《语义计算与知识检索》研究生课程

词汇语义计算（一）

万小军

北京大学语言计算与互联网挖掘研究室

2018年3月7日

http://www.icst.pku.edu.cn/lcwm/course/sckr2018/

内容

- 词汇语义计算概述
- 基于语义词典的词汇语义计算
- 基于语料统计的词汇语义计算

词汇语义计算概述

词汇语义

- 研究词语的意义
  - 怎样表示词语的意义?
  - 词语之间是怎么关联的?
    - 同义词、反义词、上位词、下位词、等等

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
  - 形态变化丰富
  - 由词的标准/基本形态(lemma)变化为多种形态（inflected forms ）
    - get -> gets, got, getting
- 本讲义以英文词为例

(英文)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…
- 词根(Stemming)
  - 删除后缀： ed, ing, ational, ation, able, ism, etc.
    - E.g. Agreements =>agree
  - 基于规则进行 (例如 Porter’s stemmer)
  - Stemming的结果可能不是词语
    - E.g. query, queries, querying => queri
  - 不相关的词可能具有相同的stem
    - E.g. police, policy => polic

(英文)词语形态规范化

- 词形还原(Lemmatization)
  - 将词语变为其语法原型(syntactic stem)
    - E.g. Agreements => agreement
  - 使用一般规则与例外处理
    - E.g. ies-> y, ed -> Ø, s-> Ø
    - sought-> seek, sheep -> sheep, feet -> foot
  - 处理结果仍然为词
  - 处理过程要考虑词性的不同
    - thought -> think if thought is a verb
    - thought 不变 if it is a noun

词义(Word Senses)

- 词义：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

一个词语有多少意义?

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

一个词语有多少意义?

- 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;
- 例如：drive
  - 作为动词：WordNet 3.1中有22种意义
  - 作为名词：WordNet3.1中有12种意义

Verb

- S: (v) **drive** (operate or control a vehicle) *"drive a car or bus"; "Can you drive this four-wheel truck?"*
- S: (v) **drive**, motor (travel or be transported in a vehicle) *"We drove to the university every morning"; "They motored to London for the theater"*
- S: (v) **drive** (cause someone or something to move by driving) *"She drove me to school every day"; "We drove the car to the garage"*
- S: (v) force, **drive**, ram (force into or from an action or state, either physically or metaphorically) *"She rammed her mind into focus"; "He drives me mad"*
- S: (v) **drive** (to compel or force or urge relentlessly or exert coercive pressure on, or motivate strongly) *"She is driven by her passion"*
- S: (v) repel, **drive**, repulse, force back, push back, beat back (cause to move back by force or influence) *"repel the enemy"; "push back the urge to smoke"; "beat back the invaders"*
- S: (v) **drive** (compel somebody to do something, often against his own will or judgment) *"She finally drove him to change jobs"*
- S: (v) **drive** (push, propel, or press with force) *"Drive a nail into the wall"*
- S: (v) **drive** (cause to move rapidly by striking or throwing with force) *"drive the ball far out into the field"*
- S: (v) tug, labor, labour, push, **drive** (strive and make an effort to reach a goal) *"She tugged for years to make a decent living"; "We have to push a little to make the deadline!"; "She is driving away at her doctoral thesis"*
- S: (v) **drive**, get, aim (move into a desired direction of discourse) *"What are you driving at?"*
- S: (v) **drive**, ride (have certain properties when driven) *"This car rides smoothly"; "My new truck drives well"*
- S: (v) **drive** (work as a driver) *"He drives a bread truck"; "She drives for the taxi company in Newark"*
- S: (v) **drive** (move by being propelled by a force) *"The car drove around the corner"*
- S: (v) **drive** (urge forward) *"drive the cows into the barn"*
- S: (v) **drive**, take (proceed along in a vehicle) *"We drive the turnpike to work"*
- S: (v) **drive** (strike with a driver, as in teeing off) *"drive a golf ball"*
- S: (v) **drive** (hit very hard, as by swinging a bat horizontally) *"drive a ball"*
- S: (v) **drive** (excavate horizontally) *"drive a tunnel"*
- S: (v) **drive** (cause to function by supplying the force or power for or by controlling) *"The amplifier drives the tube"; "steam drives the engines"; "this device drives the disks for the computer"*
- S: (v) **drive** ((hunting) search for game) *"drive the forest"*
- S: (v) **drive** ((hunting) chase from cover into more open ground) *"drive the game"*

Noun

- S: (n) **drive**, thrust, driving force (the act of applying force to propel something) *"after reaching the desired velocity the drive is cut off"*
- S: (n) **drive** (a mechanism by which force or power is transmitted in a machine) *"a variable speed drive permitted operation through a range of speeds"*
- S: (n) campaign, cause, crusade, **drive**, movement, effort (a series of actions advancing a principle or tending toward a particular end) *"he supported populist campaigns"; "they worked in the cause of world peace"; "the team was ready for a drive toward the pennant"; "the movement to end slavery"; "contributed to the war effort"*
- S: (n) driveway, **drive**, private road (a road leading up to a private house) *"they parked in the driveway"*
- S: (n) **drive** (the trait of being highly motivated) *"his drive and energy exhausted his co-workers"*
- S: (n) **drive**, driving (hitting a golf ball off of a tee with a driver) *"he sliced his drive out of bounds"*
- S: (n) **drive** (the act of driving a herd of animals overland)
- S: (n) **drive**, ride (a journey in a vehicle (usually an automobile)) *"he took the family for a drive in his new car"*
- S: (n) **drive** (a physiological state corresponding to a strong need or desire)
- S: (n) **drive** ((computer science) a device that writes data onto or reads data from a storage medium)
- S: (n) **drive**, parkway (a wide scenic road planted with trees) *"the riverside drive offers many exciting scenic views"*
- S: (n) **drive** ((sports) a hard straight return (as in tennis or squash))

词义基本关系

- 同义词(Synonymy)
- 反义词(Antonymy)
- 上位词(Hypernomy)
- 下位词(Hyponomy)
- 整体(Holonymy)
- 部分(Meronymy)

同义词(Synonym)

- Synonyms: 两个词的两个词义相同或接近相同, e.g. buy & 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, e.g. water and H_2O
- Synonymy最好基于词义而非词语进行定义

反义词(Antonym)

- Antonyms: 词义相反, e.g. 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
  - 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.

下位词(Hyponym) 与上位词(Hypernym)

- **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, e.g. *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*

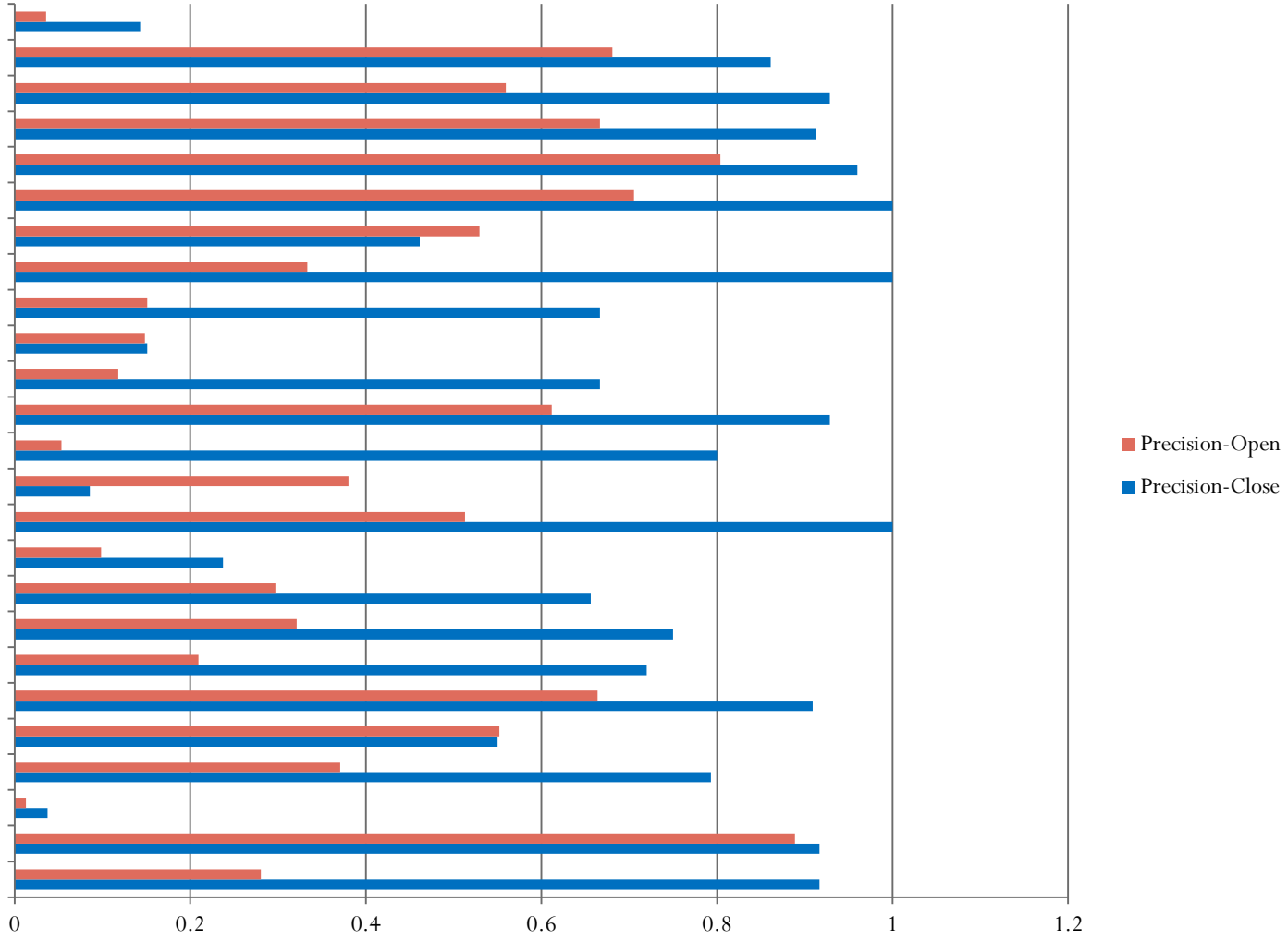
部分(Meronym)与整体(Holonym)

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

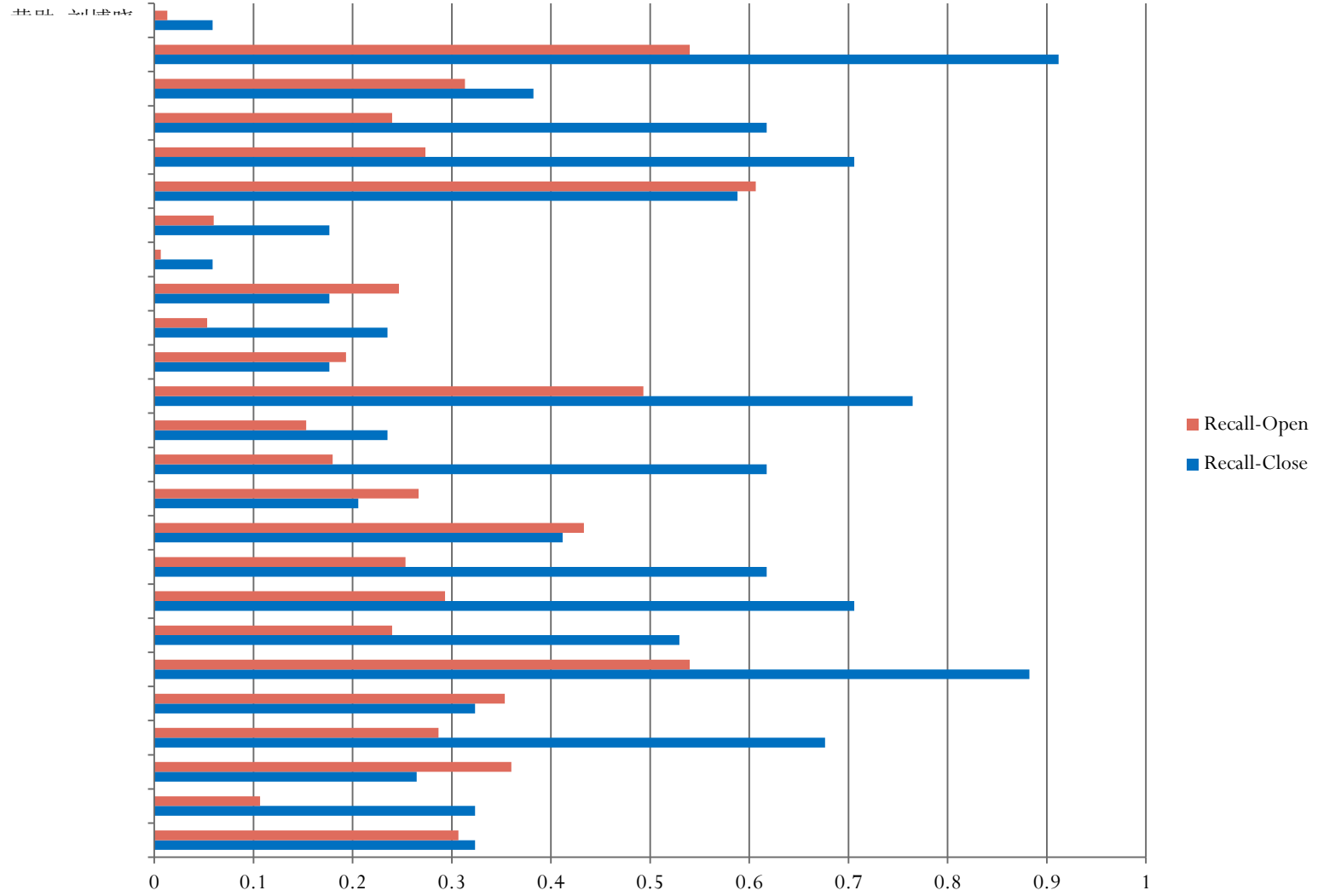
同义词抽取评测情况

- 任务
  - 词表内的同义词抽取(Close)
  - 无限制的同义词抽取(Open)
- 队伍
  - 25组
- 评测
  - 25个词条
  - 平均1.36个同义词（词表内）/ 6个同义词（无限制）

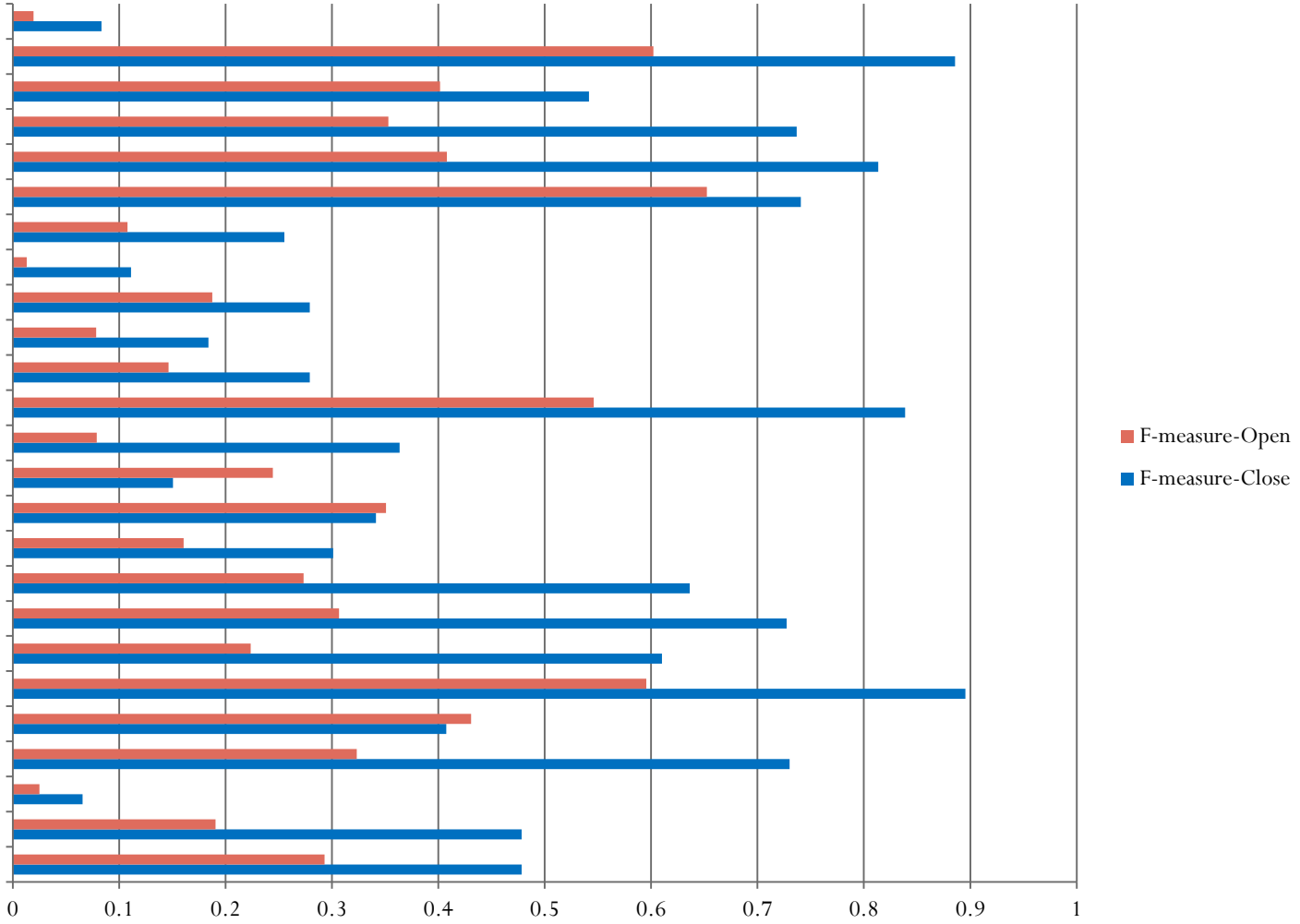
同义词抽取-Precision



同义词抽取-Recall



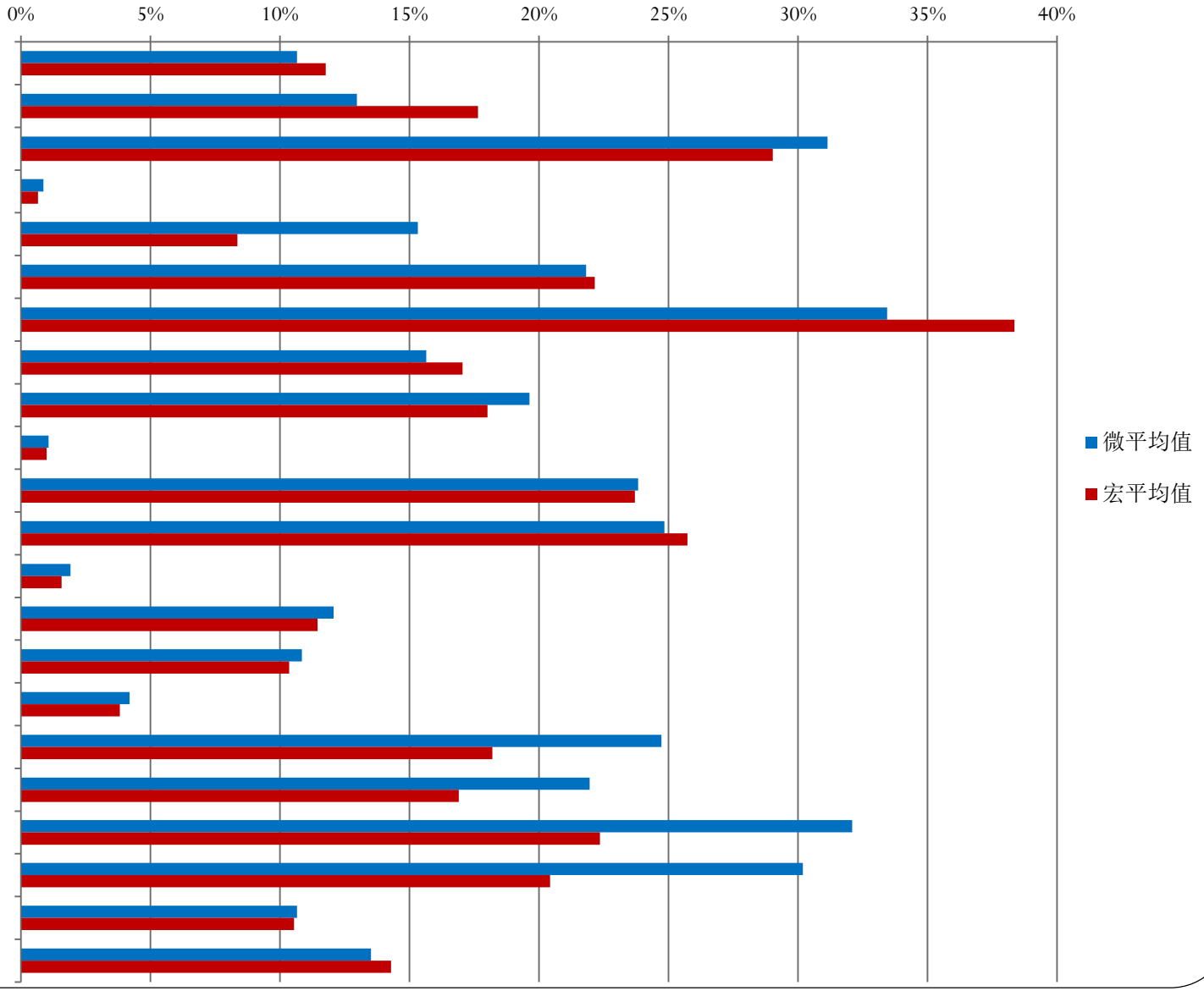
同义词抽取-F-measure



下位词发现评测情况

- 任务
  - 找出给定词的下位词
- 队伍
  - 22组
- 评测
  - 选取25个词条
  - 平均每个词对应26.8个正确下位词
  - 评测指标：F值的微平均值（对所有下位关系计算）与宏平均值（对单个词的下位关系计算后取平均）

下位词发现



思考与讨论

- 如何抽取同义词?
- 如何抽取反义词?
- 如何抽取上下位词?

思考与讨论

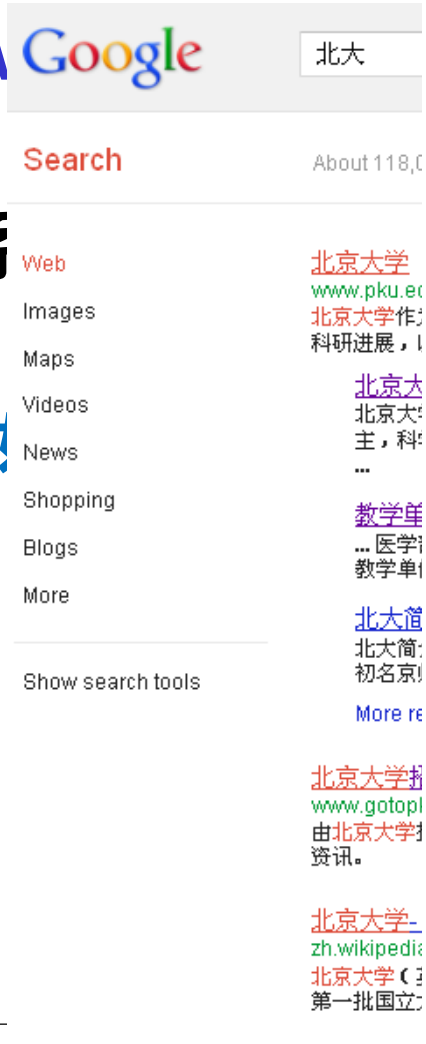
- 如何抽取同义词?
- 如何抽取反义词?
- 如何抽取上下位词?
  - 基于模板的方法
  - 基于百科的方法
  - 基于学习的方法（分类/排序)

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: 北大 vs. 北京大学
  - 上下位词（?）： 水果 vs. 苹果
- 知识/分类体系构建
- 文本推理
  - 吃水果对身体好 =》吃苹果对身体好
- 图像标注
- …

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: 北大 v
  - 上下位词（?）
- 知识/分类体系
- 文本推理
  - 吃水果对身体
- 图像标注
- …



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: a lexical database for English

GA Miller - Communications of the ACM, 1995 - dl.acm.org

Abstract Because meaningful sentences are composed of meaningful words, any system that hopes to process natural languages as people do must have information about words and their meanings. This information is traditionally provided through dictionaries, and

☆ 99 Cited by 10201 Related articles All 31 versions »

[BOOK] **WordNet**

C Fellbaum - 1998 - Wiley Online Library

Abstract **WordNet** (Miller, Beckwith, Fellbaum, Gross, & Miller 1990; Miller & Fellbaum, 1991; Miller, 1995; Fellbaum, 1998), a lexical database for English, can be thought of as a large electronic dictionary. It contains information about some 155,000 nouns, verbs, adjectives,

☆ 99 Cited by 13819 Related articles All 12 versions »

Introduction to **WordNet**: An on-line lexical database

GA Miller, R Beckwith, C Fellbaum... - International journal ..., 1990 - academic.oup.com

Abstract **WordNet** is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different

☆ 99 Cited by 5839 Related articles All 80 versions »

WordNet:: Similarity: measuring the relatedness of concepts

T Pedersen, S Patwardhan, J Michelizzi - Demonstration papers at HLT ..., 2004 - dl.acm.org

Abstract **WordNet**:: Similarity is a freely available software package that makes it possible to measure the semantic similarity and relatedness between a pair of concepts (or synsets). It provides six measures of similarity, and three measures of relatedness, all of which are

☆ 99 Cited by 1561 Related articles All 33 versions

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
  - Example synset: {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

Format of WordNet Entries

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

- Groups the meanings of English words into four categories
  - Nouns
  - Verbs
  - Adjectives
  - Adverbs

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
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

Note: *N = Nouns* *Aj = Adjectives* *V = Verbs* *Av = Adverbs*

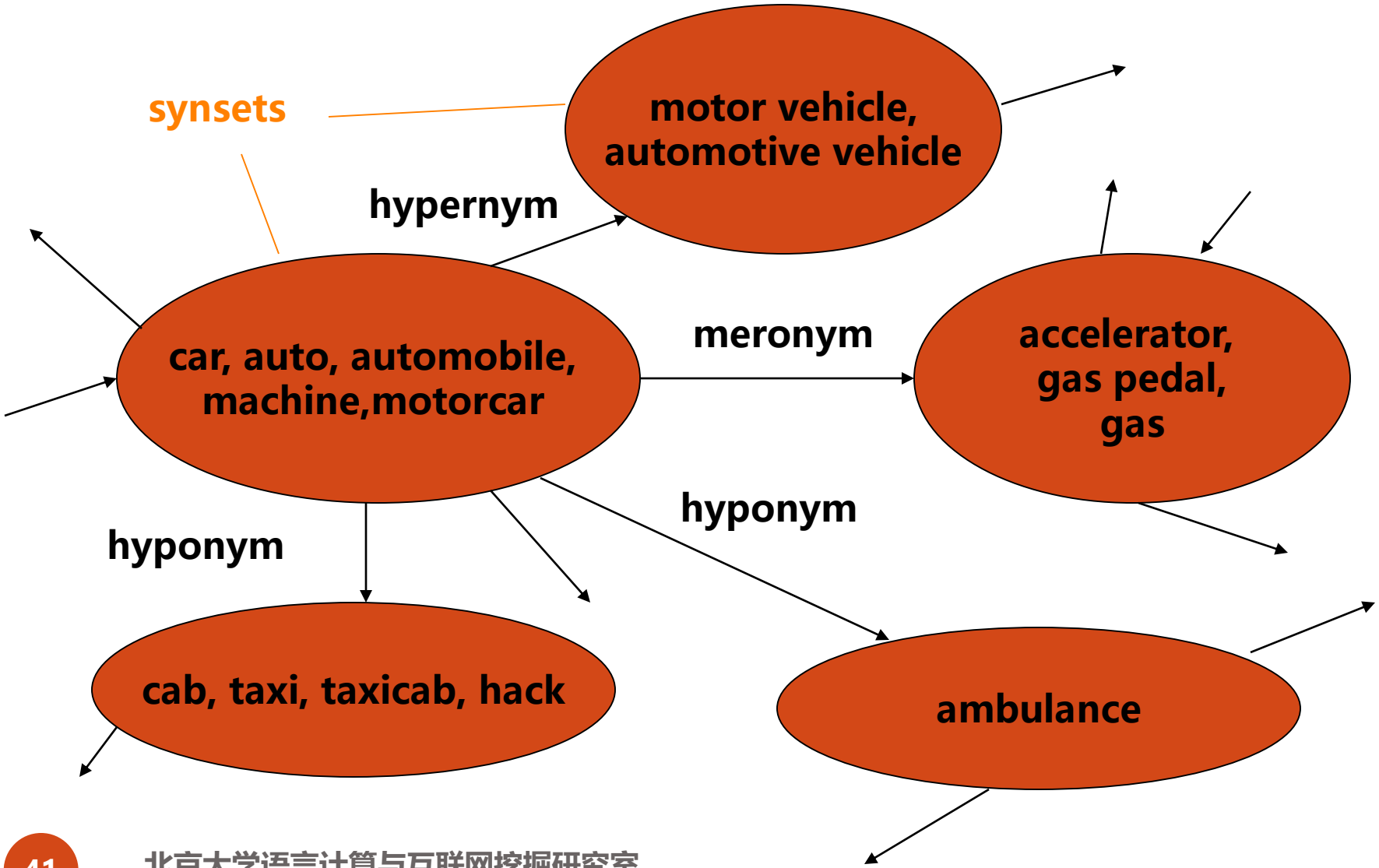
WordNet Noun Relations

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}

WordNet Verb Relations

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

A WordNet Snapshot



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 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

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

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.illc.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

Open Multilingual Wordnet

34 Open Wordnets Merged

Wordnet	Lang	Synsets	Words	Senses	Core	Licence	Data	Citation
Albanet	als	4,675	5,988	9,599	31%	CC BY 3.0	als.zip (+xml)	cite:als; (.bib)
Arabic WordNet (AWN v2)	arb	9,916	17,785	37,335	47%	CC BY SA 3.0	arb.zip (+xml)	cite:arb; (.bib)
BulTreeBank Wordnet (BTB-WN)	bul	4,959	6,720	8,936	99%	CC BY 3.0	bul.zip (+xml)	cite:bul; (.bib)
Chinese Open Wordnet	cmn	42,312	61,533	79,809	100%	wordnet	cmn.zip (+xml)	cite:cmn; (.bib)
Chinese Wordnet (Taiwan)	qcn	4,913	3,206	8,069	28%	wordnet	qcn.zip (+xml)	cite:qcn; (.bib)
DanNet	dan	4,476	4,468	5,859	81%	wordnet	dan.zip (+xml)	cite:dan; (.bib)
Greek Wordnet	ell	18,049	18,227	24,106	57%	Apache 2.0	ell.zip (+xml)	cite:ell; (.bib)
Princeton WordNet	eng	117,659	148,730	206,978	100%	wordnet	eng.zip (+xml)	cite:eng; (.bib)
Persian Wordnet	fas	17,759	17,560	30,461	41%	Free to use	fas.zip (+xml)	cite:fas; (.bib)
FinnWordNet	fin	116,763	129,839	189,227	100%	CC BY 3.0	fin.zip (+xml)	cite:fin; (.bib)
WOLF (Wordnet Libre du Français)	fra	59,091	55,373	102,671	92%	CeCILL-C	fra.zip (+xml)	cite:fra; (.bib)
Hebrew Wordnet	heb	5,448	5,325	6,872	27%	wordnet	heb.zip (+xml)	cite:heb; (.bib)
Croatian Wordnet	hrv	23,120	29,008	47,900	100%	CC BY 3.0	hrv.zip (+xml)	cite:hrv; (.bib)
IceWordNet	isl	4,951	11,504	16,004	99%	CC BY 3.0	isl.zip (+xml)	
MultiWordNet	ita	35,001	41,855	63,133	83%	CC BY 3.0	ita.zip (+xml)	cite:ita; (.bib)
ItalWordnet	ita	15,563	19,221	24,135	48%	ODC-BY 1.0	ita.zip (+xml)	cite:iwn (.bib)
Japanese Wordnet	jpn	57,184	91,964	158,069	95%	wordnet	jpn.zip (+xml)	cite:jpn; (.bib)

http://compling.hss.ntu.edu.sg/omw/

WordNet Senses

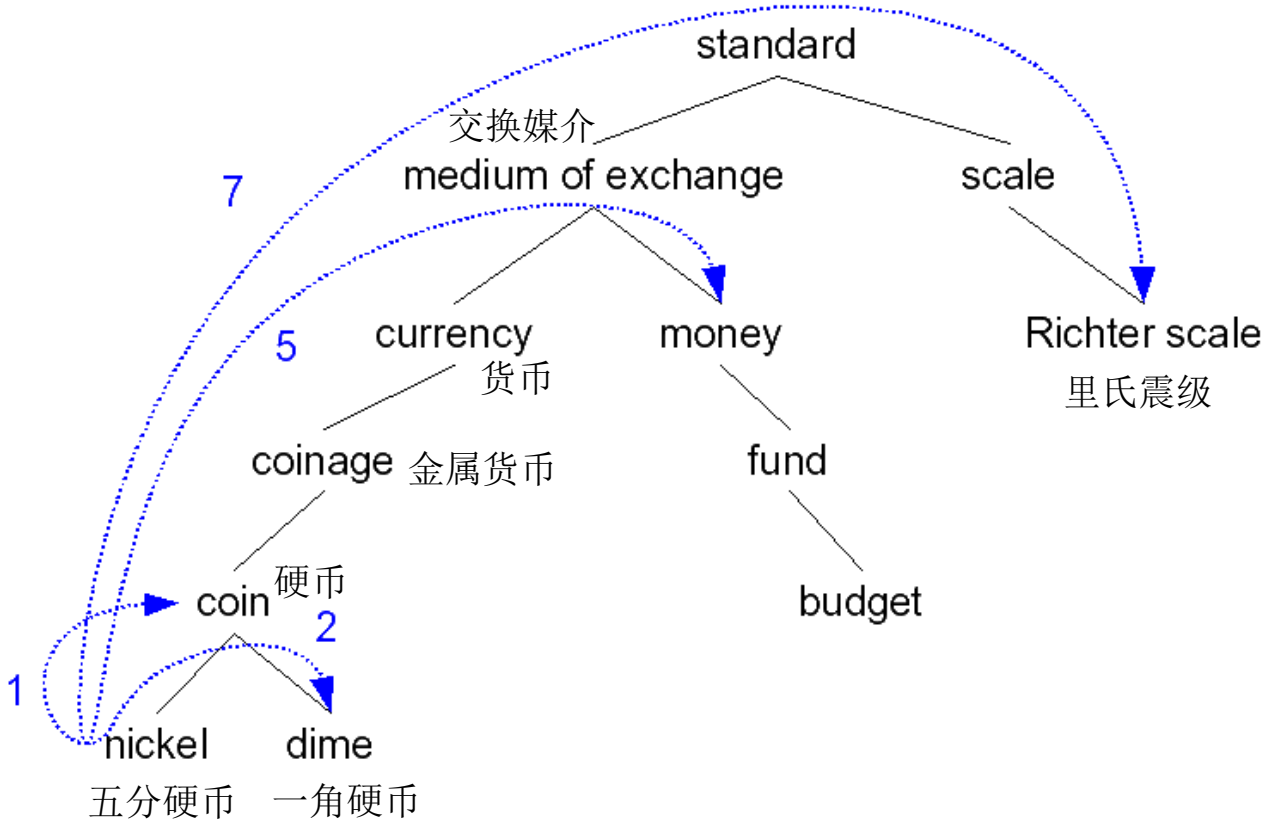
- WordNets senses倾向于细粒度
  - “play” as a verb has 35 senses, including
    - play a role or part/扮演: “Gielgud played Hamlet”
    - pretend to have certain qualities or state of mind/假装: “John played dead.”
- 人机都难以进行细粒度区分，只有词汇学专家才能有效区分
- 细粒度词义是否对NLP任务有用?
  - 不一定
- 可以考虑对细粒度词义进行归并，得到粗粒度、易于区分的词义

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)

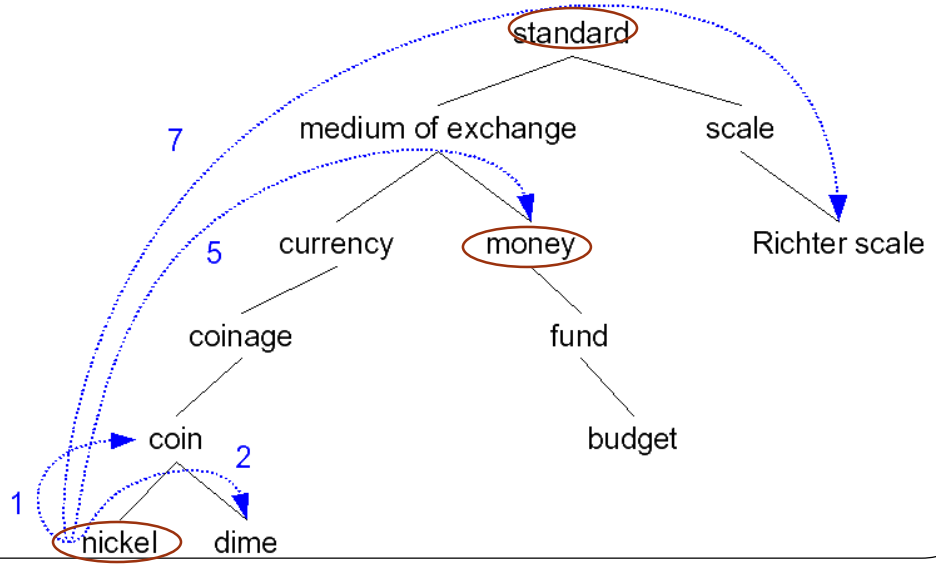


Path-based similarity的改进

- 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, *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



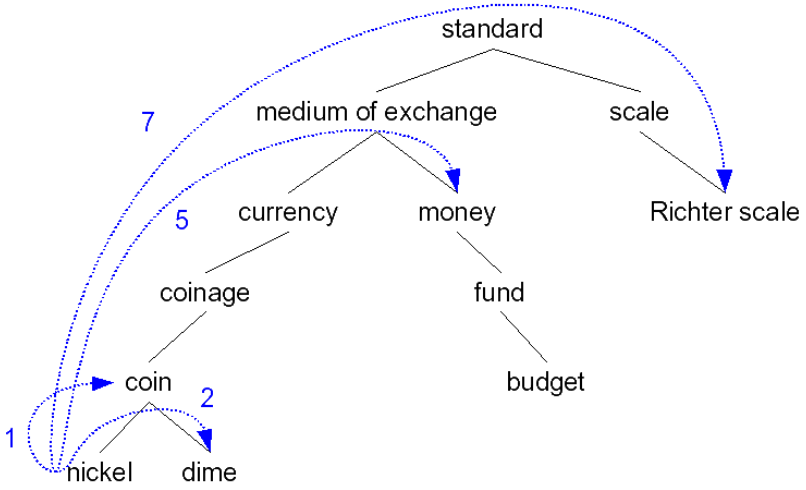
Information content similarity metrics

- 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(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

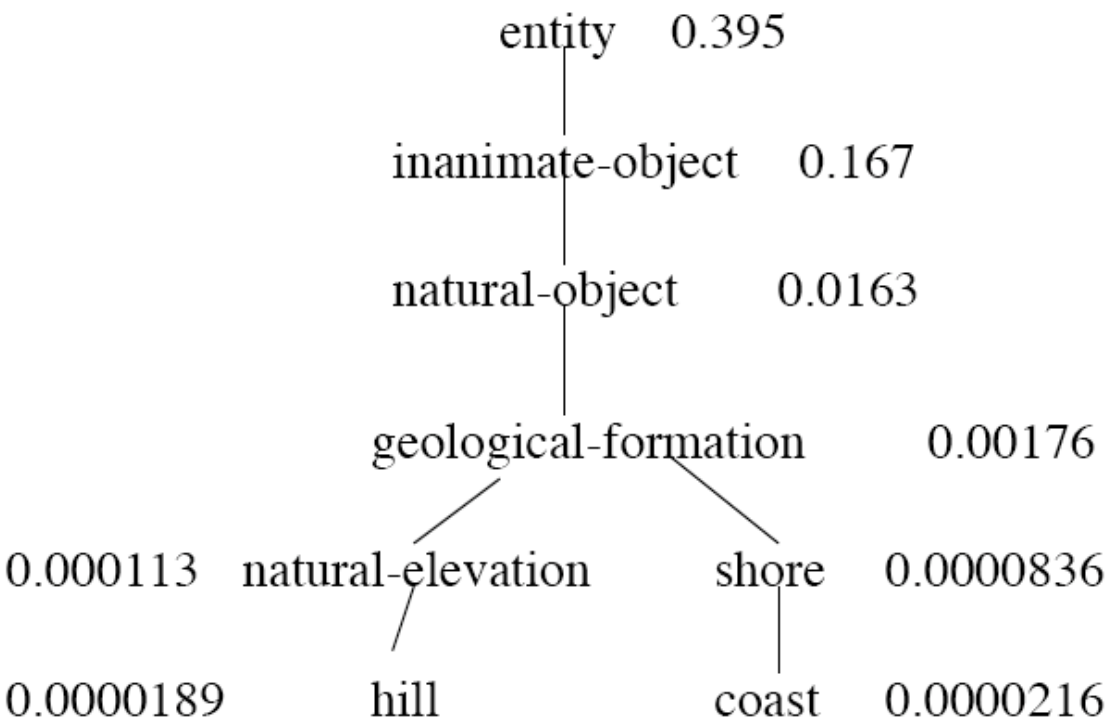
- 基于语料库进行统计
  - “dime”的出现应该被*coin, currency, standard*等词的频率所统计
  - words(c): 概念c所包容的词集（包含子孙后代节点）
  - N: 词语总数

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



Information content similarity

- WordNet结构被赋予概率P(C)



Information content: definitions

- Information content:
  - IC(c)=-logP(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

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

Lin’s Method

- SimLin(A,B)=common(A,B)/description(A,B)

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

Jiang-Conrath Method

- $Dis_{JC}(c1,c2) = 2 \log P(LCS(c1,c2)) - (\log P(c1) + \log P(c2))$
- $Sim_{JC}(c1,c2) = 1 / Dis_{JC}(c1,c2)$

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 specially prepared for use in drafting
  - *Decal*: the art of transferring designs from specially 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 *specially prepared* for 1 + 4 = 5
- RELS: 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: thesaurus-based similarity

$$\begin{aligned}
\text{sim}_{\text{path}}(c_1, c_2) &= -\log \text{pathlen}(c_1, c_2) \\
\text{sim}_{\text{Resnik}}(c_1, c_2) &= -\log P(\text{LCS}(c_1, c_2)) \\
\text{sim}_{\text{Lin}}(c_1, c_2) &= \frac{2 \times \log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \\
\text{sim}_{\text{jc}}(c_1, c_2) &= \frac{1}{2 \times \log P(\text{LCS}(c_1, c_2)) - (\log P(c_1) + \log P(c_2))} \\
\text{sim}_{\text{eLesk}}(c_1, c_2) &= \sum_{r,q \in \text{RELS}} \text{overlap}(\text{gloss}(r(c_1)), \text{gloss}(q(c_2)))
\end{aligned}$$

Comparison with human ratings of similarity

- **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.
- **Miller and Charles**: 从上述词对中抽取30词对 (10 from high level = 3-4, 10 from intermediate level = 1-3 and 10 from low level 0-1.

Similarity measure	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

其他语义词典-VerbOcean

- VerbOcean 是涵盖范围广泛的动词语义网络
  - 3,477 unique verbs
  - Unrefined 22,306 relations

SEMANTIC RELATION	EXAMPLE	Transitive	Symmetric	Num in VERBOCEAN
similarity	produce :: create	Y	Y	11,515
strength	wound :: kill	Y	N	4,220
antonymy	open :: close	N	Y	1,973
enablement	fight :: win	N	N	393
happens-before	buy :: own; marry :: divorce	Y	N	4,205

其他语义词典-BabelNet

- 语义网络多语词汇语义网络
  - 自动构建：主要将Wiki链接到WordNet，并借助机器翻译

BabelNet 4.0: General statistics

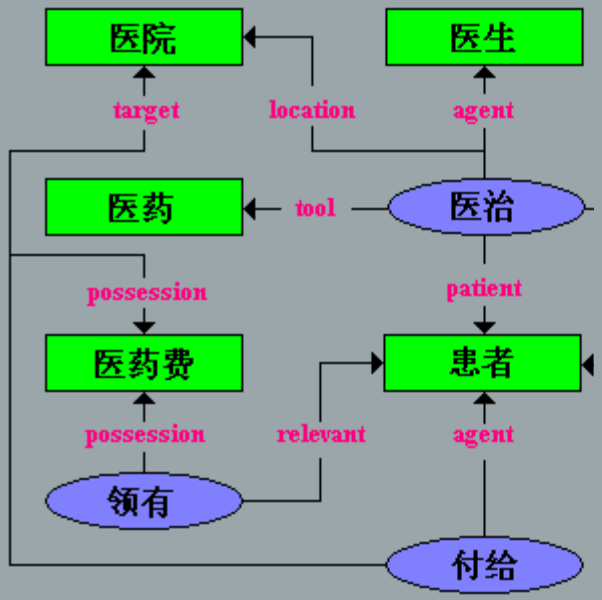
Number of languages:	284
Total number of Babel synsets:	15,788,626
Total number of Babel senses:	832,469,391
Total number of concepts:	6,117,108
Total number of Named Entities:	9,671,518
Total number of lexico-semantic relations:	1,307,706,673
Total number of glosses (textual definitions):	72,542,300
Total number of images:	53,879,884
Total number of Babel synsets with at least one domain:	2,637,414
Total number of Babel synsets with at least one picture:	10,524,280
Total number of sources:	47

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

中文语义词典

- 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）

[BOOK] **Hownet** And The Computation Of Meaning (With Cd-rom)

D Zhendong, D Qiang - 2006 - books.google.com

It is widely acknowledged that natural language processing, as an indispensable means for information technology, requires the strong support of world knowledge as well as linguistic knowledge. This book is a theoretical exploration into the extra-linguistic knowledge needed

☆ 99 Cited by 316 Related articles All 5 versions »

Semantic orientation computing based on **HowNet**

YL Zhu, J Min, Y Zhou, X Huang... - Journal of Chinese ..., 2006 - en.cnki.com.cn

Nowadays, with the development of Internet and information explosion, automated techniques for analyzing author's attitudes towards specific events will make great effort to business intelligence and public opinion survey. Semantic orientation inference has become

☆ 99 Cited by 207 Related articles »

Theoretical findings of **HowNet**

ZD Dong, Q Dong, CL Hao - Journal of Chinese Information ..., 2007 - en.cnki.com.cn

It was over 8years since the release of the first version of **HowNet**. Lots of people both at homw and abroad have already been familiar with it. Thus it is thought to be high time for us to discuss its theoretical issues. The paper elaborates the following theoretical findings:(1)

☆ 99 Cited by 68 Related articles »

HowNet-a hybrid language and knowledge resource

Z Dong, Q Dong - Natural Language Processing and ..., 2003 - ieeexplore.ieee.org

HowNet is an online common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in Chinese and English bilingual lexicons. Since it was released in 1999, **HowNet** has become more and more popular and

☆ 99 Cited by 62 Related articles

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
- 很多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

Acknowledgements

- Some slides were taken or adapted from related slides written by George A. Miller, Cosmin Adrian Bejan, Marian Olteanu, Giuseppe Carenini, Pu Wang, Keith Trnka, Danushka Bollegala, etc. Thank them for sharing their slides.