

# 词汇语义计算

车万翔

社会计算与信息检索研究中心

2017年春季学期

# 词汇语义计算概述

# 词汇语义

- 研究词语的意义
  - 如何表示词语的意义？
  - 词语之间是如何关联的？
    - 同义词、反义词、上位词、下位词等

# 词语 ( Word )

- 中文词
  - 单字词、多字词
  - 部分词的意义由字的意义组成
  - 一般不具有形态变化
- 英文词
  - 形态变化较为丰富
  - 由词的标准/基本形态 ( lemma ) 变化为多种形态 ( inflected forms )
    - get → gets, got, gotten, getting
- 本课程以英文词为例

# ( 英语 ) 词语形态规范化

- 词语形态规范化
  - 如何匹配 company 与 companies ? sell 与 sold ?
  - 删除词语的形态信息：时态、数量等
- 词根 ( stemming )
  - 删除后缀：ed, ing, ational, ation, able, ism, ...
  - 基于规则的方法 ( 如：Porter's stemmer )
  - Stemming 的结果可能不是词
    - 如：query, queries, querying → queri
  - 不相关的词可能具有相同的 stem
    - 如：police, policy → polic

# ( 英语 ) 词语形态规范化

- 词形还原 ( Lemmatization )
  - 将词语变为其语法原型 ( syntactic stem )
    - 如 : agreements → agreement
  - 使用一般规则与例外处理
    - 如 : ies → y, ed → ∅, s → ∅
    - sought → seek, sheep → sheep, feet → foot
  - 处理结果仍为词
  - 处理过程要考虑词性的不同
    - thought → think 如果 thought 是动词
    - thought → thought 如果是名词

# 词义 ( Word Senses )

- 一个词语的特定意义
- 一个词语可能有多个词义
- 一个词义能被一个注释 ( gloss ) 所描述
  - apple: fruit with red or yellow or green skin and sweet to tart crisp whitish flesh
- 一词多义
  - homonyms: 词义完全不相关
    - bank: money bank, river bank
  - polysemes: 词义之间有关联
    - Bank: financial institute, building of the financial institute, storage of blood (blood bank)
  - 两者之间界限模糊

# 一个词有多少意义？

- 难以回答，比较主观

- Drive the car
- Drive to school
- Drive me mad





# 一个词有多少意义？

- 不同词典和不同人对一个词的意义数量会有不同看法
- 通常词典和语言资源会给出一个词的细粒度的意义，但对于很多NLP任务来说可能并不需要
  - 对于单词 drive，WordNet 有 34 意义

# 词义基本关系

- 同义 (Synonymy)
- 反义 (Antonymy)
- 上位 (Hypernymy)
- 下位 (Hyponymy)
- 整体 (Holonymy)
- 部分 (Meronymy)

# 同义 (Synonym)

- 两个词的两个词义相同或接近相同
  - 如：buy 和 purchase
- 可用代入法检测
  - I bought/purchased a car.
- 不存在完美的同义词，同义词可能在某些上下文中有所不同
  - 如：water 和 H<sub>2</sub>O

# 反义 (Antonym)

- 词义相反
  - 如 : long/short, rise/fall
- 尽管反义词具有相反的意义，但它们在某种角度仍非常相似，具有一定的共性
  - 如 : long and short are degree of lengths
- 利用基于语料库的上下文相似性度量难以区分同义词与反义词
  - This is good.
  - This is nice.
  - This is bad.

# 上位 (Hypernym) 与下位(Hyponym)

- Hyponyms: Y is a hyponym of X if every Y is a (kind of) X
  - 一个词的词义比另一个词的词义更加具体
    - 如 : *apple* is a hyponym of *fruit*
- Hypernyms: Y is a hypernym of X if every X is a (kind of) Y
  - Opposite of hyponym, e.g. *fruit* is a hypernym of *apple*

# 部分 (Meronym)与整体 (Holonym)

- Meronyms: Y is a meronym of X if Y is a part of X
  - Part-whole relation
    - 如 : *wheel* is a meronym of *car*
- Holonyms: Y is a holonym of X if X is a part of Y
  - Opposite of meronyms
    - 如 : *car* is a holonym of *wheel*

# 词义关系在信息检索等领域中的作用

- 查询扩展与智能匹配
  - 同义词: 哈工大 vs. 哈尔滨工业大学

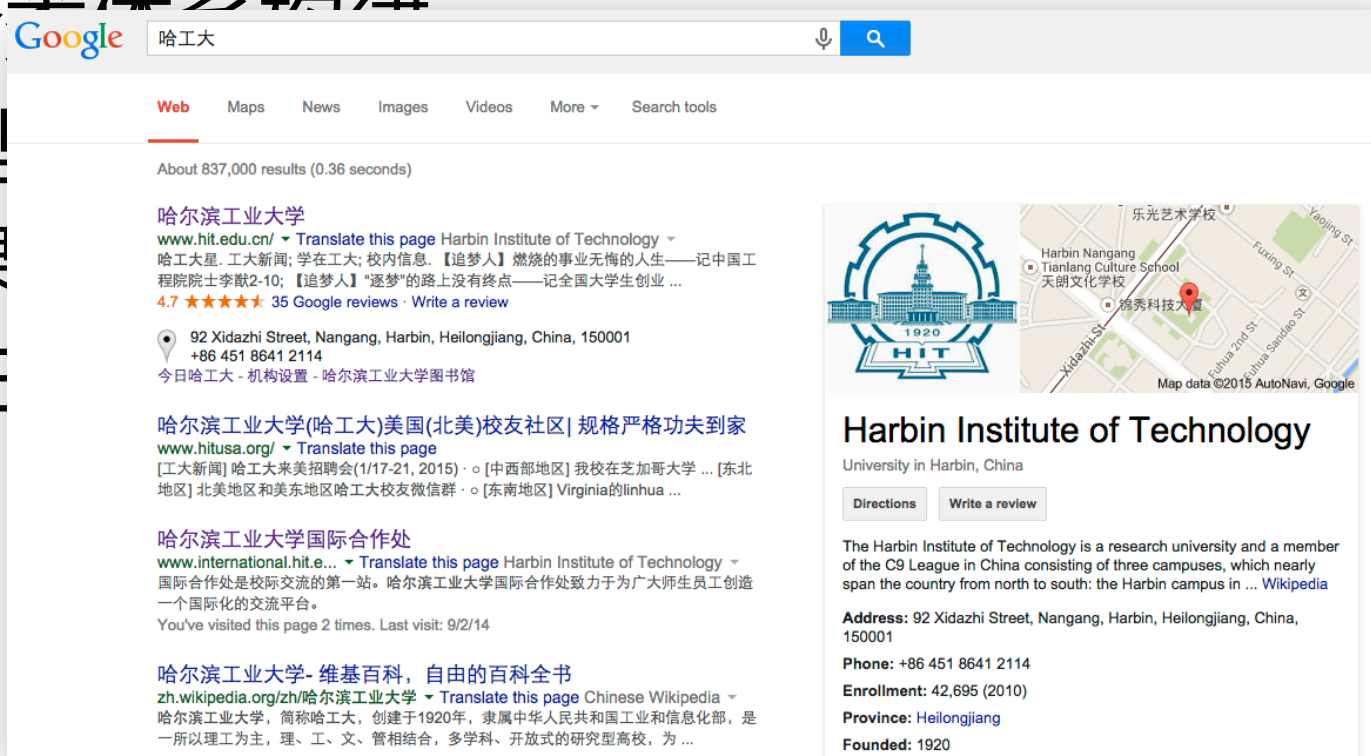
- 知识/分

- 文本推理

- 吃水果

- 图像标注

- ...



The screenshot shows a Google search for '哈工大' (Harbin Institute of Technology). The search bar at the top contains the text '哈工大'. Below the search bar, the results are displayed. The first result is for '哈尔滨工业大学' (Harbin Institute of Technology), with the URL 'www.hit.edu.cn/' and a brief description. The second result is for '哈尔滨工业大学(哈工大)美国(北美)校友社区', with the URL 'www.hitusa.org/'. The third result is for '哈尔滨工业大学国际合作处', with the URL 'www.international.hit.edu.cn/'. The fourth result is for '哈尔滨工业大学- 维基百科, 自由的百科全书', with the URL 'zh.wikipedia.org/zh/哈尔滨工业大学'. On the right side of the search results, there is a map showing the location of Harbin Institute of Technology in Harbin, China. The map includes a street view and a list of nearby locations, such as 'Harbin Nangang', 'Tianlang Culture School', and '锦秀科技大厦'. Below the map, the text 'Harbin Institute of Technology' is displayed, along with the address '92 Xidazhi Street, Nangang, Harbin, Heilongjiang, China, 150001' and the phone number '+86 451 8641 2114'. The text also mentions that the university is a member of the C9 League in China and provides information about its enrollment and founding year.

# 词汇相似度 (Word Similarity)

- 同义词关系是二值关系
  - 两个词是/不是同义关系
- 更宽松的准则
  - 词汇相似度/语义距离 (Word similarity or word semantic distance)
- 两个词之间具有越多的共性越相似
- 实际上是基于词义的关系
- 可以基于词义和词进行计算



# 词语相似度两类计算方法

- 基于语义词典的方法 (Thesaurus-based)
  - 基于两个词在 WordNet 等语义词典中是否 “相邻”
- 基于语料统计的方法  
(Distributional/statistical algorithms)
  - 比较词语在语料库中的上下文

# 基于语义词典的词汇语义计算

# WordNet

- 著名的英文词义关系计算资源，词义数据库
  - 包含词义及其关系
- 免费浏览和下载
  - <http://wordnet.princeton.edu/>
- Developed in the mid-1980s by famous cognitive psychologist George Miller and a team at Princeton University
  - George A. Miller passed away on July 22, 2012 at the age of 92.

# WordNet

- Synset (synonym set): (近似)同义集合
  - WordNet的基本单元
  - 每一个synset表示一个语义概念
  - 如: {hit, strike, impinge on, run into, collide with}
- 每个词条包括多个 synsets , 注释 , 使用样例等信息
- Synsets 通过不同的词义关系相连
- 四个词性类别
  - Nouns、Verbs、Adjectives、Adverbs

# WordNet 示例

The noun “bass” has 8 senses in WordNet.

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

The adjective “bass” has 1 sense in WordNet.

1. bass<sup>1</sup>, deep<sup>6</sup> - (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 中的语义关系

- A semantic relation is represented by a pointer between word forms or between synsets.

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponymy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry
<i>Note: N = Nouns Aj = Adjectives V = Verbs Av = Adverbs</i>		

# WordNet Noun Relations

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

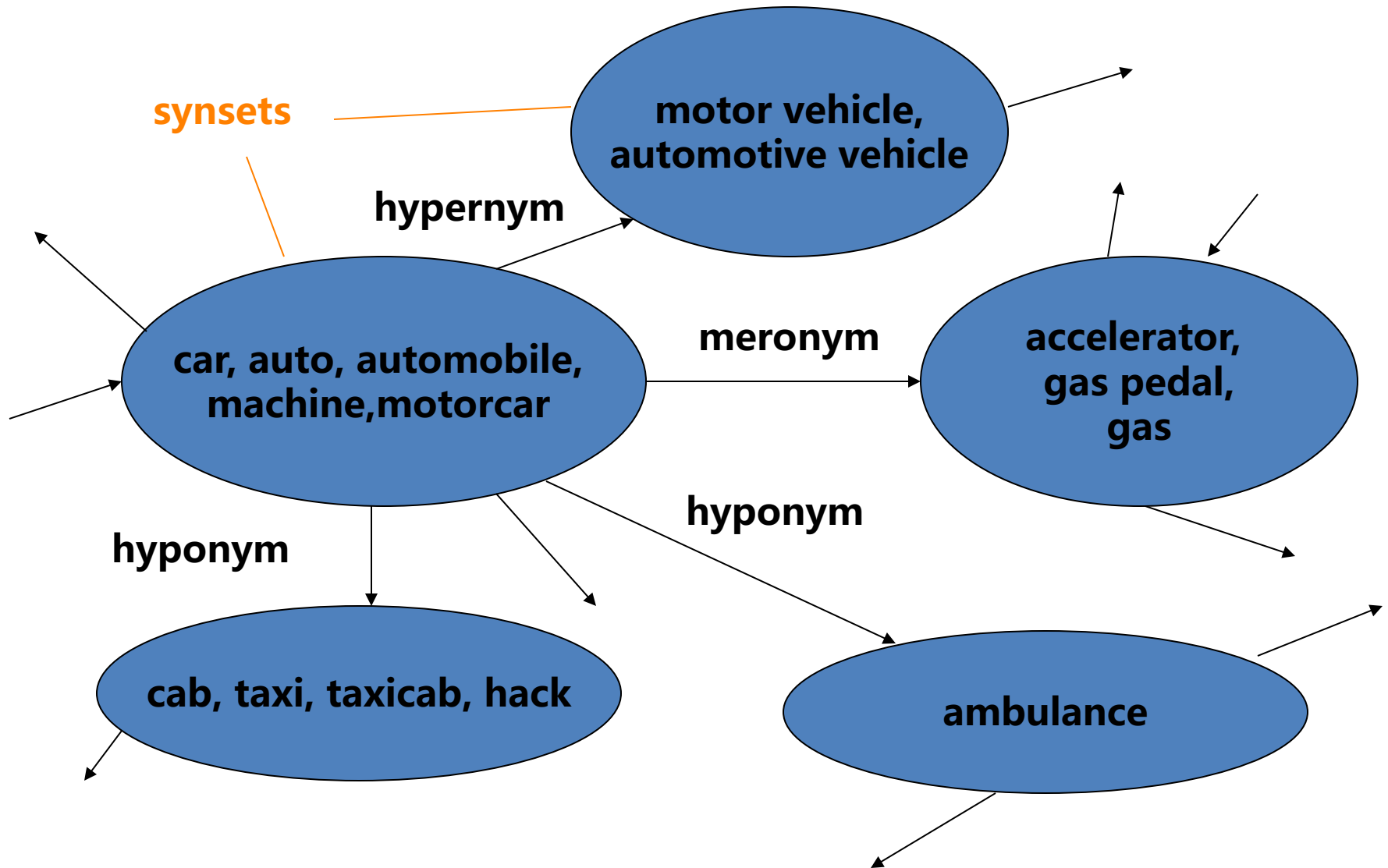
# WordNet Verb Relations

Relation	Definition	Example
Hypernym	From events to superordinate events	<i>fly</i> <sup>9</sup> → <i>travel</i> <sup>5</sup>
Troponym	From a verb (event) to a specific manner elaboration of that verb	<i>walk</i> <sup>1</sup> → <i>stroll</i> <sup>1</sup>
Entails	From verbs (events) to the verbs (events) they entail	<i>snore</i> <sup>1</sup> → <i>sleep</i> <sup>1</sup>
Antonym	Opposites	<i>increase</i> <sup>1</sup> ⇔ <i>decrease</i> <sup>1</sup>



# A WordNet Snapshot

**synsets**



# WordNet Hierarchies

Sense 3

bass, basso --

(an adult male singer with the lowest voice)

=> singer, vocalist, vocalizer, vocaliser

=> musician, instrumentalist, player

=> performer, performing artist

=> entertainer

=> person, individual, someone...

=> organism, being

=> living thing, animate thing,

=> whole, unit

=> object, physical object

=> physical entity

=> entity

=> causal agent, cause, causal agency

=> physical entity

=> entity

Sense 7

bass --

(the member with the lowest range of a family of musical instruments)

=> musical instrument, instrument

=> device

=> instrumentality, instrumentation

=> artifact, artefact

=> whole, unit

=> object, physical object

=> physical entity

=> entity

# WordNet 3.0 Statistics

## Number of words, synsets, and senses

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

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

# WordNets for Other Languages

- EuroWordNet
  - Individual WordNets for some European languages (Dutch, Italian, Spanish, German, French, Czech, and Estonia) which are also interconnected by interlingual links  
<http://www.ilc.uva.nl/EuroWordNet/>
- WordNets for some asian languages
  - Hindi
    - <http://www.cfilt.iitb.ac.in/wordnet/webhwn/>
  - Marathi
    - <http://www.cfilt.iitb.ac.in/wordnet/webmwn/>
  - Japanese
    - <http://nlpwww.nict.go.jp/wn-ja/index.en.html>

# WordNet Senses

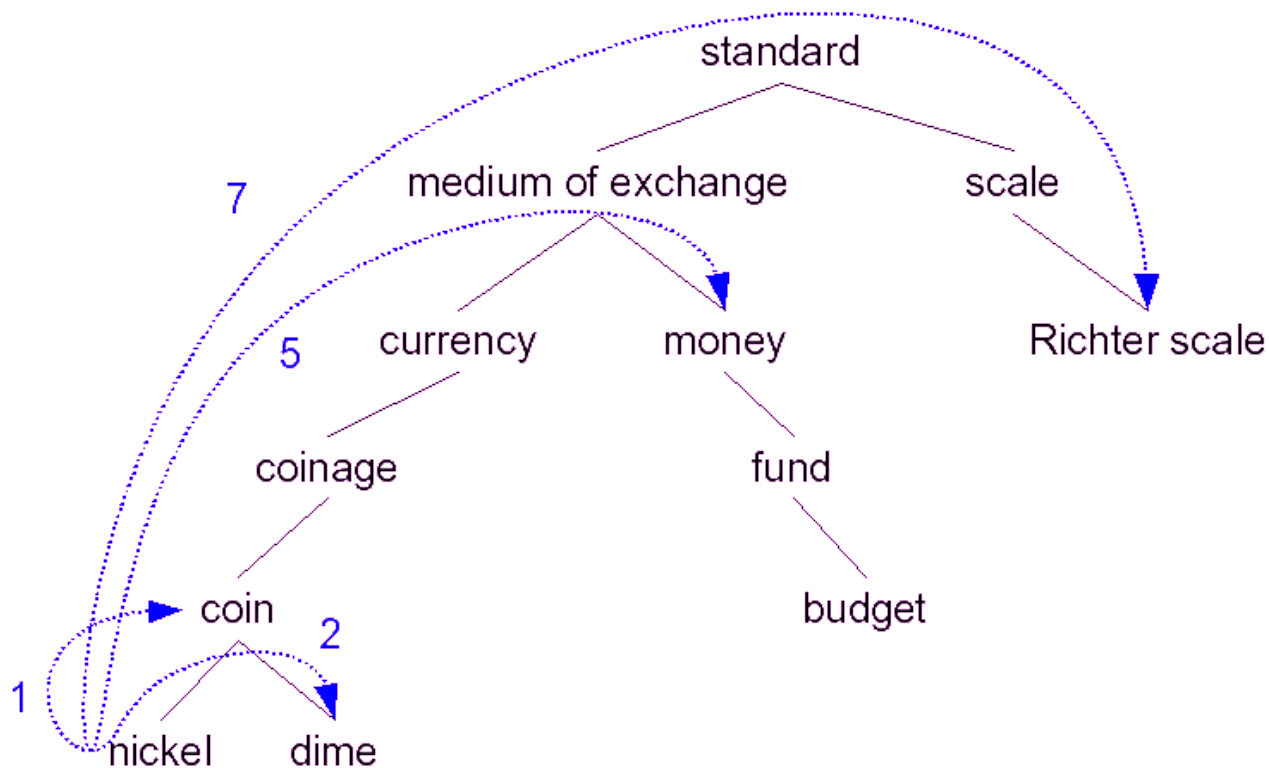
- WordNets senses倾向于细粒度
  - “play” as a verb has 35 senses, including
    - play a role or part: “Gielgud played Hamlet”
    - pretend to have certain qualities or state of mind: “John played dead.”
- 人机都难以进行细粒度区分，只有词汇学专家才能有效区分
- 细粒度词义是否对NLP任务有用？
  - 不一定
- 可以考虑对细粒度词义进行归并，得到粗粒度、易于区分的词义

# 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

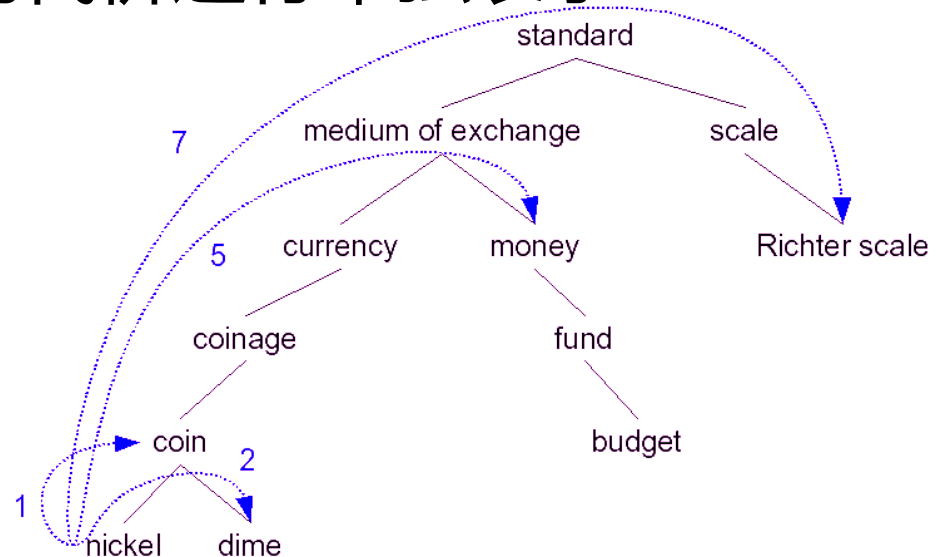
# Path-based Similarity

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



# 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, *nickel to money* 与 *nickel to standard* 具有相同的相似度
  - 然而, *Nickel to money* 看起来应该比 *nickel to standard* 更相似
- 因此, 需要对每条边的代价进行单独表示





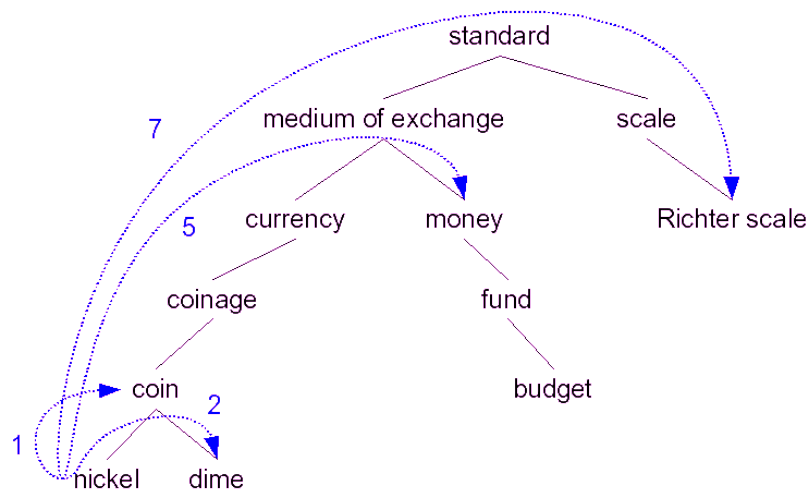
# Information Content Similarity Metrics

- 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
  - $P(\text{root})=1$
  - 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

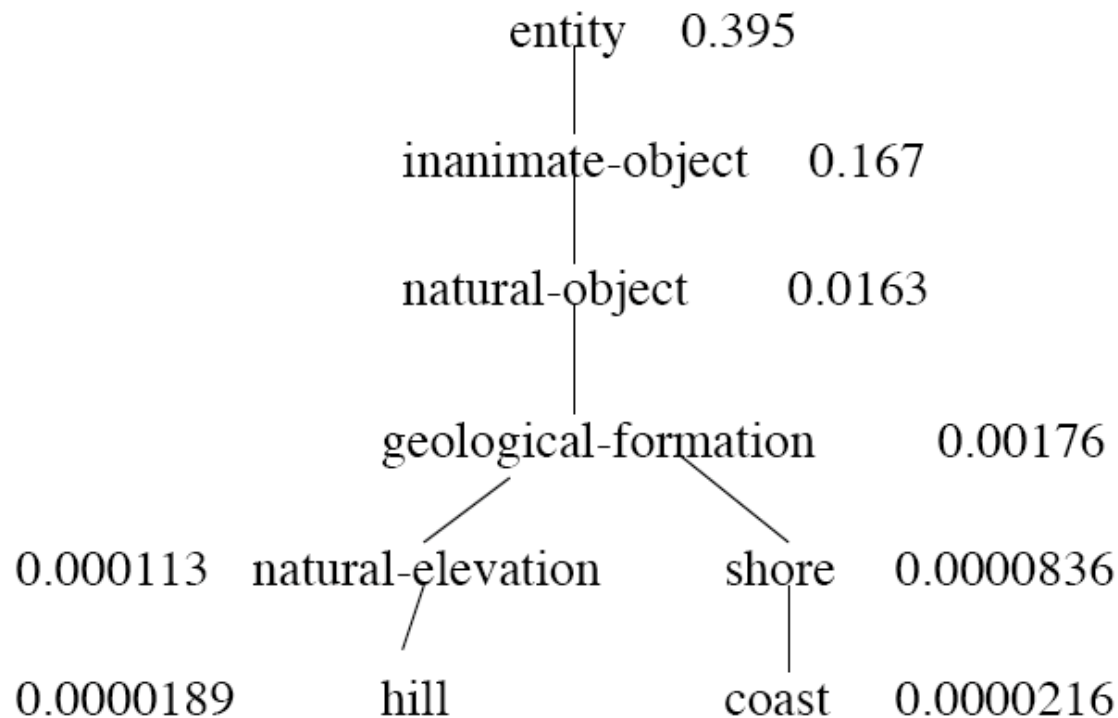
- 基于语料库进行统计
  - “dime” 的出现应该被coin, currency, standard等词的频率所统计
  - words(c): 概念c所包容的词集 ( 包含子孙后代节点 )
  - N: 词语总数

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



# Information Content Similarity

- WordNet结构被赋予概率 $P(C)$



# Information Content 定义

- 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  $c2$

# Resnik Method

- 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)
  - $\text{sim}_{\text{resnik}}(c1, c2) = -\log P(\text{LCS}(c1, c2))$
  - 公共包容节点在层次结构中位置越低，相似性越大

# Lin's Method

- $\text{Sim}_{\text{Lin}}(c1, c2) = \frac{2 \log P(\text{LCS}(c1, c2))}{\log P(c1) + \log P(c2)}$ 
  - $\text{Sim}_{\text{Lin}}(\text{hill}, \text{coast}) = \frac{2 \log P(\text{geological-formation})}{\log P(\text{hill}) + \log P(\text{coast})} = .59$

# Jiang-Conrath Method

- $\text{Dis}_{\text{JC}}(c1, c2) = 2 \log P(\text{LCS}(c1, c2)) - (\log P(c1) + \log P(c2))$
- $\text{Sim}_{\text{JC}}(c1, c2) = 1 / \text{Dis}_{\text{JC}}(c1, c2)$

# Extended Lesk

- 两个概念的注释中包含越多的相似词语，它们越相似
  - *Drawing paper*: paper that is specialy prepared for use in drafting
  - *Decal*: the art of transferring designs from specialy prepared paper to a wood or glass or metal surface
- 对于共同出现的n个词组成的词组，加上值 $n^2$ 
  - paper and specialy prepared for  $1 + 4 = 5$
- 需要考虑的 WordNet 关系，基于此关系得到的其他词的 gloss 作为扩充进行比较

$$\text{sim}_{\text{eLesk}}(c_1, c_2) = \sum_{r, q \in \text{RELS}} \text{overlap}(\text{gloss}(r(c_1)), \text{gloss}(q(c_2)))$$



# Summary

$$\text{sim}_{\text{path}}(c_1, c_2) = -\log \text{pathlen}(c_1, c_2)$$

$$\text{sim}_{\text{Resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2))$$

$$\text{sim}_{\text{Lin}}(c_1, c_2) = \frac{2 \times \log P(\text{LCS}(c_1, c_2))}{\log P(c_1) + \log P(c_2)}$$

$$\text{sim}_{\text{jc}}(c_1, c_2) = \frac{1}{2 \times \log P(\text{LCS}(c_1, c_2)) - (\log P(c_1) + \log P(c_2))}$$

$$\text{sim}_{\text{eLesk}}(c_1, c_2) = \sum_{r, q \in \text{RELS}} \text{overlap}(\text{gloss}(r(c_1)), \text{gloss}(q(c_2)))$$

# 实验结果（与人对比）

- **Miller and Charles**: 抽取30词对 (10 from high level = 3-4, 10 from intermediate level = 1-3 and 10 from low level 0-1).
- **Rubenstein and Goodenough**: 65个词对，包含高度同义词对与语义不相关词对，词对相似性值由人工标注，范围在 0.0到4.0.

<i>Similarity measure</i>	M&C	R&G
Hirst and St-Onge ( $rel_{HS}$ )	.744	.786
Leacock and Chodorow ( $sim_{LC}$ )	.816	.838
Resnik ( $sim_R$ )	.774	.779
Jiang and Conrath ( $dist_{JC}$ )	.850	.781
Lin ( $sim_L$ )	.829	.819

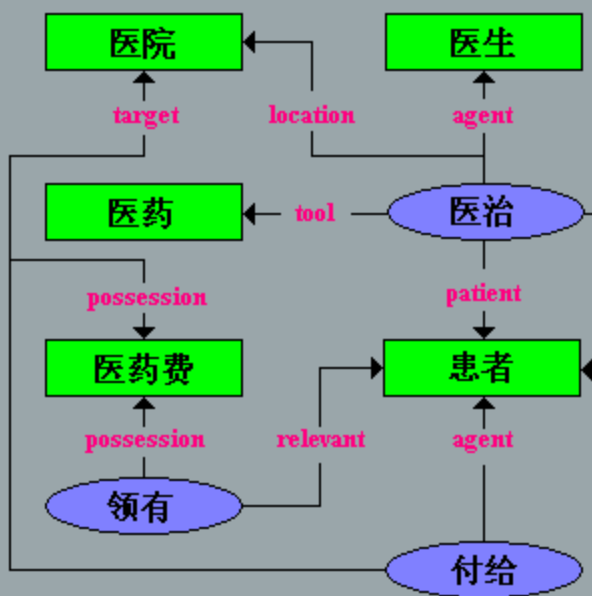
# 软件包

- WordNet::Similarity
  - <http://wn-similarity.sourceforge.net/>

# 中文语义词典

- 同义词词林
  - 总词汇量仅5.3万多
  - 哈工大扩展版包含77,458条词语
- 中文概念辞书(CCD)
  - 基于WordNet框架
- 知网 (HowNet)
  - 揭示概念与概念之间以及概念所具有的属性之间的关系

# 知网 ( HowNet )



(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)

(p) 相关关系 (由在相关概念前标注 # 体现, 如“谷物”, “煤田”等)

# 语义词典方法的缺点

- 对于很多语言并没有好用的语义词典
- 有些词不被语义词典所包含
  - 实体、新词
- 大部分方法依赖于上下位层次关系
  - 限于名词，对于形容词和动词并不完善

# 大词林

- <http://www.bigcilin.com/>
- 一个开放域命名实体知识库自动构建系统，系统从Web搜索结果、在线百科和命名实体字面等多个信息源挖掘命名实体的类别，并从Apriori关联项、后缀上位词、分类层次化和词汇分布表示等多个角度学习获取类别之间的层次化关系

