协会/international association of constitutional law/ International
- 国际香料集团/international spice group/ International
- 经济和外贸部/ministry of economy and external trade of syria/ Syria
- 经济和外贸部/ministry of economy and foreign trade of egypt/ Egypt

3.2 知网(Hownet)

- 作者：董振东董强
- 网站：http://www.keenage.com
- 知网（英文名称为HowNet）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。
- 知网系统包括下列数据文件和程序：
  - (01) 知网管理系统
  - (02) 中英双语知识词典

知网(Hownet)

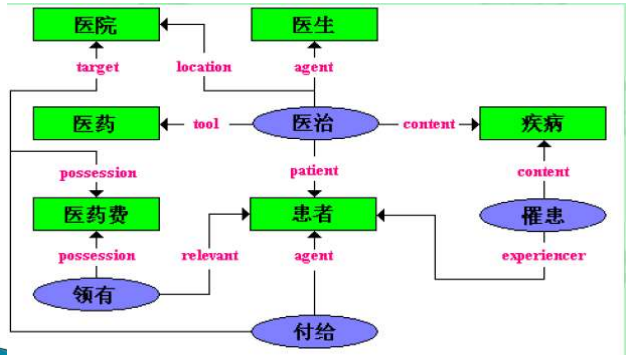
- 概念描述举例

  NO.=017144
  W_C=打
  G_C=V
  E_C=~网球，~牌，~秋千，~太极，球~得很棒
  W_E=play
  G_E=V
  E_E=

  DEF=exercise|锻练,sport|体育
- 其中DEF是核心，采用特定的“知识描述语言”

- 部件和属性，这两个基本单位在知网的哲学体系中占有着重要的地位。关于对部件的认识是：每一个事物都可能是另外一个事物的部件，同时每一个事物也可能是另外一个事物的整体
- 关于对属性的认识是：任何一个事物都一定包含着多种属性，事物之间的异或同是由属性决定的，没有了属性就没有了事物。
- 知网还着力要反映概念之间和概念的属性之间的各种关系。

知网（Hownet）5



- 总的来说，知网描述了下列各种关系：
  - (a) 上下位关系 (由概念的主要特征体现，请参看《知网管理工具》)
  - (b) 同义关系（可通过《同义、反义以及对义组的形成》获得）
  - (c) 反义关系（可通过《同义、反义以及对义组的形成》获得）
  - (d) 对义关系（可通过《同义、反义以及对义组的形成》获得）
  - (e) 部件-整体关系（由在整体前标注 % 体现，如"心"，"CPU"等）
  - (f) 属性-宿主关系（由在宿主前标注 & 体现，如"颜色"，"速度"等）
  - (g) 材料-成品关系（由在成品前标注 ? 体现，如"布"，"面粉"等）
  - (h) 施事/经验者/关系主体-事件关系（由在事件前标注 * 体现，如"医生"，"雇主"等）
  - (i) 受事/内容/领属物等-事件关系（由在事件前标注 $ 体现，如"患者"，"雇员"等）
  - (j) 工具-事件关系（由在事件前标注 * 体现，如"手表"，"计算机"等）
  - (k) 场所-事件关系（由在事件前标注 @ 体现，如"银行"，"医院"等）
  - (l) 时间-事件关系（由在事件前标注 @ 体现，如"假日"，"孕期"等）
  - (m) 值-属性关系（直接标注无须借助标识符，如"蓝"，"慢"等）
  - (n) 实体-值关系（直接标注无须借助标识符，如"矮子"，"傻瓜"等）
  - (o) 事件-角色关系（由加角色名体现，如"购物"，"盗墓"等）

知网（Hownet）2

- 义原是最基本的、不易于再分割的意义的最小单位
- 例:

治：医治 管理 处罚 „„
处：处在 处罚 处理 „„
理：处理 整理 理睬 „„

知网（Hownet）2

- 义原总数：1500多个
- 义原分类：共8类
- -基本义原
  - 事件、实体、次要特征
  - 属性、属性值、数量、数量值
  - 语法义原：描述语法特征，如POS
- 语法
  - -关系义原：描述意义关系，类似于格关系
- 动态角色
- 动态属性

知网（Hownet）4

- 义原的上下位关系构成树结构
- -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
…

3.3WordNet

- 网址：
  -http://www.cogsci.princeton.edu/~wn/
  http://wordnet.princeton.edu/
- 开发单位：
  -普林斯顿大学心理语言学实验室
  -初衷是作为研究人类词汇记忆的心理语言学成果
  -在自然语言处理中得到广泛的应用
- 免费的在线词汇数据库
- 世界很多语种都开发了相应的版本
  -各种欧洲语言：EuroNet
  -汉语：CCD（Chinese Concept Dictioanry）

WordNet发展概况

- 1978年，Miller描述了一种“自动化词典”（automated dictionary）的想法。
- 1985年，WordNet真正成为普林斯顿新成立的认知科学实验室几项研究计划中的一个，并开始实际运作。
- 1986年，Bienkowski用LISP语言写了Grinder的第一个版本。
- 20世纪70—90年代添加词表并对词进行分类。
- 1989年年初 WordNet从一个简单的“词典浏览器”（dictionary browser）发展成一个自足的词汇数据库（self-contained lexical database）。
- 1991年7月 WordNet 1.0正式公布，之后WordNet一系列版本发布，迄今最新版本为WordNet2.1版本。

WordNet词汇来源

- 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）。

二 WordNet中的词汇组织关系

- 同义词集（Sysnsets）
  WordNet将英语英语的名词、动词、形容词和副词组织为synsets,每一个Synset表示一个基本的词汇概念，并在这些概念之间建立了包括同义关系（synonymy）、反义关系（antonymy）、上下位关系（hypernymy&hyponymy）、部分关系（meronymy）等多种语义关系。

Example synset: {hit, strike, impinge on, run into, collide with}

Format of WordNetEntries

The noun “bass” has 8 senses in WordNet.
1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso[1] - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

The adjective “bass” has 1 sense in WordNet.
1. bass[1], deep[6] - (having or denoting a low vocal or instrumental range)
”a deep voice”; ”a bass voice is lower than a baritone voice”; ”a bass clarinet”

WordNet中词语间的关系

- 主要的词汇关系
- 同义关系（构成Synsets）
- 反义关系（指针！）
- 上位关系（指针@）
- 下位关系（指针~）

- 整体关系（名词、指针#m/#s/#p）
- 部分关系（名词、指针%m/%s/%p）
- 蕴含关系（动词、指针*）
- 因果关系（动词、指针>）
- 近似关系（形容词、指针&）

注：形容词如果是动词分词，用指针（<）指向该动词；

副词如果由形容词的派生而来，用指针（\）指向。

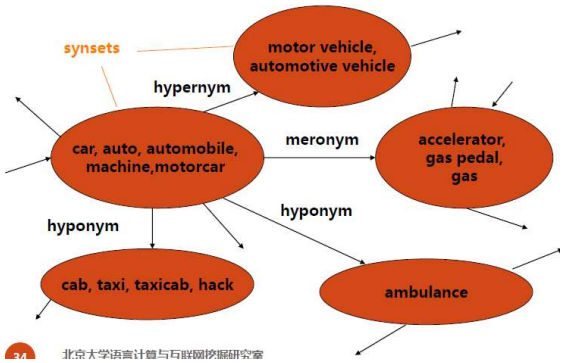
WordNetNoun Relations

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} → *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} → *lunch*^{1}
Member Meronym	Has-Member	From groups to their members	*faculty*^{2} → *professor*^{1}
Has-Instance		From concepts to instances of the concept	*composer*^{1} → *Bach*^{1}
Instance		From instances to their concepts	*Austen*^{1} → *author*^{1}
Member Holonym	Member-Of	From members to their groups	*copilot*^{1} → *crew*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} → *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} → *meal*^{1}
Antonym		Opposites	*leader*^{1} → *follower*^{1}

WordNetVerb Relations

Relation	Definition	Example
Hypernym	From events to superordinate events	*fly*^{9} → *travel*^{5}
Troponym	From a verb (event) to a specific manner elaboration of that verb	*walk*^{1} → *stroll*^{1}
Entails	From verbs (events) to the verbs (events) they entail	*snore*^{1} → *sleep*^{1}
Antonym	Opposites	*increase*^{1} ⟺ *decrease*^{1}

A WordNetSnapshot



WordNet3.0 Statistics

Number of words, synsets, and senses

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Polysemy information

POS	Monosemous Words and Senses	Polysemous Words
Noun	101863	15935
Verb	6277	5252
Adjective	16503	4976
Adverb	3748	733
Totals	128391	26896

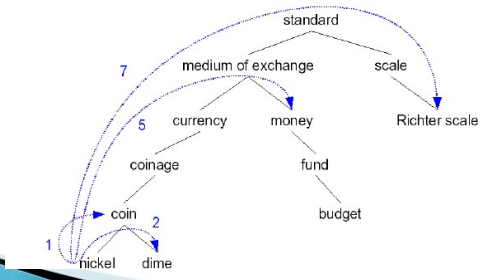
POS	Average Polysemy Including Monosemous Words	Average Polysemy Excluding Monosemous Words
Noun	1.24	2.79
Verb	2.17	3.57
Adjective	1.40	2.71
Adverb	1.25	2.50

WordNet-based Word Similarity

- 可以使用**WordNet的任意信息**
  - Relation
  - Glosses
  - Example sentences
- **Word similarity vs. word relatedness**
  - **Similar words are near-synonyms**
    - Car, bicycle: similar
    - Related could be related any way
- **Car, gasoline: related, not similar**

Path based similarity

- 两个词在词典层次结构中越相邻，这两个词越相似 **(i.e.具有比较短的路径)**



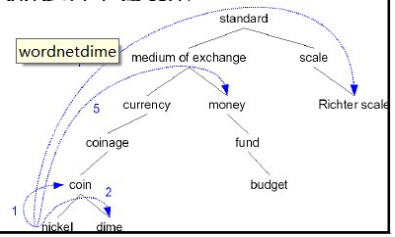
Path-based similarity的改进

- pathlen(c1,c2)= 词义节点c1 and c2之间最短路径上边的数量
- simpath(c1,c2) = -log pathlen(c1,c2)
- wordsim(w1,w2)=

$$\max_{c1 \in senses(w1), c2 \in senses(w2)} sim(c1,c2)$$

Path-based similarity的问题

- 假设每条链接(边）表示同样的距离
  - 基于Path-based similarity, Nickelto money与nickelto standard 具有相同的相似度
  - 然而，Nickel to money看起来应该比nickelto standard 更相似
- 因此，需要对每条边的代价讲行单独表示



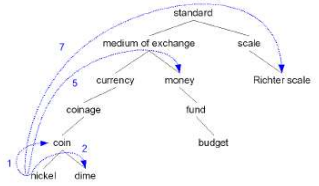
Information content similarity metrics

- 定义P(C)：
- 从一个语料库中随机选择一个词，这个词属于概念C的概率
- P(root)=1
- 在词典层次结构中，一个概念节点位置越低，那么相应的概率也越低

Information content similarity

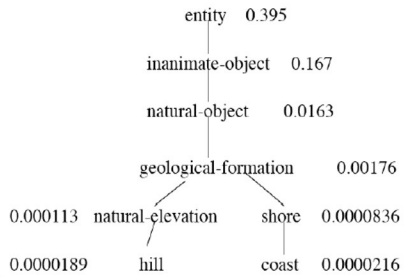
- 基于语料库进行统计
  - “dime” 的出现应该被*coin, currency, standard*等词的频率所统计
  - words(c): 概念c所包容的词集（包含子孙后代节点）
  - N: 词语总数

$$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$$



Information content similarity

- WordNet结构被赋予概率P(C)



Information content: definitions

- Information content:
  - $IC(c) = -\log P(c)$
- Lowest common subsumer LCS(c1,c2)
  - The lowest node in the hierarchy that subsumes (is a hypernym of) both c1 and c2

Resnikmethod

- Resnik:衡量两个词的共性为
  - 两个词节点的最低共同祖先节点的信息内容(info content)
  - 公共包容[illegible]大

$$sim_{resnik}(c1,c2) = -\log P(LCS(c1,c2))$$

Lin's Method

- **SimLin(A,B)=common(A,B)/description(A,B)**
- $Sim_{Lin}(c1,c2) = 2 \log P(LCS(c1,c2)) / (\log P(c1) + \log P(c2))$
  - $Sim_{Lin}(hill,coast) = 2\log P(geological\text{-}formation)) / (\log P(hill) + \log P(coast)) = .59$

3.4基于NLTK的Wordnet接口

- >>> from nltk.corpus import wordnet as wn

```
>>> wn.synsets('dog')
[Synset('dog.n.01'), Synset('frump.n.01'), Synset('dog.n.03'),
Synset('cad.n.01'),Synset('frank.n.02'), Synset('pawl.n.01'),
Synset('andiron.n.01'), Synset('chase.v.01')]
>>> wn.synsets('dog', pos=wn.VERB)
[Synset('chase.v.01')]
```

pos可为：NOUN、VERB、ADJ、ADV...

```
>>> wn.synset('dog.n.01')
>>> Synset('dog.n.01')                    #指出同义词集
>>> wn.synset('dog.n.01').definition #指出同义词集的定义
'a member of the genus Canis (probably descended from the common
wolf) that has been domesticated by man since prehistoric times;
occurs in many breeds'
>>> wn.synset('dog.n.01').examples   #举一个同义词集的一个例子
['the dog barked all night']
##词条集
>>> wn.synset('dog.n.01').lemmas     #列出一个同义词集的所有词条
[Lemma('dog.n.01.dog'), Lemma('dog.n.01.domestic_dog'),
Lemma('dog.n.01.Canis_familiaris')]
>>> wn.synset('dog.n.01').lemma_names #列出一个同义词集的所有词条的名字
['dog', 'domestic_dog', 'Canis_familiaris']
##词条
>>> wn.lemma('dog.n.01.dog').synset   #指出词条所属的同义词集
Synset('dog.n.01')
>>> wn.lemma('dog.n.01.dog').name
'dog'
##综合举例
>>> [lemma.name for lemma in wn.synset('dog.n.01').lemmas]
['dog', 'domestic_dog', 'Canis_familiaris']
```

```
>>> dog = wn.synset('dog.n.01')
>>> dog.hypernyms()        #上位词集合
[Synset('domestic_animal.n.01'), Synset('canine.n.02')]
>>> dog.root_hypernyms() #得一个最一般的上位（或根上位）同义词集
[Synset('entity.n.01')]
>>> dog.hyponyms()         #下位词集合
[Synset('puppy.n.01'), Synset('great_pyrenees.n.01'),
Synset('basenji.n.01'), Synset('newfoundland.n.01'),
Synset('lapdog.n.01'), Synset('poodle.n.01'),
Synset('leonberg.n.01'), Synset('toy_dog.n.01'),
Synset('spitz.n.01'), Synset('pooch.n.01'), Synset('cur.n.01'),
Synset('mexican_hairless.n.01'), Synset('hunting_dog.n.01'),
Synset('working_dog.n.01'), Synset('dalmatian.n.02'),
Synset('pug.n.01'), Synset('corgi.n.01'),
Synset('griffon.n.02')]
```

```
>>> dog.member_holonyms()    #从物品到它们的部件（部分）或到
它们被包含其中的东西（整体）。例如：一棵树的部分是它的树干，树冠等
[Synset('pack.n.06'), Synset('canis.n.01')]
```

```
>>> dog = wn.synset('dog.n.01')
>>> cat = wn.synset('cat.n.01')
```

```
>>> dog.path_similarity(cat)
0.2000000000000001
>>> cat.path_similarity(cat)
1.0
```