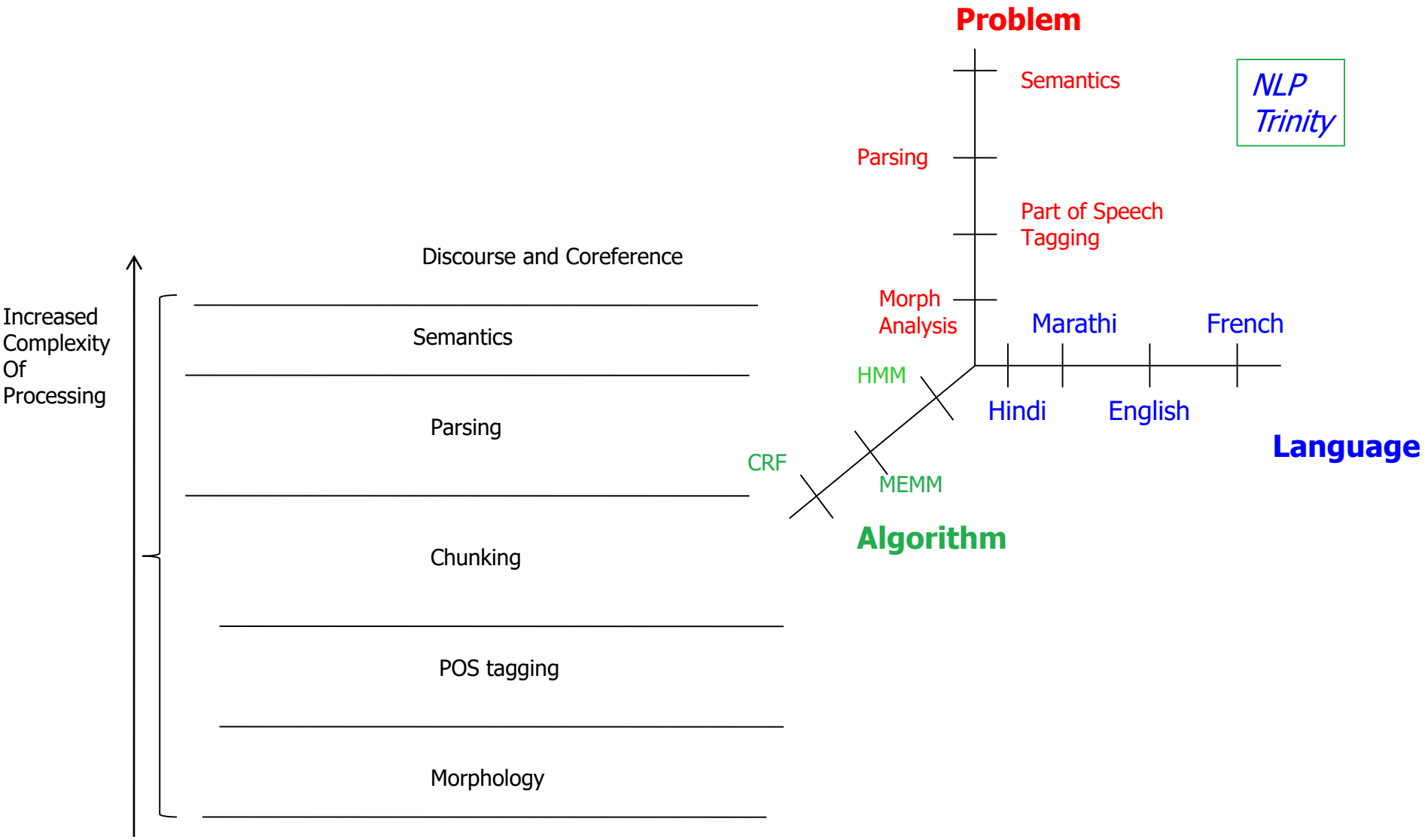
CS 563: Lexical Knowledge Network


Problem
Semantics
NLP
Trinity
Parsing
Part of Speech
Tagging
Morph
Analysis
Marathi
French
HMM
Hindi
English
Language
CRF
MEMM
Algorithm
Discourse and Coreference
Increased
Complexity
Of
Processing
Semantics
Parsing
Chunking
POS tagging
Morphology

Use of Lexical Knowledge Networks

- Proliferation of statistical ML based methods
- Still the need for deep knowledge based methods is acutely felt
- ML methods capture surface phenomena and can do limited inference

- Deeper knowledge needed for hard tasks like Textual Entailment, Question Answering and other high end NLP tasks

Consider the following problems

- How do you disambiguate 'web' in '*the spider spun a web*' from '*go surf the web*'?
- How do you summarise a long paragraph?
- How do you automatically construct language phrasebooks for tourists?
- Can a search query such as *"a game played with bat and ball"* be answered as *"cricket"*?
- Can the emotional state of a person who blogs "*I didn't expect to win the prize!*" be determined?

Some foundational points

Three Perspectives on Meaning

1. **Lexical Semantics**
   - The meanings of individual words
2. **Formal Semantics** (or Compositional Semantics or Sentential Semantics)
   - How those meanings combine to make meanings for individual sentences or utterances ?
3. **Discourse or Pragmatics**
   - How those meanings combine with each other and with other facts about various kinds of context to make meanings for a text or discourse
   - **Dialog or Conversation** is often lumped together with Discourse

The unit of meaning is a *sense*

- One word can have multiple meanings:
  - *Instead, a **bank** can hold the investments in a custodial account in the client's name*
  - *But as agriculture burgeons on the east **bank**, the river will shrink even more*

- A **word sense** is a representation of one aspect of the meaning of a word

- **bank** here has two senses

Terminology

- **Lexeme:** a pairing of meaning and form
- **Lemma:** the word form that represents a **lexeme**
  - ***Carpet*** is the lemma for ***carpets***
  - ***Dormir*** is the lemma for ***duermes***
- The lemma *bank* has two **senses:**
  - *Financial insitution*
  - *Soil wall next to water*
- A **sense** is a discrete representation of one aspect of the meaning of a word

Relations between words/senses

- Homonymy
- Polysemy
- Synonymy
- Antonymy
- Hypernymy
- Hyponymy
- Meronymy

Homonymy

- Homonymy:
  - Lexemes that share a form
    - Phonological, orthographic or both
  - But have unrelated, distinct meanings
  - Clear example:
    - Bat (wooden stick-like thing) vs Bat (flying scary mammal thing)
    - Or bank (financial institution) versus bank (riverside)
  - Can be homophones, homographs, or both:
    - Homophones:
      - Write and right
      - Piece and peace
    - Homograph: bat vs bat

Homonymy causes problems for NLP applications

- Text-to-Speech
  - Same orthographic form but different phonological form
    - bass vs bass (*a deep voice or tone/a kind of fish*)

- Information retrieval
  - Different meanings, same orthographic form
    - QUERY: bat
- Machine Translation
- Speech recognition
  - Why? (*to, two* and *too*; *bank*-gets higher probabilities even appear in different contexts)

Polysemy

Polysemy: when a single word has multiple related meanings

bank: the building
bank: the financial institution
bank: the biological repository

Most non-rare words have multiple meanings

Polysemy

1. *The **bank** was constructed in 1875 out of local red brick.*
2. *I withdrew the money from the **bank**.*

Are those the same meaning?

- We might define meaning 1 as: "The building belonging to a financial institution"
- And meaning 2: "A financial institution"

Synonyms

- Word that have the same meaning in some or all contexts
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - water / H_20

Synonyms

- But there are few (or no) examples of perfect synonymy
  - Why should that be?
  - Even if many aspects of meaning are identical
  - Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.
- Example:
  - Big/large
  - Water and H_20

Tropes, or Figures of Speech

- Metaphor: one entity is given the attributes of another (tenor/vehicle/ground)
  - Life is a bowl of cherries. Don't take it serious….
  - We are the eyelids of defeated caves. ??
- Metonymy: one entity used to stand for another (replacive)
  - GM killed the Fiero. (kill- *put an end of an ongoing activity*)
  - The ham sandwich wants his check.
- Both extend existing sense to new meaning
  - Metaphor: completely different concept
  - Metonymy: related concepts

Antonyms

- Senses that are opposites **with respect to one feature of their meaning**

- Otherwise, they are very similar!
  - dark / light
  - short / long
  - hot / cold
  - up / down
  - in / out

Hyponyms and Hypernyms

- **Hyponym:** the sense is a subclass of another sense
  - *car* is a hyponym of *vehicle*
  - *dog* is a hyponym of *animal*
  - *mango* is a hyponym of *fruit*
- **Hypernym**: the sense is a superclass
  - *vehicle* is a hypernym of *car*
  - *animal* is a hypernym of *dog*
  - *fruit* is a hypernym of *mango*

hypernym	vehicle	fruit	furniture	mammal
hyponym	car	mango	chair	dog

Selectional Preferences (Indian Tradition)

- "Desire" of some words in the sentence ("aakaangksha").
  - *I **saw** the boy with long hair.*
  - *The verb **"saw"** and the noun **"boy"** desire an object here.*
- "Appropriateness" of some other words in the sentence to fulfil that desire ("yogyataa").
  - *I saw the boy with long hair.*
  - *The PP **"with long hair"** can be appropriately connected only to **"boy"** and not **"saw".***

- In case, the ambiguity is still present, "proximity" ("sannidhi") can determine the meaning.
  - *E.g. I saw the boy with a telescope.*
  - *The PP **"with a telescope"** can be attached to both **"boy"** and **"saw"**, so ambiguity still present. It is then attached to "boy" using the proximity check.*

Selectional Preference (Recent Linguistic Theory)

- There are words which demand arguments, like, verbs, prepositions, adjectives and sometimes nouns. These arguments are typically nouns.
- Arguments must have the property to fulfil the demand. They must satisfy selectional preferences
  - Example
    - Give (verb)
      - agent – animate
      - obj – direct
      - obj – indirect
    - *I **gave** him the **book***
    - *I **gave** him the **book** (yesterday in the school) -> adjunct*
- How does this help in WSD?
  - One type of contextual information is the information about the type of arguments that a word takes.

Verb Argument frame

- Structure expressing the desire of a word is called the *Argument Frame*

- Selectional Preference
  - Properties of the "Supply Words" meeting the desire of the previous set

Argument frame (example)

Sentence: *I am fond of X*

Fond

{

Arg1: *Prepositional Phrase (PP)*

{

PP: *<u>of</u> NP*

{

N: *somebody/something*

}

}

}

Verb Argument frame (example)

Verb: ***give***

Give

{

agent: *<the give>animate*

direct object: *<the thing given>*

indirect object:
<beneficiary>animate/organization

}

[I]$_{\textbf{agent}}$ *gave a [book]*$_{\textbf{dobj}}$ *to [Ram]*$_{\textbf{iobj}}$*.*

Parameters for Lexical Knowledge Networks

1. Domains addressed by the structure

2. Principles of Construction

3. Methods of Construction

4. Representation

5. Quality of database

6. Applications

7. Usability mechanisms for software applications and users: APIs, record structure, User interfaces

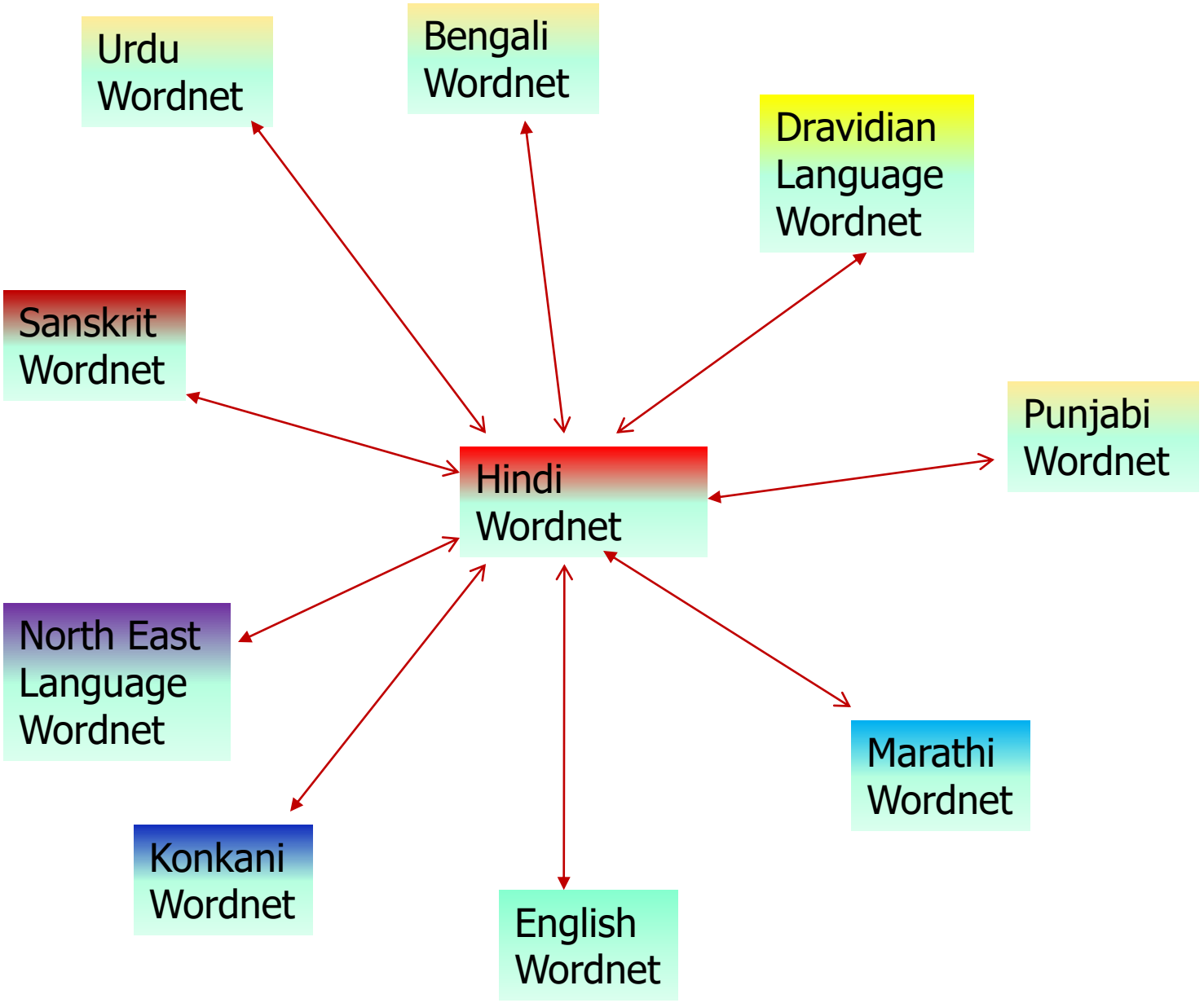
8. Size and coverage

Wordnet

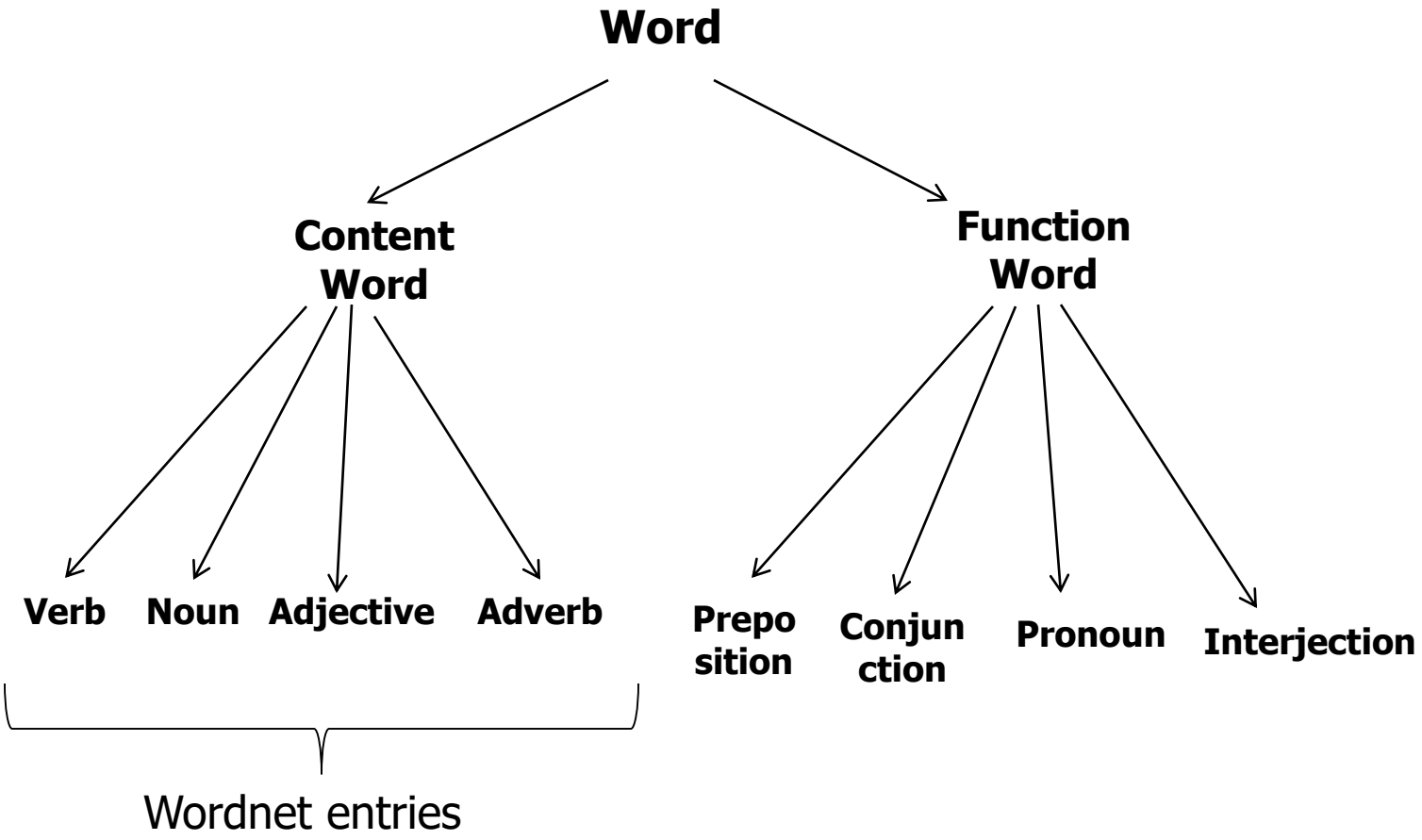
Wordnet: main purpose

- Disambiguation: **Sense Disambiguation**
- Main instrument: Relational Semantics
- Disambiguate *by other words*
  - *{house}: "house" as a kind of "physical structure"*
  - *{family, house}: "family" as an abstract concept*
  - *{house, astrological position}: "astrological place" as a concept*
    - Most horoscopic traditions of astrology systems divide the horoscope into a number (usually twelve) of **houses** whose positions depend on time and location rather than on date

INDOWORDNET



Classification of Words



Sense tagged corpora (*task: sentiment analysis*)

- I have enjoyed_21803158 #LA#_18933620 every_42346474 time_17209466 I have been_22629830 there_3110157 , regardless_3119663 if it was for work_1578942 or pleasure_11057430.
- I usually_3107782 fly_21922384 into #LA#_18933620, but this time_17209466 we decided_2689493 to drive_21912201 .
- Interesting_41394947, to say_2999158 the least_3112746 .

Senses of “pleasure”

The noun pleasure has 5 senses (first 2 from tagged texts)

1. (21) pleasure, pleasance -- (a fundamental feeling that is hard to define but that people desire to experience; "he was tingling with pleasure")
2. (4) joy, delight, pleasure -- (something or someone that provides pleasure; a source of happiness; "a joy to behold"; "the pleasure of his company"; "the new car is a delight")
3. pleasure -- (a formal expression; "he serves at the pleasure of the President")
4. pleasure -- (an activity that affords enjoyment; "he puts duty before pleasure")

etc

Basic Principle

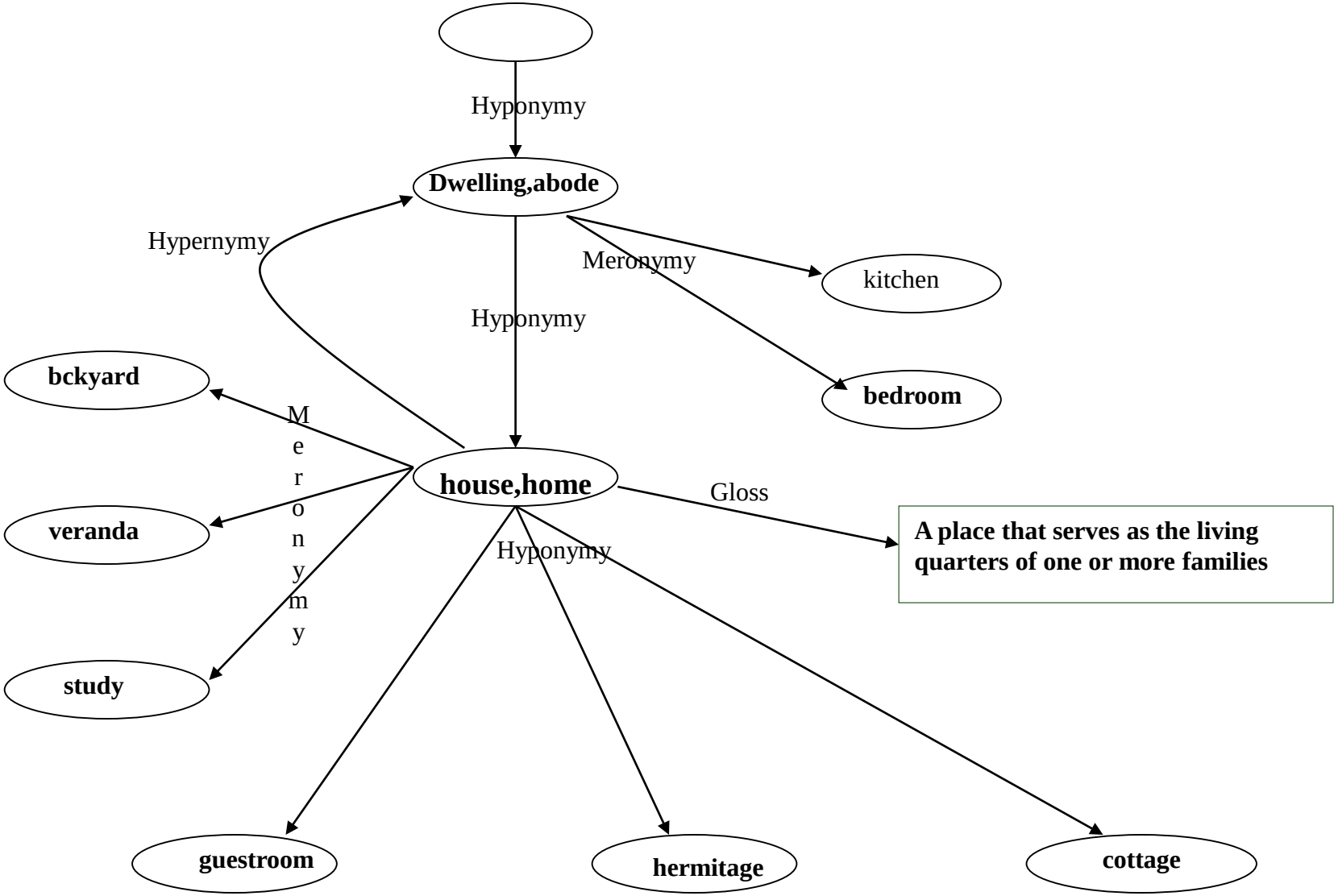
- Words in natural languages are polysemous
- However, when synonymous words are put together, a unique meaning often emerges
- Use is made of *Relational Semantics*

Lexical and Semantic relations in WordNet

1. Synonymy
2. Hypernymy / Hyponymy
3. Antonymy
4. Meronymy / Holonymy
5. Gradation
6. Entailment
7. Troponymy

1, 3 and 5 are lexical (*word to word)*, rest are semantic (*synset to synset).*

WordNet Sub-Graph



Lexical entailment

- **A verb V1 logically entails a verb V2 when the sentence** « Someone V1 » (logically) entails the sentence « Someone V2 ».
  - Ex. "**snore" lexically entails "sleep".**
  - The first sentence "presuppose" the second.
- Negation reverses the direction of entailment:
  - Ex. Not sleeping entails not snoring.
- Lexical entailment is a non-symmetric relation:
  - Only synonymous verbs can be mutually entailing
    - Ex. "A defeated B" and "A beat B".

Temporal inclusion

- A verb V1 will be said to temporally include a verb V2 if there is some stretch of time during which the activities denoted by the two verbs co-occur, but no time during which V2 occurs and V1 does not
  - Ex. “snore” entails “sleep” and is properly included by it.
- If V1 entails V2 and if a temporal inclusion relation holds between V1 and V2, then people will accept a part-whole statement relating V2 and V1

Troponymy

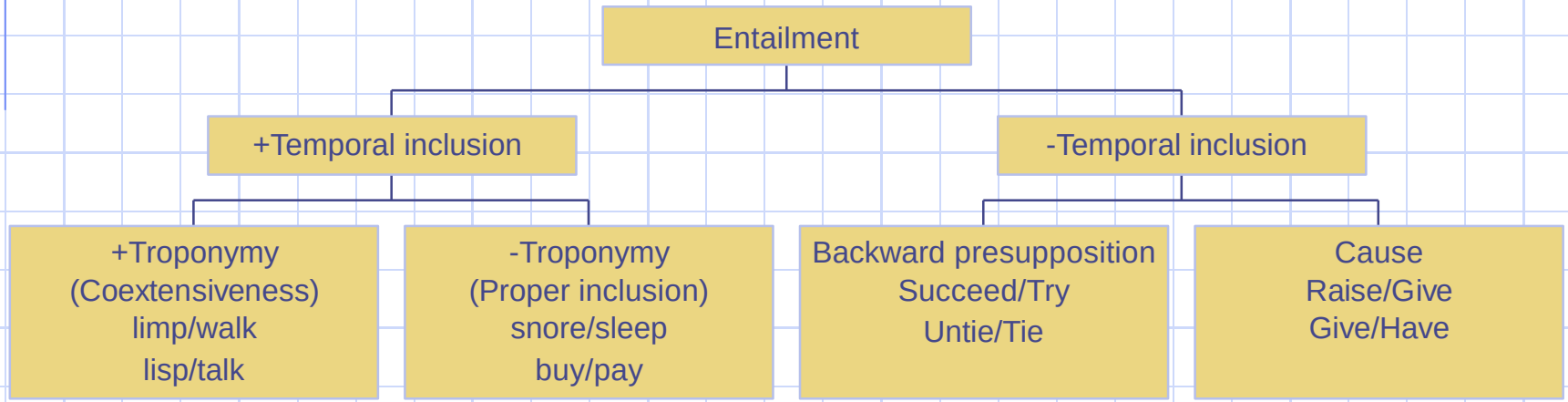
- The troponymy relation between two verbs V1 and V2 can be expressed by the formula:
  - To V1 is to V2 in some particular "manner"
  - Ex. Troponyms of communication:
    - Encode the speaker's intention like in
      - Examine, confess, preach, ...
    - Encode the medium of communication like in
      - Fax, email, phone, telex, ...
  - Troponymy is a particular kind of entailment:
    - Every troponym V1 of a (more general) verb V2 also entails V2.
    - **The activity referred by a troponym and its more general hypernym are always temporally coextensive**
    - Obs. "snore" is not a troponym of "sleep" (because of proper temporal inclusion).

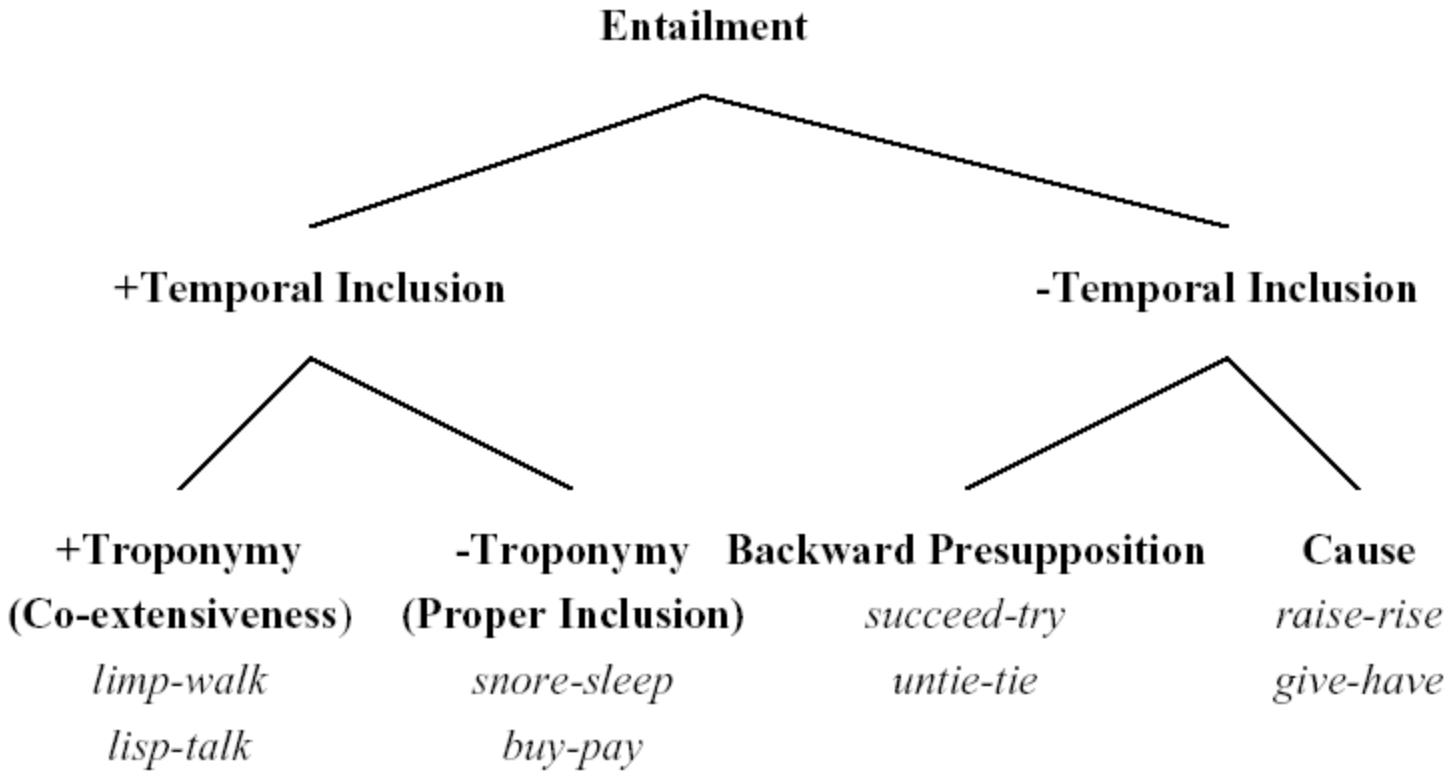
Backward presupposition

- The activity denoted by the entailed verb always preceeds in time the activity denoted by the entailing verb
  - Succeed-try (entailing-entailed)
  - Untie-tie (entailing-entailed)

Entailment Relations for Verbs

Entailment relations among verbs




Entailment
+Temporal Inclusion
-Temporal Inclusion
+Troponymy
(Co-extensiveness)
limp-walk
lisp-talk
-Troponymy
(Proper Inclusion)
snore-sleep
buy-pay
Backward Presupposition
succeed-try
untie-tie
Cause
raise-rise
give-have

Lexical Relation

- Antonymy
  - Oppositeness in meaning
  - Relation between word forms
  - Often determined by phonetics, word length etc. ({rise, ascend} *vs.* {fall, descend})

Kinds of Antonymy

Size	**Small - Big**
Quality	**Good – Bad**
State	**Warm – Cool**
Personality	**Dr. Jekyl- Mr. Hyde**
Direction	**East- West**
Action	**Buy – Sell**
Amount	**Little – A lot**
Place	**Far – Near**
Time	**Day - Night**
Gender	**Boy - Girl**

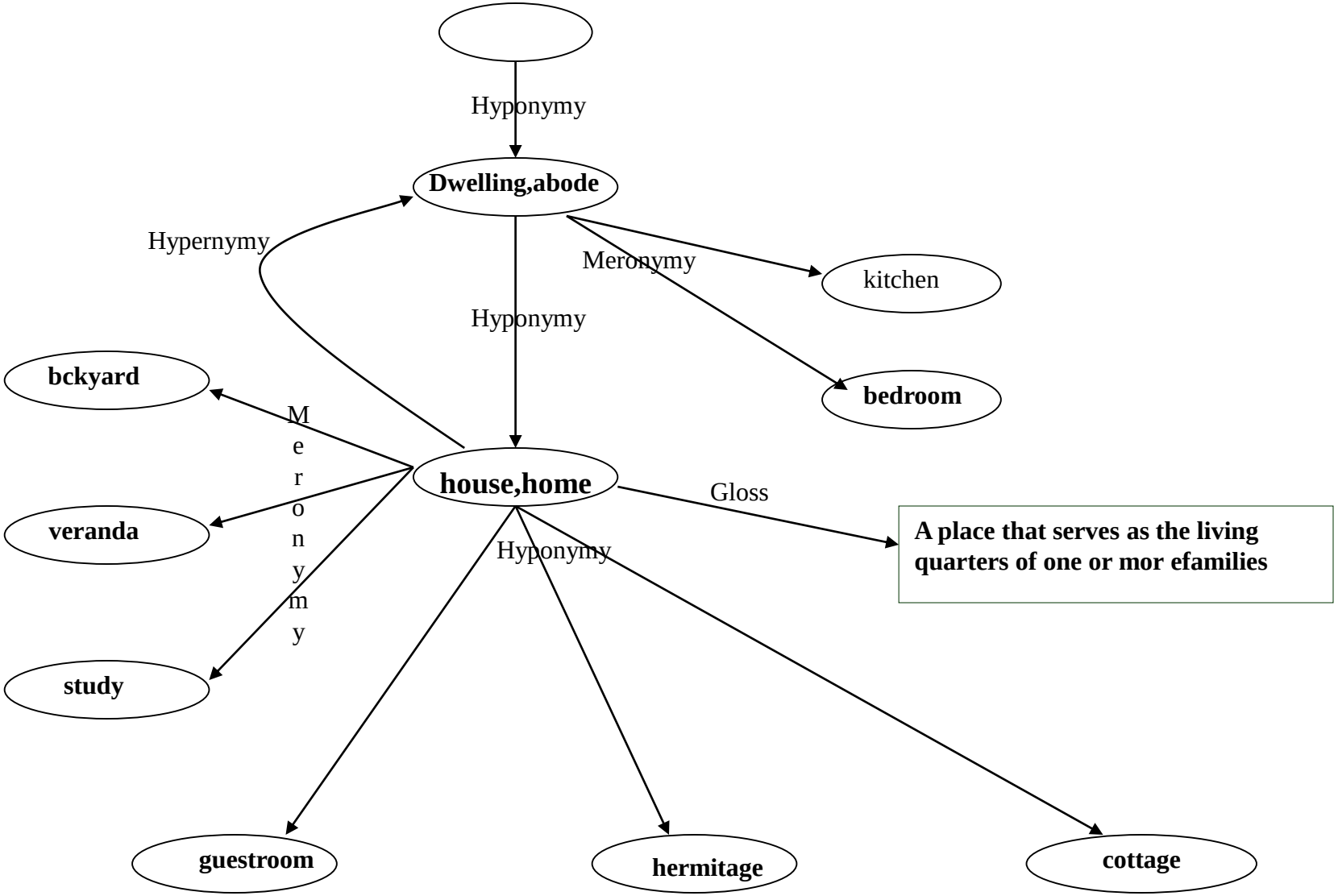
Kinds of Meronymy

Component-object	**Head - Body**
Staff-object	**Wood - Table**
Member-collection	**Tree - Forest**
Feature-Activity	**Speech - Conference**
Place-Area	**Palo Alto - California**
Phase-State	**Youth - Life**
Resource-process	**Pen - Writing**
Actor-Act	**Physician - Treatment**

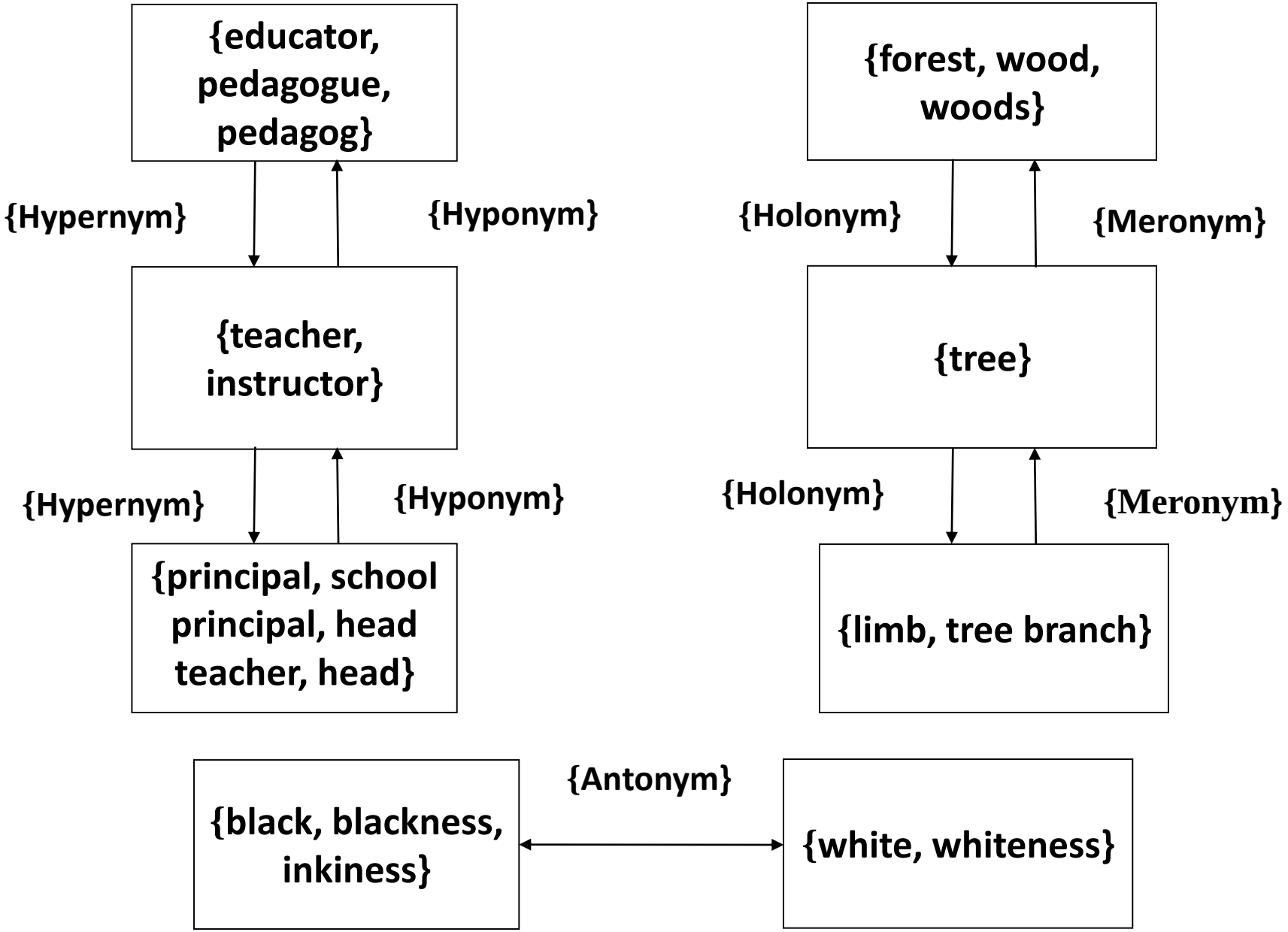
Gradation

State	Childhood, Youth, Old age
Temperature	Hot, Warm, Cold
Action	Sleep, Doze, Wake

WordNet Sub-Graph

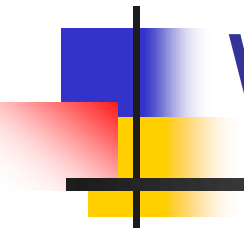


WordNet – Illustration of relations



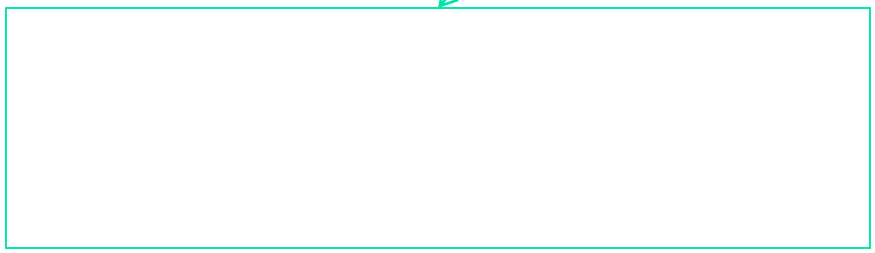
WordNet: limitations

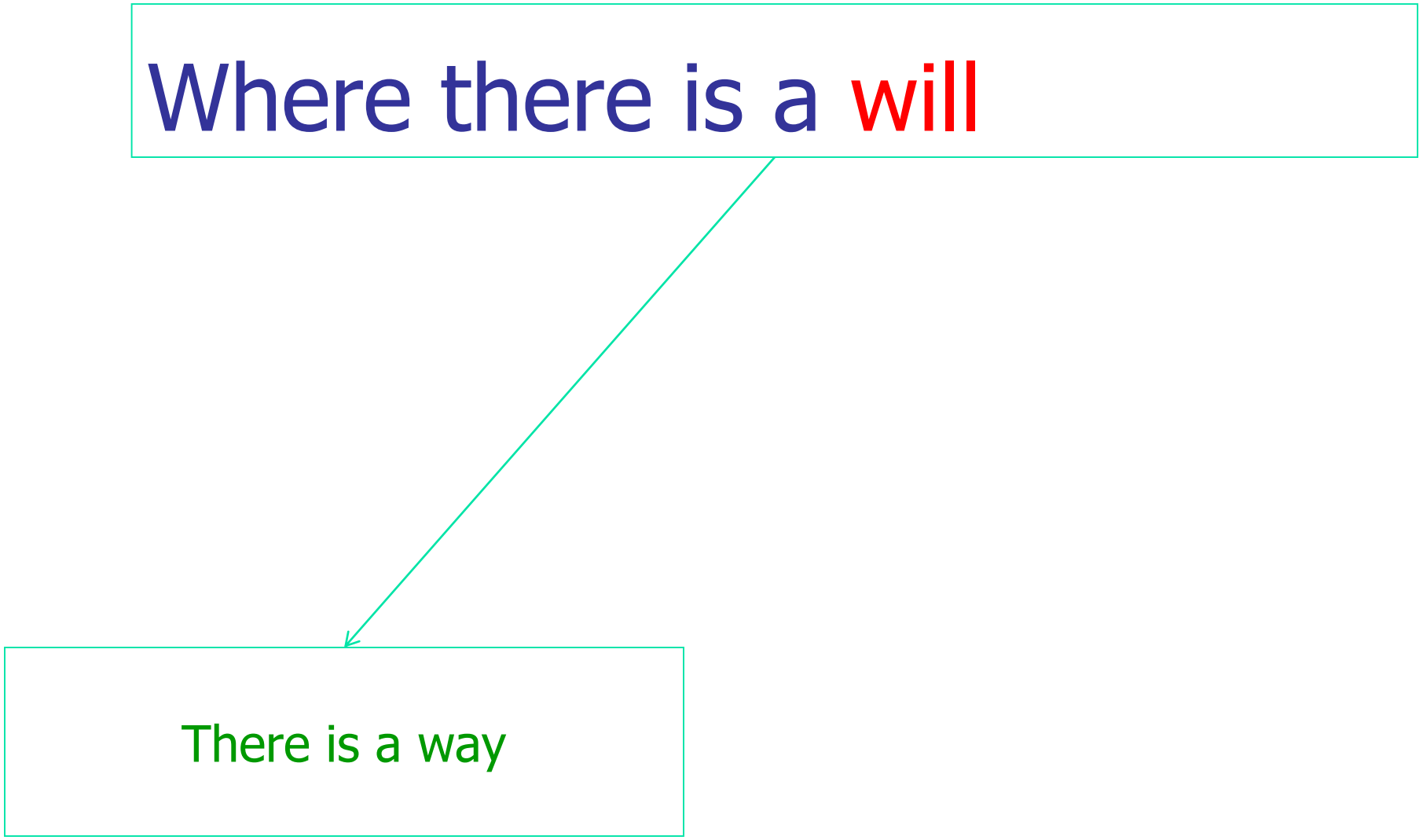
- Contains little syntactic information
- No explicit predicate argument structures
- No systematic extension of basic senses
- Sense distinctions are very fine-grained, **IAA 73%**
- No hierarchical entries

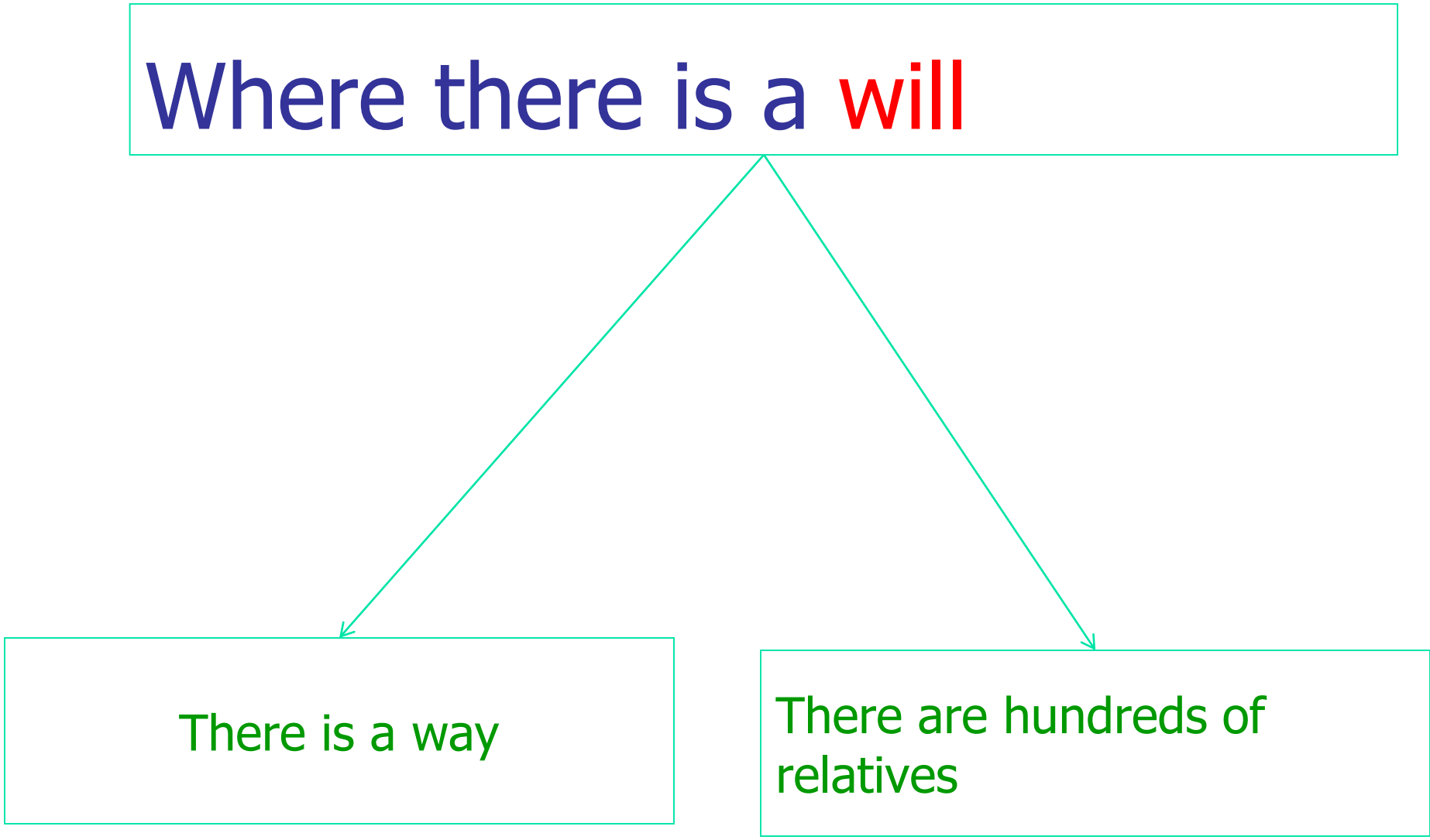


WSD algorithms

Where there is a will




Where there is a will
There is a way


Where there is a will
There is a way
There are hundreds of relatives

Example of WSD

- **Operation**, surgery, surgical operation, surgical procedure, surgical process -- (a medical procedure involving an incision with instruments; performed to repair damage or arrest disease in a living body; "they will schedule the operation as soon as an operating room is available"; "he died while undergoing surgery") TOPIC->(noun) surgery#1
- **Operation**, military operation -- (activity by a military or naval force (as a maneuver or campaign); "it was a joint operation of the navy and air force") TOPIC->(noun) military#1, armed forces#1, armed services#1, military machine#1, war machine#1
- **Operation** -- ((computer science) data processing in which the result is completely specified by a rule (especially the processing that results from a single instruction); "it can perform millions of operations per second") TOPIC->(noun) computer science#1, computing#1
- mathematical process, mathematical **operation**, **operation** -- ((mathematics) calculation by mathematical methods; "the problems at the end of the chapter demonstrated the mathematical processes involved in the derivation"; "they were learning the basic operations of arithmetic") TOPIC->(noun) mathematics#1, math#1, maths#1

IS WSD NEEDED IN LARGE APPLICATIONS?

WSD clue can be distant: Sense of word “bank” ?

I went to the bank that was closed due to a public holiday with government fearing massive influx of crowd who would want to withdraw money or mass.

Essential Resource for WSD: ***Wordnet***

Word Meanings	**Word Forms**				
	F_1	**F_2**	**F_3**	**…**	**F_n**
M_1	(*depend*) $E_{1,1}$	(*bank*) $E_{1,2}$	(rely) $E_{1,3}$		
M_2		*(bank)* $E_{2,2}$		(*embankment*) $E_{2,…}$	
M_3		*(bank)* $E_{3,2}$	$E_{3,3}$		
…				…	
M_m					$E_{m,n}$

Example of sense marking: its need

एक_4187 नए शोध_1138 के अनुसार_3123 जिन लोगों_1189 का सामाजिक_43540 जीवन_125623 व्यस्त_48029 होता है उनके दिमाग_16168 के एक_4187
हिस्से_120425 में अधिक_42403 जगह_113368 होती है।

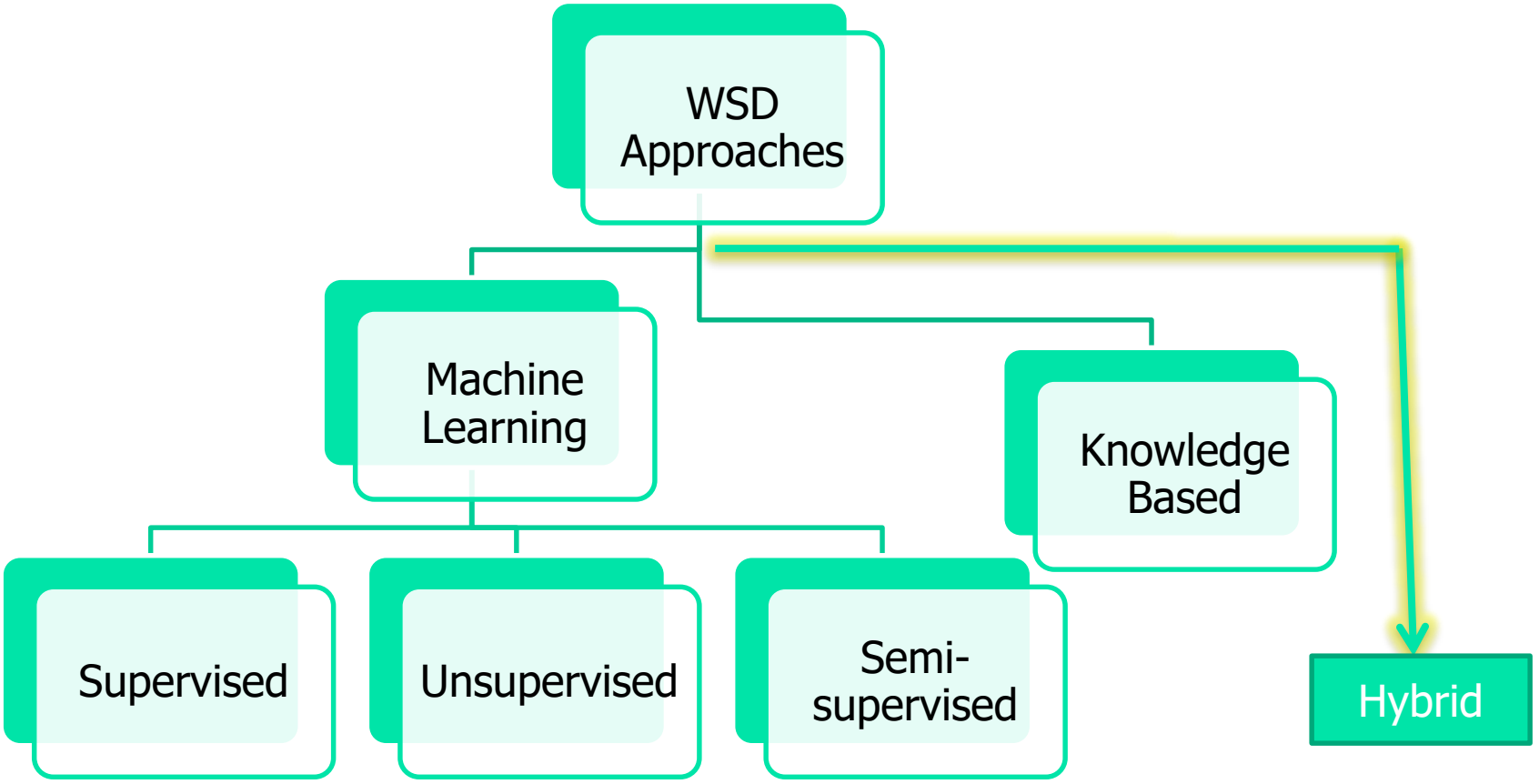
(According to a new research, those people who have a busy social life, have larger space in a part of their brain).

नेचर न्यूरोसाइंस में छपे एक_4187 शोध_1138 के अनुसार_3123 कई_4118 लोगों_1189 के दिमाग_16168
के स्कैन से पता_11431 चला कि दिमाग_16168 का एक_4187 हिस्सा_120425 एमिगडाला सामाजिक_43540 व्यस्तताओं_1438 के साथ_328602 सामंजस्य_166
के लिए थोड़ा_38861 बढ़_25368 जाता है। यह शोध_1138 58 लोगों_1189 पर किया गया जिसमें उनकी उम्र_13159 और दिमाग_16168 की साइज़ के आँकड़े_128065
लिए गए। अमरीकी_413405 टीम_14077 ने पाया_227806 कि जिन लोगों_1189 की सोशल नेटवर्किंग अधिक_42403 है उनके दिमाग_16168 का एमिगडाला
वाला हिस्सा_120425 बाकी_130137 लोगों_1189 की तुलना_में_38220 अधिक_42403 बड़ा_426602 है। दिमाग_16168 का एमिगडाला वाला हिस्सा_120425
भावनाओं_1912 और मानसिक_42151 स्थिति_1652 से जुड़ा हुआ माना_212436 जाता है।

Ambiguity of लोगों (People)

- **लोग**, **जन**, **लोक**, **जनमानस**, **पब्लिक** - एक से अधिक व्यक्ति *"लोगों के हित में काम करना चाहिए"*
  - **(English synset) multitude, masses, mass, hoi_polloi, people, the_great_unwashed - the common people generally *"separate the warriors from the mass" "power to the people"***
- **दुनिया**, **दुनियाँ**, **संसार**, **विश्व**, **जगत**, **जहाँ**, **जहान**, **ज़माना**, **जमाना**, **लोक**, **दुनियावाले**, **दुनियाँवाले**, **लोग** - संसार में रहने वाले लोग *"महात्मा गाँधी का सम्मान पूरी दुनिया करती है / मैं इस दुनिया की परवाह नहीं करता / आज की दुनिया पैसे के पीछे भाग रही है"*
  - **(English synset) populace, public, world - people in general considered as a whole *"he is a hero in the eyes of the public"***

Bird's eye view of WSD



OVERLAP BASED APPROACHES

- Require a ***Machine Readable Dictionary*** *(MRD).*
- Find the overlap between the features of different senses of an ambiguous word (**sense bag**) and the features of the words in its context (**context bag**)
- These features could be sense definitions, example sentences, hypernyms etc
- The features could also be given weights
- The sense which has the maximum overlap is selected as the contextually appropriate sense

LESK'S ALGORITHM

Sense Bag: *contains the words in the definition of a candidate sense of the ambiguous word.*

Context Bag: *contains the words in the definition of each sense of each context word.*

E.g. "On burning ***coal*** we get ***ash***."

From Wordnet

- The noun ash has 3 senses (first 2 from tagged texts)
- 1. (2) ash -- (the residue that remains when something is burned)
- 2. (1) ash, ash tree -- (any of various deciduous pinnate-leaved ornamental or timber trees of the genus Fraxinus)
- 3. ash -- (strong elastic wood of any of various ash trees; used for furniture and tool handles and sporting goods such as baseball bats)
- The verb ash has 1 sense (no senses from tagged texts)
- 1. ash -- (convert into ashes)

CRITIQUE

- Proper nouns in the context of an ambiguous word can act as strong disambiguators.

  E.g. **"Sachin Tendulkar"** will be a strong indicator of the category **"sports".**

  Sachin Tendulkar plays **cricket.**

- Proper nouns are not present in the thesaurus. Hence this approach fails to capture the strong clues provided by proper nouns.
- Accuracy
  - 50% when tested on 10 highly polysemous English words.

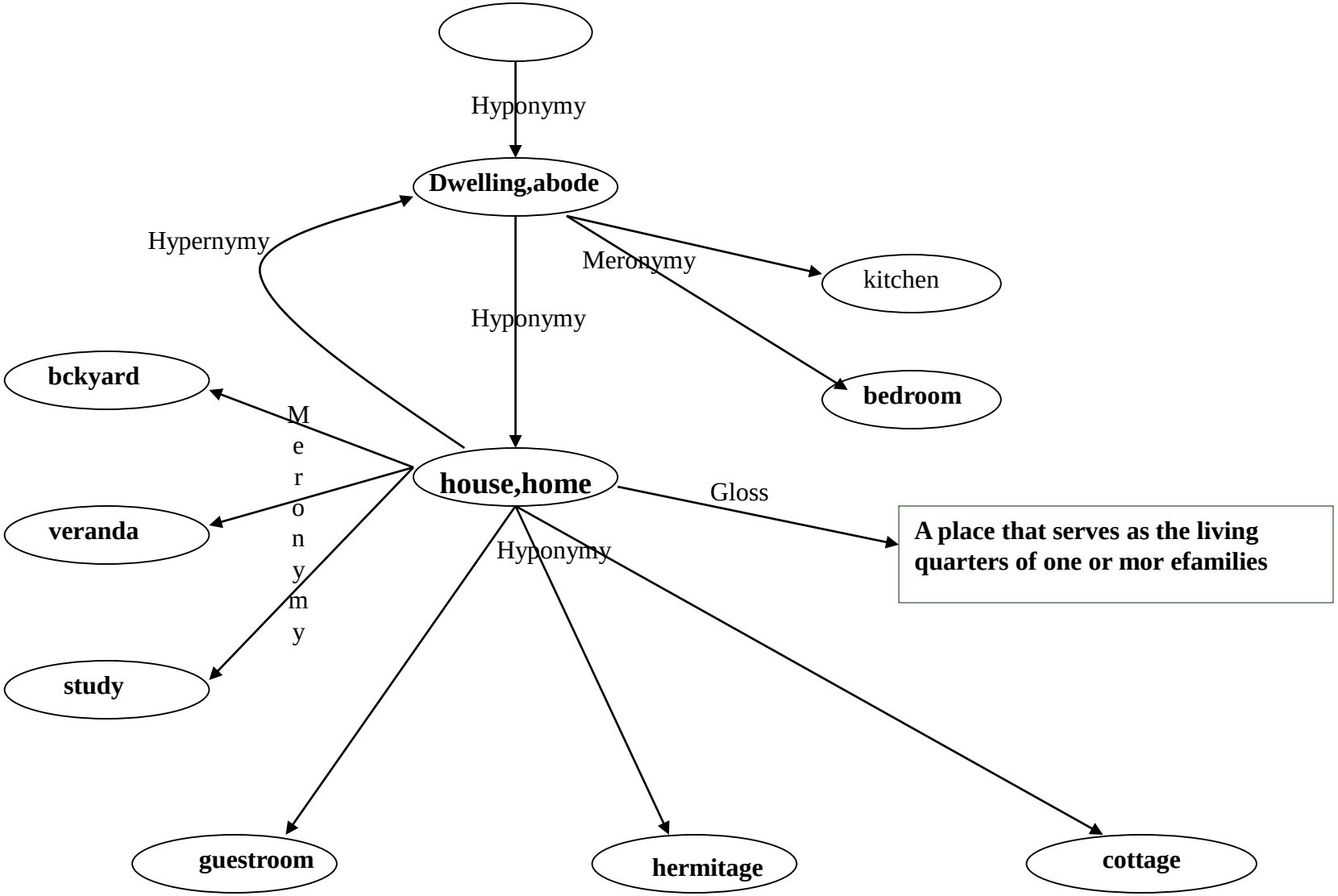
Extended Lesk's algorithm

- Original algorithm is sensitive towards exact words in the definition
- Extension includes glosses of semantically related senses from WordNet (e.g. *hypernyms*, *hyponyms*, etc.)
- The scoring function becomes:

$$score_{ext}(S) = \sum_{s' \in rel(s) \text{ or } s \equiv s'} |context(w) \cap gloss(s')|$$

  - where,
    - *gloss(S) is the gloss of sense S from the lexical resource.*
    - *Context(W) is the gloss of each sense of each context word.*
    - *rel(s)* gives the senses related to s in WordNet under some relations.

WordNet Sub-Graph



Example: Extended Lesk

- *"On combustion of coal we get ash"*

From Wordnet

- The noun ash has 3 senses (first 2 from tagged texts)
- 1. (2) ash -- (the residue that remains when something is burned)
- 2. (1) ash, ash tree -- (any of various deciduous pinnate-leaved ornamental or timber trees of the genus Fraxinus)
- 3. ash -- (strong elastic wood of any of various ash trees; used for furniture and tool handles and sporting goods such as baseball bats)
- The verb ash has 1 sense (no senses from tagged texts)
- 1. ash -- (convert into ashes)

Example: Extended Lesk (cntd)

- *"On combustion of coal we get ash"*

From Wordnet (through hyponymy)

- ash -- (the residue that remains when something is burned)

  => fly ash -- (fine solid particles of ash that are carried into the air when fuel is combusted)

  => bone ash -- (ash left when bones burn; high in calcium phosphate; used as fertilizer and in bone china)

Critique of Extended Lesk

- Larger region of matching in WordNet
  - Increased chance of Matching

    BUT
  - Increased chance of Topic Drift

WALKER'S ALGORITHM

- A Thesaurus Based approach
- ***Step 1:*** *For each sense of the target word find the thesaurus category to which that sense belongs.*
- ***Step 2:*** *Calculate the score for each sense by using the context words. A context word will add 1 to the score of the sense if the thesaurus category of the word matches that of the sense.*
  - *E.g. The money in this **bank** fetches an interest of 8% per annum*
  - Target word: ***bank***
  - Clue words from the context: ***money, interest, annum, fetch***

	Sense1: Finance	Sense2: Location
Money	+1	0
Interest	+1	0
Fetch	0	0
Annum	+1	0
Total	3	0

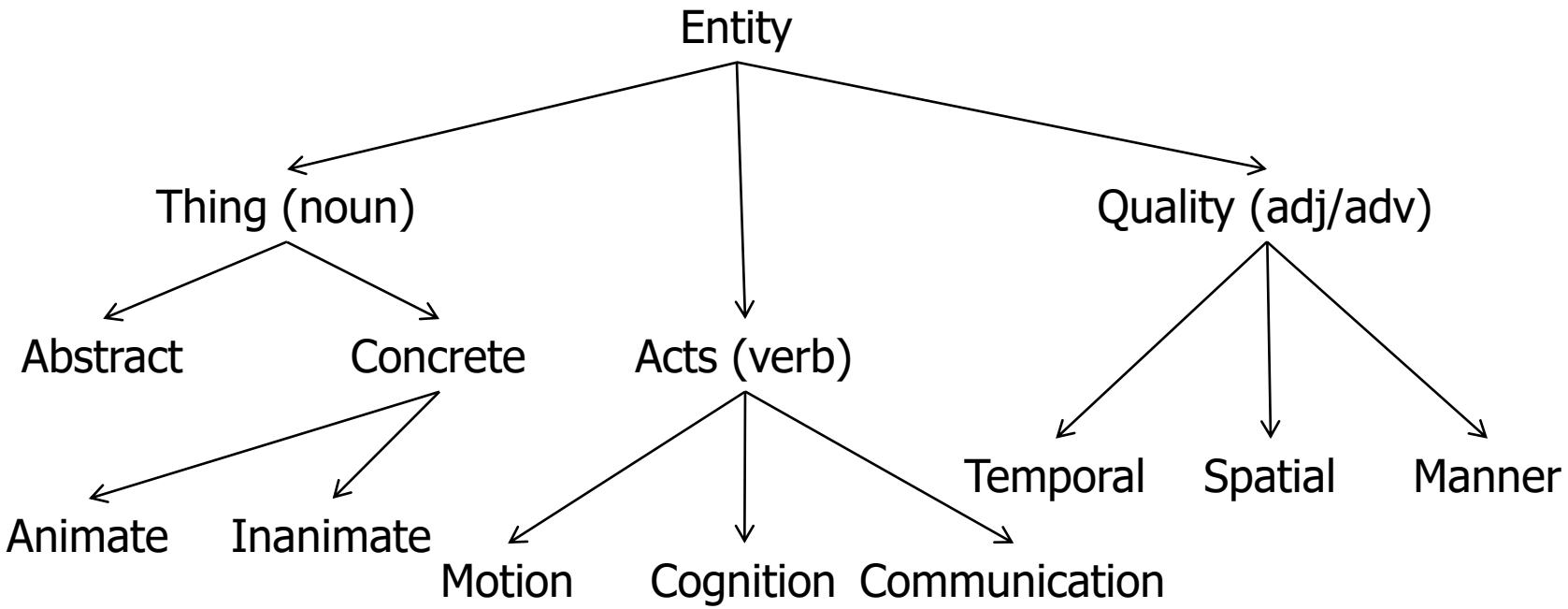
Context words add 1 to the sense when the topic of the word matches that of the sense

WSD USING CONCEPTUAL DENSITY

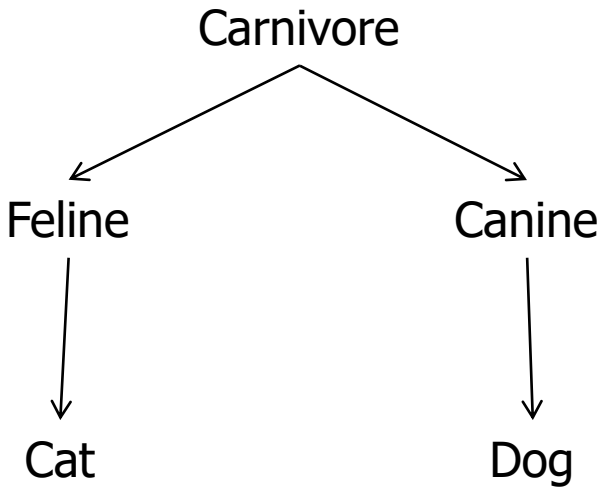
(Agirre and Rigau, 1996)

- Select a sense based on the *relatedness* of that word-sense to the context.
- Relatedness is measured in terms of conceptual distance
  - (i.e. how close the concept represented by the ***word*** and the concept represented by its ***context words*** are)
- This approach uses a structured hierarchical semantic net (*WordNet*) for finding the conceptual distance.
- Smaller the conceptual distance higher will be the conceptual density
  - (i.e. if all words in the context are strong indicators of a particular concept then that concept will have a higher density.)

Fundamental ontology (starting part)



Path length and concept "height"



path_length(cat, dog) = 4

path_length(animate, inanimate) = 2

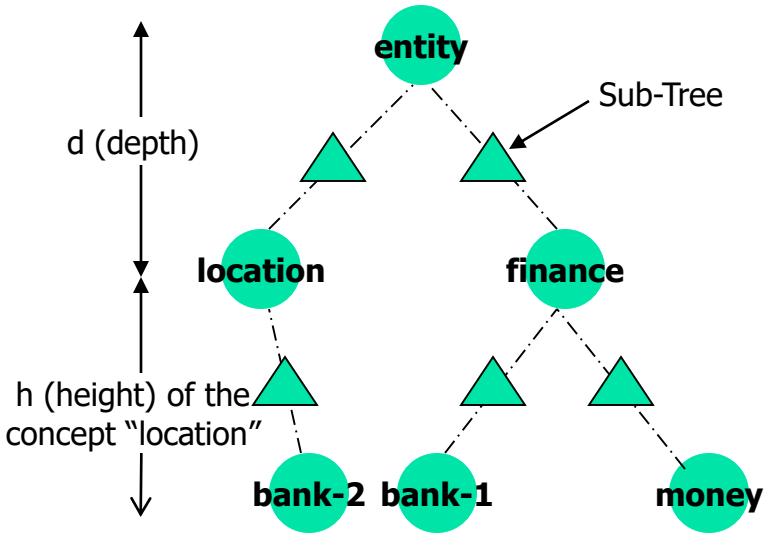
Animate and inanimate are more similar?

- Higher the concept, less specific it is
- Feature vector has less number of components
- Child concept inherits everything of parent plus adds its own
- Entropy is higher at higher levels of conceptual hierarchy
- Semantic similarity will reduce at higher levels

CONCEPTUAL DENSITY FORMULA

Wish list

- The conceptual distance between two word senses should be proportional to the length of the path between the two words in the hierarchical tree (WordNet).
- The conceptual distance between two word senses should be inversely proportional to the depth of the common ancestor concept in the hierarchy.



$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendants_c}$$

where,

c= concept

nhyp = mean number of hyponyms

h= height of the sub-hierarchy

m= no. of senses of the word and senses of context words contained in the sub-hierarchy

CD= Conceptual Density

and 0.2 is the smoothing factor

CONCEPTUAL DENSITY (cntd)

- The dots in the figure represent the senses of the word to be disambiguated or the senses of the words in context.
- The CD formula will yield highest density for the sub-hierarchy containing more senses.
- The sense of W contained in the sub-hierarchy with the highest CD will be chosen.

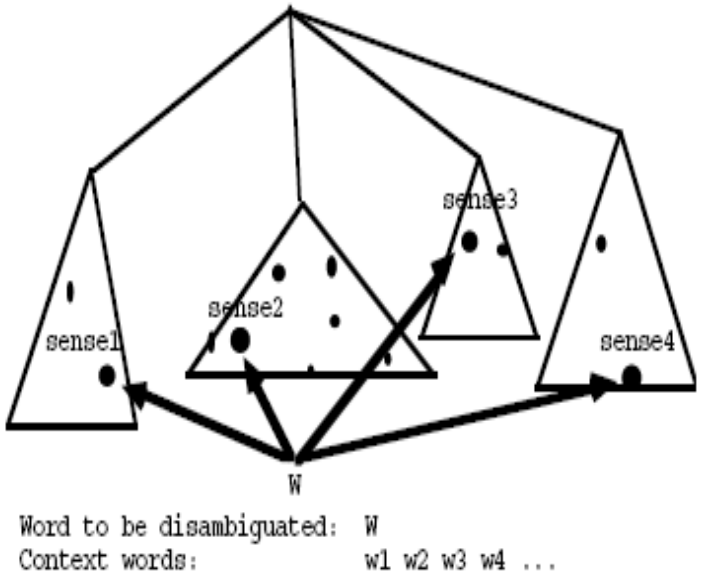
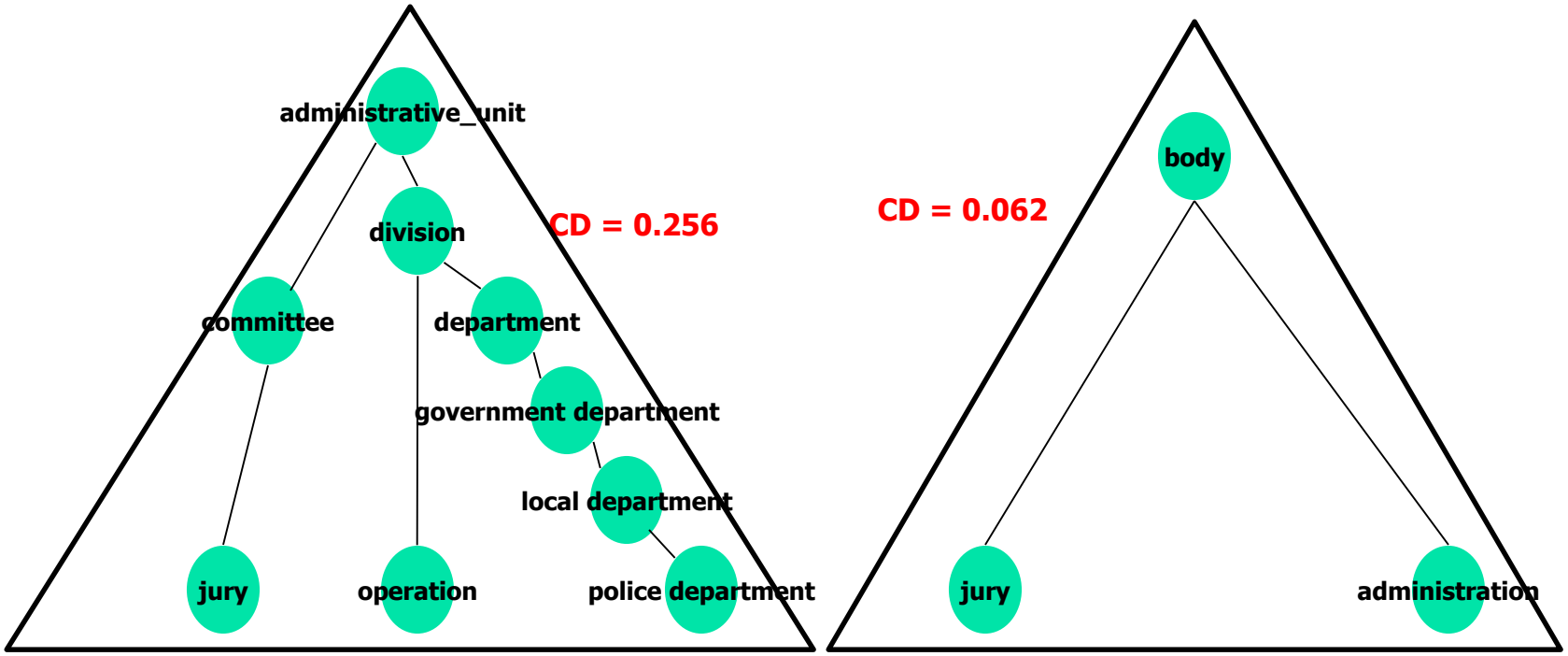


Figure 1: senses of a word in WordNet

CONCEPTUAL DENSITY (EXAMPLE)



The jury(2) praised the administration(3) and **operation** (8) of Atlanta Police Department(1)

Step 1: Make a lattice of the nouns in the context, their senses and hypernyms.

Step 2: Compute the conceptual density of resultant concepts (sub-hierarchies).

Step 3: The concept with the highest CD is selected.

Step 4: Select the senses below the selected concept as the correct sense for the respective words.

CRITIQUE

- Resolves lexical ambiguity of ***nouns*** by finding a combination of senses that maximizes the total Conceptual Density among senses.
- The Good
  - Does not require a tagged corpus.
- The Bad
  - Fails to capture the strong clues provided by proper nouns in the context.
- Accuracy
  - 54% on Brown corpus.

WSD USING RANDOM WALK ALGORITHM (Page Rank) *(sinha and Mihalcea, 2007)*

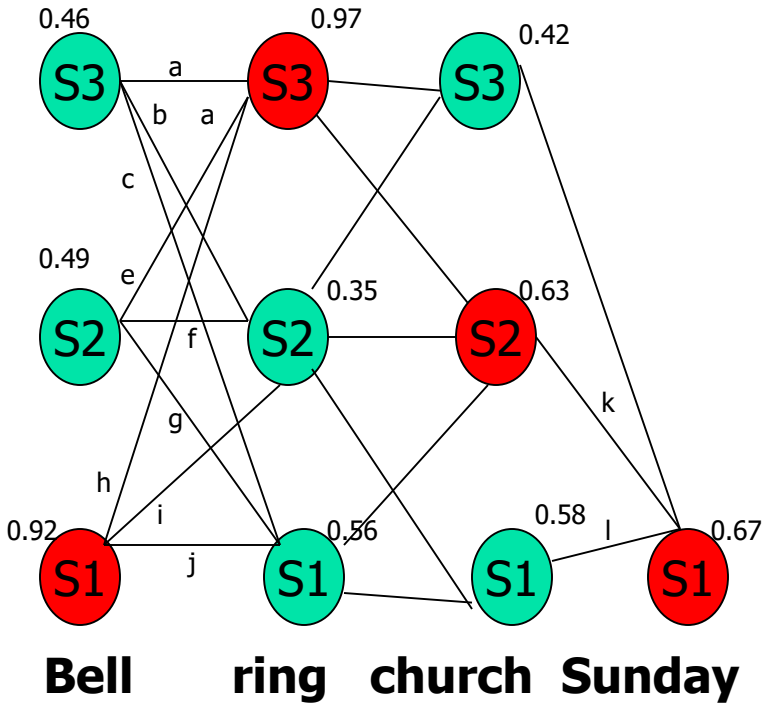
The **church bells** no longer **rung** on **Sundays**.

church
1: one of the groups of Christians who have their own beliefs and forms of worship
2: a place for public (especially Christian) worship
3: a service conducted in a church

bell
1: a hollow device made of metal that makes a ringing sound when struck
2: a push button at an outer door that gives a ringing or buzzing signal when pushed
3: the sound of a bell

ring
1: make a ringing sound
2: ring or echo with sound
3: make (bells) ring, often for the purposes of musical edification

Sunday
1: first day of the week; observed as a day of rest and worship by most Christians



Step 1: Add a vertex for each possible sense of each word in the text.

Step 2: Add weighted edges using definition based semantic similarity (Lesk's method).

Step 3: Apply graph based ranking algorithm to find score of each vertex (i.e. for each word sense).

Step 4: Select the vertex (sense) which has the highest score.

A look at Page Rank (from Wikipedia)

Developed at Stanford University by Larry Page (hence the name *Page*-Rank) and Sergey Brin as part of a research project about a new kind of search engine

The first paper about the project, describing PageRank and the initial prototype of the Google search engine, was published in 1998

Shortly after, Page and Brin founded Google Inc., the company behind the Google search engine

While just one of many factors that determine the ranking of Google search results, PageRank continues to provide the basis for all of Google's web search tools

A look at Page Rank (cntd)

PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page.

Assume a small universe of four web pages: **A**, **B**, **C** and **D**.

The initial approximation of PageRank would be evenly divided between these four documents. Hence, each document would begin with an estimated PageRank of 0.25.

If pages **B**, **C**, and **D** each only link to **A**, they would each confer 0.25 PageRank to **A**. All PageRank **PR( )** in this simplistic system would thus gather to **A** because all links would be pointing to **A**.

PR(A)=PR(B)+PR(C)+PR(D)

This is 0.75.

A look at Page Rank (cntd)

Suppose that page **B** has a link to page **C** as well as to page **A**, while page **D** has links to all three pages

The *value of the link-votes is divided among all the outbound links on a page*.

Thus, page **B** gives a vote worth 0.125 to page **A** and a vote worth 0.125 to page **C**.

Only one third of **D**'s PageRank is counted for A's PageRank (approximately 0.083).

PR(A)=PR(B)/2+PR(C)/1+PR(D)/3

In general,

$$\mathbf{PR(U)= \sum_{V \varepsilon B(U)} PR(V)/L(V),}$$ where B(u) is the set of pages u is linked to, and L(V) is the number of links from V

A look at Page Rank (damping factor)

The PageRank theory holds that even an imaginary surfer who is randomly clicking on links will eventually stop clicking.

The probability, at any step, that the person will continue is a damping factor *d*.

$$PR(U) = (1-d)/N + d.\sum_{V \varepsilon B(U)} PR(V)/L(V),$$

N=size of document collection

For WSD: Page Rank

- Given a graph G = (V,E)
  - In(V_i) = predecessors of V_i
  - Out(V_i) = successors of V_i

$$S(V_i) = \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j)$$

- In a weighted graph, the walker randomly selects an outgoing edge with higher probability of selecting edges with higher weight.

$$WS(V_i) = \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

Other Link Based Algorithms

- *HITS* algorithm invented by Jon Kleinberg (used by Teoma and now Ask.com)
- IBM *CLEVER project*
- *TrustRank* algorithm.

CRITIQUE

- Relies on random walks on graphs encoding label dependencies.
- The Good
  - Does not require any tagged data (a wordnet is sufficient).
  - The weights on the edges capture the definition based semantic similarities.
  - Takes into account global data recursively drawn from the entire graph.
- The Bad
  - Poor accuracy
- Accuracy
  - 54% accuracy on SEMCOR corpus which has a baseline accuracy of 37%.

KB Approaches – Comparisons

Algorithm	Accuracy
WSD using Selectional Restrictions	44% on Brown Corpus
Lesk's algorithm	50-60% on short samples of *"Pride and Prejudice"* and some *"news stories"*.
Extended Lesk's algorithm	32% on Lexical samples from Senseval 2 (Wider coverage).
WSD using conceptual density	54% on Brown corpus.
WSD using Random Walk Algorithms	54% accuracy on SEMCOR corpus which has a baseline accuracy of 37%.
Walker's algorithm	50% when tested on 10 highly polysemous English words.

KB Approaches – Conclusions

- Drawbacks of WSD using Selectional Restrictions
  - Needs exhaustive Knowledge Base.
- Drawbacks of Overlap based approaches
  - Dictionary definitions are generally very small.
  - Dictionary entries rarely take into account the distributional constraints of different word senses (e.g. selectional preferences, kinds of prepositions, etc. → **c*igarette*** and ***ash*** never co-occur in a dictionary).
  - Suffer from the problem of sparse match.
  - Proper nouns are not present in a MRD. Hence these approaches fail to capture the strong clues provided by proper nouns.

SUPERVISED APPROACHES

NAÏVE BAYES

- The Algorithm find the winner sense using

$$\hat{s} = \text{argmax}_{s \in \text{senses}} \Pr(s|V_w)$$

- 'V_w' is a feature vector consisting of:
  - POS of w
  - Semantic & Syntactic features of w
  - Collocation vector (set of words around it) → typically consists of next word(+1), next-to-next word(+2), -2, -1 & their POS's
  - Co-occurrence vector (number of times w occurs in bag of words around it)

- Applying Bayes rule and naive independence assumption

$$\hat{s} = \text{argmax}_{s \in \text{senses}} \Pr(s).\Pi_{i=1}^{n} \Pr(V_w^i|s)$$

BAYES RULE AND INDEPENDENCE ASSUMPTION

$$\hat{s} = \text{argmax}_{s \in \text{senses}} \Pr(s|V_w)$$

where V_w is the feature vector.

- Apply Bayes rule:

$$\Pr(s|V_w) = \Pr(s).\Pr(V_w|s)/\Pr(V_w)$$

$$\hat{s} = \text{argmax}_{s \in \text{senses}} \Pr(s|V_w)$$

- $\Pr(V_w|s)$ can be approximated by independence assumption:

$$\Pr(V_w|s) = \Pr(V_w^1|s).\Pr(V_w^2|s,V_w^1)...\Pr(V_w^n|s,V_w^1,..,V_w^{n-1})$$

$$= \Pi_{i=1}^{n} \Pr(V_w^i|s)$$

Thus,

$$\hat{s} = \text{argmax}_{s \hat{I} \text{senses}} \Pr(s).\Pi_{i=1}^{n} \Pr(V_w^i|s)$$

ESTIMATING PARAMETERS

- Parameters in the probabilistic WSD are:
  - **Pr(s)**
  - **Pr(V_w^i|s)**
- Senses are marked with respect to sense repository (WORDNET)

Pr(s) = count(s,w) / count(w)

Pr(V_w^i|s) = Pr(V_w^i,s)/Pr(s)

$$= \frac{\text{count}(V_w^i, s, w)/\text{count}(w)}{\text{count}(s, w)/\text{count}(w)}$$

= c(V_w^i,s,w)/c(s,w)

DECISION LIST ALGORITHM

- Based on 'One sense per collocation' property
  - Nearby words provide strong and consistent clues as to the sense of a target word.
- Collect a large set of collocations for the ambiguous word
- Calculate word-sense probability distributions for all such collocations.
- Calculate the log-likelihood ratio

$$\text{Log}\left(\frac{\Pr(\text{Sense-A} \mid \text{Collocation}_i)}{\Pr(\text{Sense-B} \mid \text{Collocation}_i)} \right)$$

Assuming there are only two senses for the word. Of course, this can easily be extended to 'k' senses.

- Higher log-likelihood = more predictive evidence
- Collocations are ordered in a ***decision list***, with most predictive collocations ranked highest.

DECISION LIST ALGORITHM (CONTD.)

Training Data

Sense	Training Examples (Keyword in Context)
A	used to strain microscopic *plant* life from the ...
A	... zonal distribution of *plant* life . ...
A	close-up studies of *plant* life and natural ...
A	too rapid growth of aquatic *plant* life in water ...
A	... the proliferation of *plant* and animal life ...
A	establishment phase of the *plant* virus life cycle ...
A	... ...
B	... ...
B	computer **manufacturing** *plant* and adjacent ...
B	discovered at a St. Louis *plant* **manufacturing**
B	... copper **manufacturing** *plant* found that they
B	copper wire **manufacturing** *plant* , for example ...
B	's cement **manufacturing** *plant* in Alpena ...
B	polystyrene **manufacturing** *plant* at its Dow ...
B	company **manufacturing** *plant* is in Orlando ...

Resultant Decision List

Final decision list for *plant* (abbreviated)

LogL	Collocation	Sense
10.12	*plant* growth	⇒ A
9.68	car (within ±*k* words)	⇒ B
9.64	*plant* height	⇒ A
9.61	union (within ±*k* words)	⇒ B
9.54	equipment (within ±*k* words)	⇒ B
9.51	assembly *plant*	⇒ B
9.50	nuclear *plant*	⇒ B
9.31	flower (within ±*k* words)	⇒ A
9.24	job (within ±*k* words)	⇒ B
9.03	fruit (within ±*k* words)	⇒ A
9.02	*plant* species	⇒ A
...	...	

Classification of a test sentence is based on the highest ranking collocation found in the test sentence.

E.g.

…plucking **flowers** affects ***plant growth***…

CRITIQUE

- Harnesses powerful, empirically-observed properties of language.
- The Good
  - Does not require large tagged corpus. Simple implementation.
  - Simple semi-supervised algorithm which builds on an existing supervised algorithm.
  - Easy understandability of resulting decision list.
  - Is able to capture the clues provided by Proper nouns from the corpus.
- The Bad
  - The classifier is word-specific.
  - A new classifier needs to be trained for every word that you want to disambiguate.
- Accuracy
  - Average accuracy of **96%** when tested on a set of 12 highly polysemous words.

Exemplar Based WSD (k-nn)

- An exemplar based classifier is constructed for each word to be disambiguated.
- **Step1:** From each ***sense marked sentence*** containing the ambiguous word , a training example is constructed using:
  - POS of ***w*** as well as POS of neighboring words.
  - Local collocations
  - Co-occurrence vector
  - Morphological features
  - Subject-verb syntactic dependencies
- **Step2:** Given a test sentence containing the ambiguous word, a test example is similarly constructed.
- **Step3:** The test example is then compared to all training examples and the k-closest training examples are selected.
- **Step4:** The sense which is most prevalent amongst these "k" examples is then selected as the correct sense.

WSD Using SVMs

- **Training Phase:** Using a tagged corpus, f or every sense of the word a SVM is trained using the following features:
  - POS of ***w*** as well as POS of neighboring words.
  - Local collocations
  - Co-occurrence vector
  - Features based on syntactic relations (e.g. headword, POS of headword, voice of head word etc.)
- **Testing Phase:** Given a test sentence, a test example is constructed using the above features and fed as input to each binary classifier.
- The correct sense is selected based on the label returned by each classifier.

WSD Using Perceptron Trained HMM

- WSD is treated as a sequence labeling task.
- The class space is reduced by using WordNet's super senses instead of actual senses.
- A discriminative HMM is trained using the following features:
  - POS of ***w*** as well as POS of neighboring words.
  - Local collocations
  - Shape of the word and neighboring words

    E.g. for s = "Merrill Lynch & Co shape(s) =Xx*Xx*&Xx
- Lends itself well to NER as labels like "person", location", "time" etc are included in the super sense tag set.

Supervised Approaches – Comparisons

Approach	Average Precision	Average Recall	Corpus	Average Baseline Accuracy
Naïve Bayes	64.13%	Not reported	Senseval3 – All Words Task	60.90%
Decision Lists	96%	Not applicable	Tested on a set of 12 highly polysemous English words	63.9%
Exemplar Based disambiguation (k-NN)	68.6%	Not reported	WSJ6 containing 191 content words	63.7%
SVM	72.4%	72.4%	Senseval 3 – Lexical sample task (Used for disambiguation of 57 words)	55.2%
Perceptron trained HMM	67.60	73.74%	Senseval3 – All Words Task	60.90%

Supervised Approaches – Conclusions

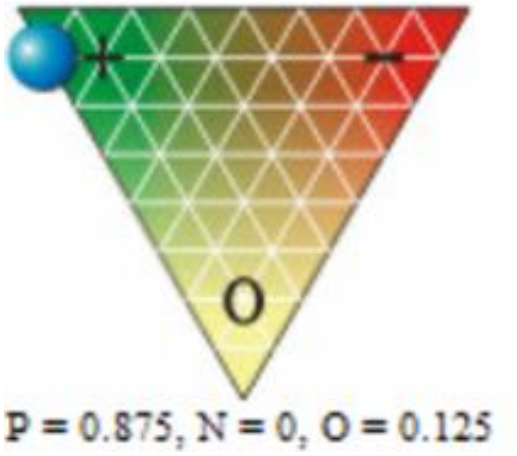
- General Comments
  - Use corpus evidence instead of relying of dictionary defined senses.
  - Can capture important clues provided by proper nouns because proper nouns do appear in a corpus.
- Naïve Bayes
  - Suffers from data sparseness.
  - Since the scores are a product of probabilities, some weak features might pull down the overall score for a sense.
  - A large number of parameters need to be trained.
- Decision Lists
  - A word-specific classifier. A separate classifier needs to be trained for each word.
  - Uses the single most predictive feature which eliminates the drawback of Naïve Bayes.

Sentiwordnet

Need for sentiment resource

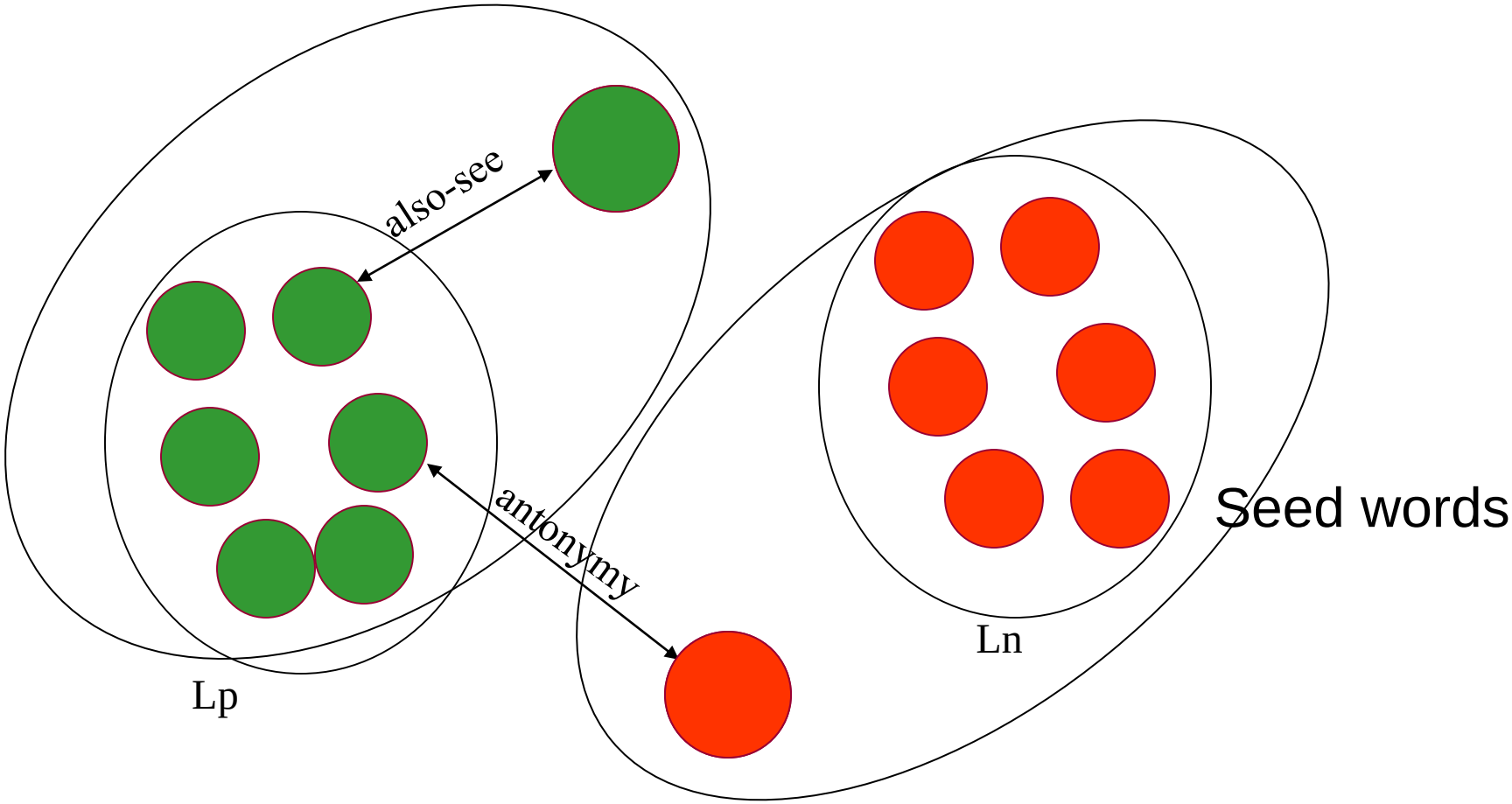
Reference : [Esuli et al,2006]

SentiWordnet



Happy

Seed-set expansion in SWN



The sets at the end of kth step are called Tr(k,p) and Tr(k,n)

Tr(k,o) is the set that is not present in Tr(k,p) and Tr(k,n)

Building SentiWordnet

Using SentiWordnet scores

pestering
P = 0,
N = 0.625,
O = 0.375

Freebase: A Collaboratively Created Database

Bollacker, Kurt, et al. "Freebase: a collaboratively created graph database for structuring human knowledge." *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008.

http://wiki.freebase.com/wiki/T1:Freebase_Tutorial

(ack: Naman for his slides)

Motivation

- **Web** : Repository of unstructured(unorganised) information.

- Unstructured → Structured Data
  - Example : Relational Database.

- Problem with Relational Databases
  - Suppose we want to store data about the person say **X** --who composes songs, sings and performs, write books and acts in movies-- which relational table should we use?

Data in many tables

- Following Relational tables would be required :
  - ✓ Song Composer
  - ✓ Singer
  - ✓ Book Author
  - ✓ Film Actor

- Data about same person in different tables

- **COMPLEX** Query

Freebase

- Open, structured, linked knowledge graph captures human Knowledge.
- Build around entities. An entity can be person, place or thing.
- Information of over 30 million entities.
- Information extracted from communities like wikipedia, university Libraries, Wikimedia Commons etc.

Terminologies

- **TOPIC** : Analogous to article in Wikipedia
  - ✓ Physical Entities like TAJ Hotel, IIT Bombay, Dominos Restaurant .
  - ✓ Abstract concepts like love, color etc.

- **TYPES** : **Multi-faceted** nature of topics. Multiple aspects to the same topic.
  - ✓ Farhan Akhtar is a song writer, singer, performer, book author, and a film actor.
  - ✓ Leonardo da Vinci was a painter, a sculptor, an architect, an engineer.

- **PROPERTIES** : A type has a set of properties corresponding to a particular topic**.**
  - ✓ Properties of **company** type : Company's founders, Board members, Employees, Products, Revenue etc.

Terminologies Contd..

- **Domains and ID's :** Types are grouped into *domains*. Each domain is given an ID (identifier), e.g.,
  - ✓ /business - The Business domain
  - ✓ /music - The Music domain

Query Freebase

- Freebase uses MQL(MetaWeb Query Language) : API for making programmatic queries to freebase

- It uses **JSON (Javascript Object Notation)** objects as queries.

```
[{
"id":"/en/swades",
"/film/film/directed_by":[{
"name":null
}]
}]
```

```
{
  "result": [{
    "/film/film/directed_by": [{
      "name": "Ashutosh Gowariker"
    }],
    "id": "/en/swades"
  }]
}
```

ConceptNet

ConceptNet

- From MIT (Liu and Singh, 2004)
- Capture common sense information
- Emphasis on everyday knowledge rather than rigorous linguistic lexical differentiations (unlike wordnet)

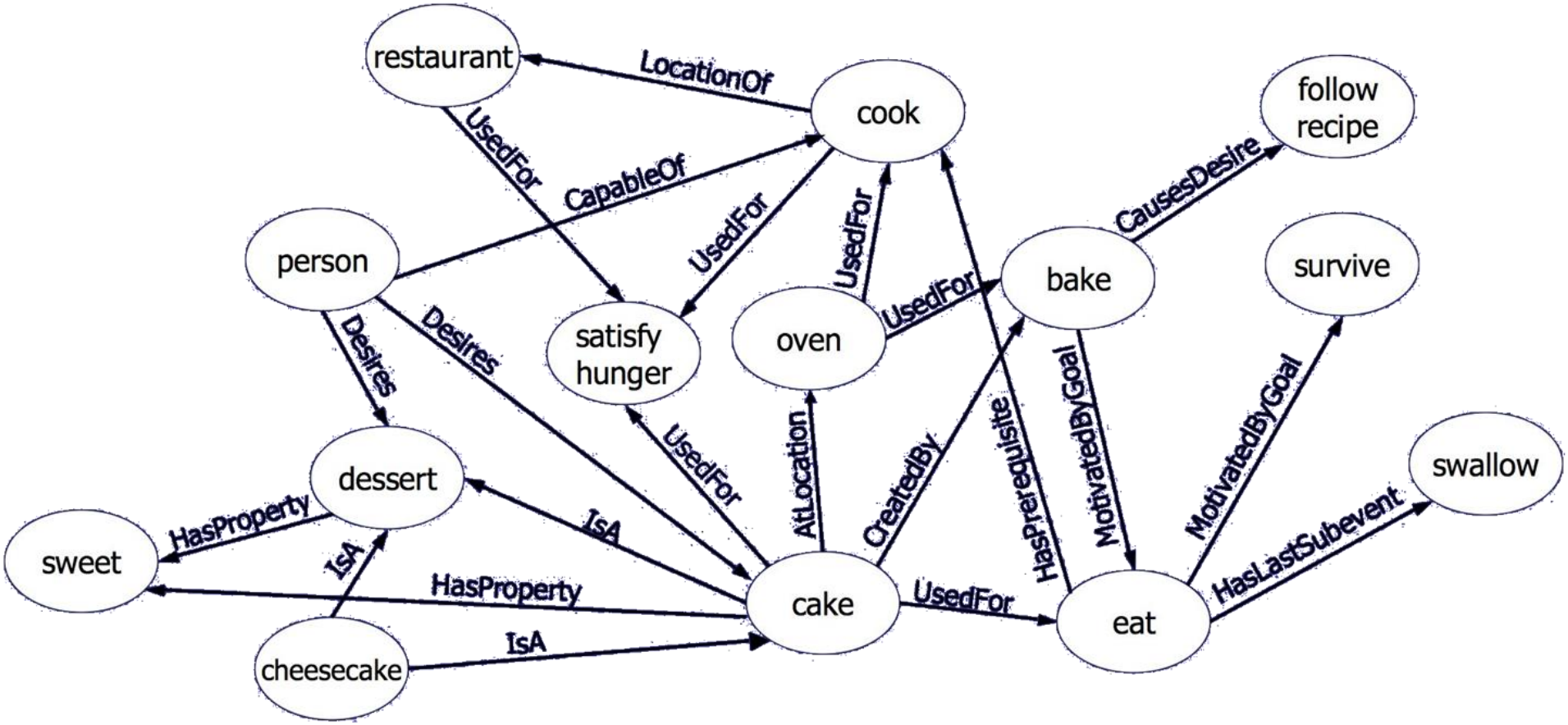
Projects to Collect Commonsense[1]

- Cyc
  - Started in 1984 by Dr. Doug Lenat
  - Developed by CyCorp, with 3.2 millions of assertions linking over 280000 concepts and using thousands of micro-theories.
  - Cyc-NL is still a “potential application”, knowledge representation in frames is quite complicated and thus difficult to use.

Projects to Collect Commonsense^2

- Open Mind Common Sense Project
  - Started in 2000 at MIT by Push Singh
  - WWW collaboration with over 20,123 registered users, who contributed 812,769 items
  - Used to generate *ConceptNet*, very large semantic network.

ConceptNet: Example1



ConceptNet: Example2

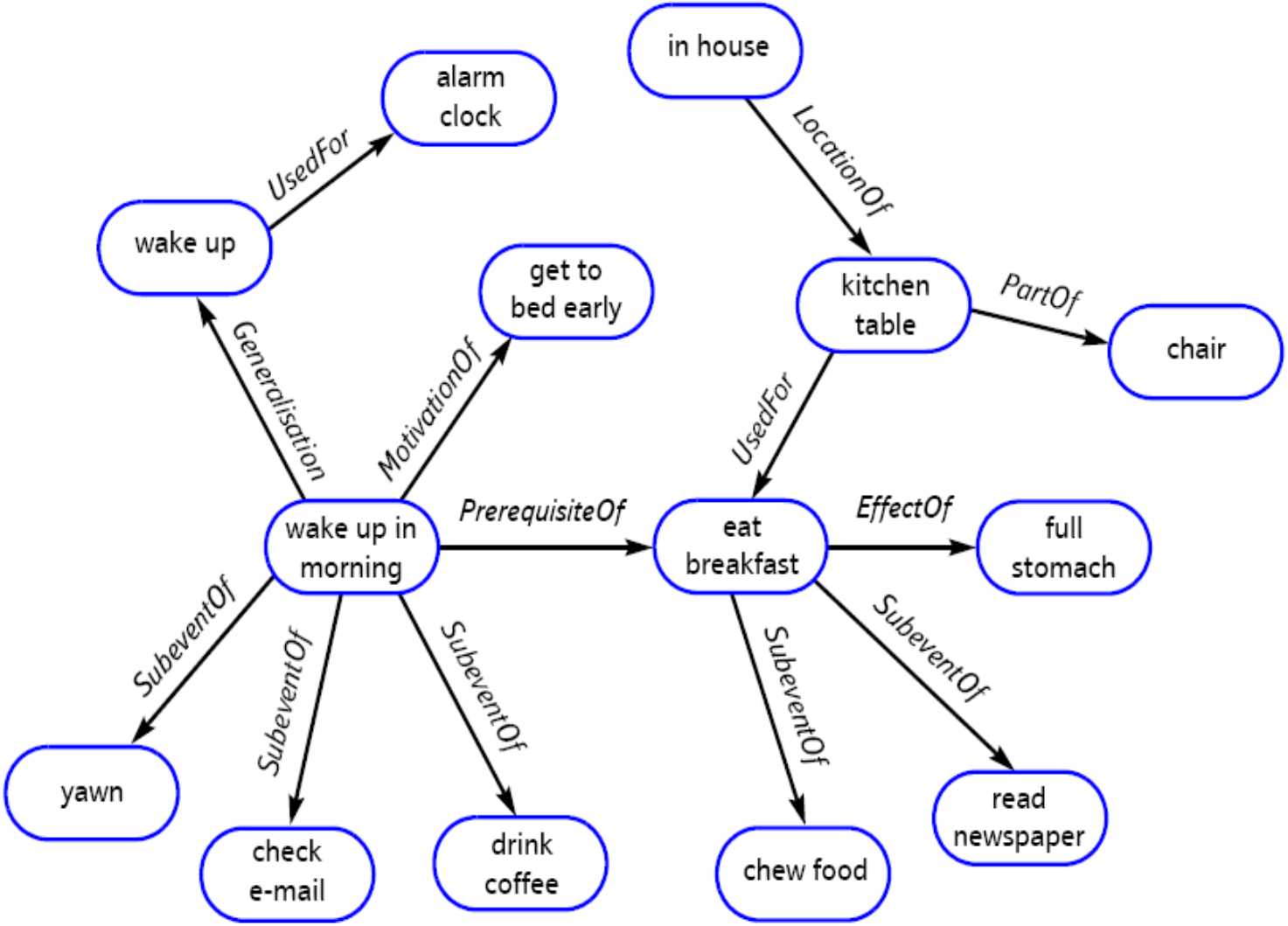


Fig 2 A subset of ConceptNet.

“I borrowed ‘Treasure Island’ for a couple of weeks”

- 1. `Treasure Island' is the name of a book
- 2. People borrow books to read
- 3. The book was most likely borrowed from a library
- 4. The book has to be returned to the lender in 14 days time

Flavors of common sense knowledge

- Emotive (“I feel awful")
- Functional (“Cups hold liquids")
- Causal (“Extracting a tooth causes pain")
- spatial (“Horses are usually found in stables")

OMCS *(Singh et al, 2004)*

- Open Mind Common Sense Project
- Volunteers contribute assertions (crowdsourcing?)
- 30 different activities of everyday life
- Short semi structured sentences (as opposed to synsets)
- Volunteers follow “patterns” in the “help” menu

Example of patterns

- *'Treasure Island' is a book: is-a pattern*
- *Books are found in a library: is-located-in pattern*
- Patterns make it possible to create machine processable data

ConceptNet: structure

- Directed Acyclic Graph formed by linking over 1.5 million assertions into a semantic network of about 300000 nodes
- Each node: a fragment of text (unlike synset) *aka* "concept"
- Nodes
  - NP: "watermelon"
  - VP: "breath air"

ConceptNet: Relations[1]

- 20 relations grouped into 8 thematic types
  - 1. K-Lines: ConceptuallyRelatedTo, ThematicKLine, SuperThematicK-Line
  - 2. Things: IsA, Prop ertyOf, PartOf, MadeOf, DefnedAs
  - 3. Agents: CapableOf

  (note the difference from wordnet lexicon semantic relations like antonymy, hypernymy etc.)

ConceptNet: Relations[2]

- Themes:
  - 4. Events: PrerequisiteEvent, FirstSubEventOf, LastSubEventOf, SubEventOf
  - 5. Spatial: LocationOf
  - 6. Causal: EffectOf, DesirousEffectOf
  - 7. Functional: UsedFor, CapableOfReceivingAction
  - 8. Affective: MotivationOf, DesireOf

Twenty Semantic Relation Types in ConceptNet (Liu and Singh, 2004)	
THINGS (52,000 assertions)	IsA: (IsA "apple" "fruit") Part of: (PartOf "CPU" "computer") PropertyOf: (PropertyOf "coffee" "wet") MadeOf: (MadeOf "bread" "flour") DefinedAs: (DefinedAs "meat" "flesh of animal")
EVENTS (38,000 assertions)	PrerequisiteeventOf: (PrerequisiteEventOf "read letter" "open envelope") SubeventOf: (SubeventOf "play sport" "score goal") FirstSubeventOF: (FirstSubeventOf "start fire" "light match") LastSubeventOf: (LastSubeventOf "attend classical concert" "applaud")
AGENTS (104,000 assertions)	CapableOf: (CapableOf "dentist" "pull tooth")
SPATIAL (36,000 assertions)	LocationOf: (LocationOf "army" "in war")
TEMPORAL time & sequence	
CAUSAL (17,000 assertions)	EffectOf: (EffectOf "view video" "entertainment") DesirousEffectOf: (DesirousEffectOf "sweat" "take shower")
AFFECTIONAL (mood, feeling, emotions) (34,000 assertions)	DesireOf (DesireOf "person" "not be depressed") MotivationOf (MotivationOf "play game" "compete")
FUNCTIONAL (115,000 assertions)	IsUsedFor: (UsedFor "fireplace" "burn wood") CapableOfReceivingAction: (CapableOfReceivingAction "drink" "serve")
ASSOCIATION K-LINES (1.25 million assertions)	SuperThematicKLine: (SuperThematicKLine "western civilization" "civilization") ThematicKLine: (ThematicKLine "wedding dress" "veil") ConceptuallyRelatedTo: (ConceptuallyRelatedTo "bad breath" "mint")

Table 1 ConceptNet's relational ontology of 20 link types.

ConceptuallyRelatedTo	IsA	FirstSubeventOf	DesirousEffectOf
ThematicKLine	MadeOf	SubeventOf	UsedFor
SuperThematicKLine	DefinedAs	LastSubeventOf	LocationOf
CapableOfReceivingAction	CapableOf	PrerequisiteEventOf	MotivationOf
PropertyOf	PartOf	EffectOf	DesireOf

Table 2 Ontology of concept types.

Events	Things	Places	Properties
Eat sandwich	Orange juice	At zoo	Furry
Sell car	Morning coffee	On table	Very expensive
Tell story	Policeman	Near school	Dark
Go to zoo	Leaf blower	Inside oven	Quickly
Type letter	Laptop computer	In closet	Dark

Sentence→ConceptNet

- Extraction
  - 50 regular expression rules run on OMCS sentences
  - Database creation
- Normalization
  - Spell correction
  - Stop word removal if needed
  - Lemmatization
- Relaxation

Database of ConceptNet

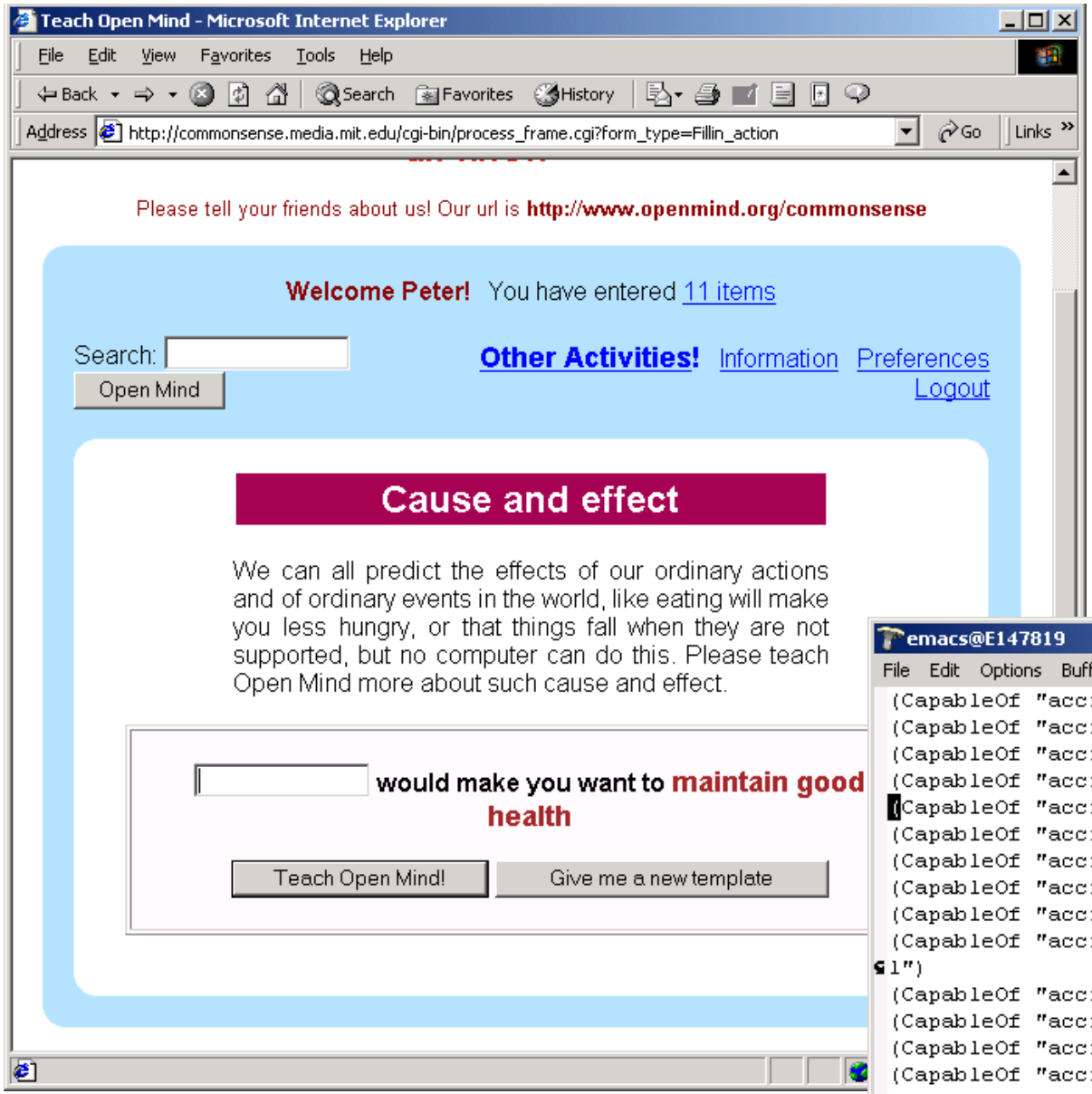
- Relations and facts are stored
  - (IsA “spider" “bug" “f=3; i=0;")
  - (LocationOf “waitress" “in restaurant" “f=2; i=0;")
- Frequency of seeing the assertion recorded as the number of times this relation was found through an inferencing procedure

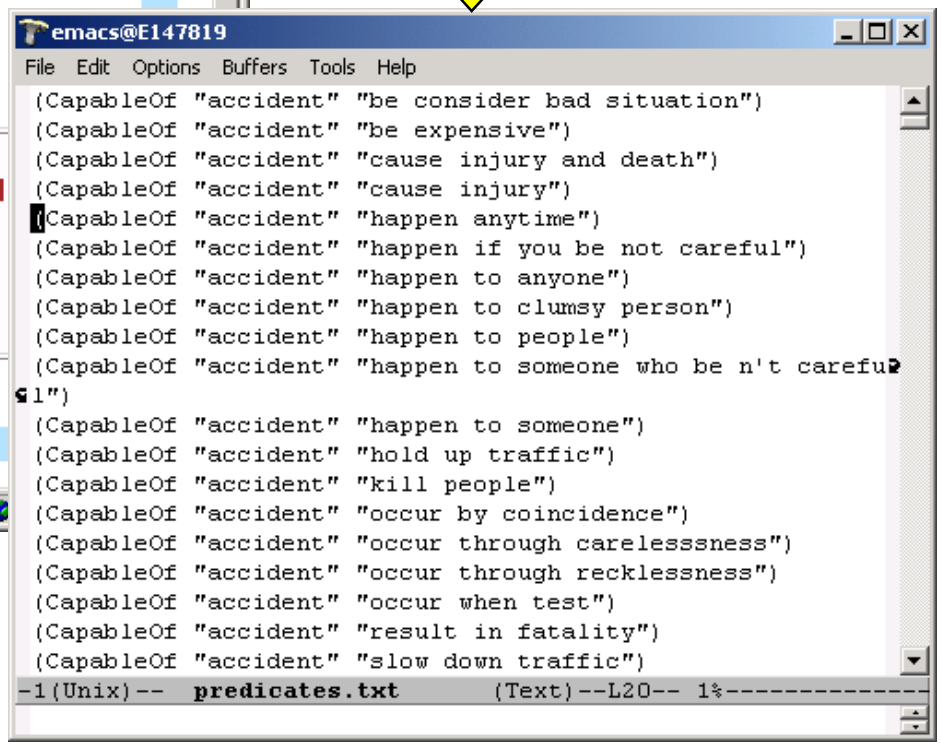
Inference in ConceptNet

- Multiple assertions inferred from a single Open Mind sentence
- Example: ‘A lime is a sour fruit’
- Infer *IsA(lime, fruit)*
- Additionally infer *PropertyOf(lime, sour)*
- Infer Generalisations
  - if the majority of fruits have the property ‘sweet’,
  - then this property is lifted to the parent class, as: Property Of(fruit, sweet).

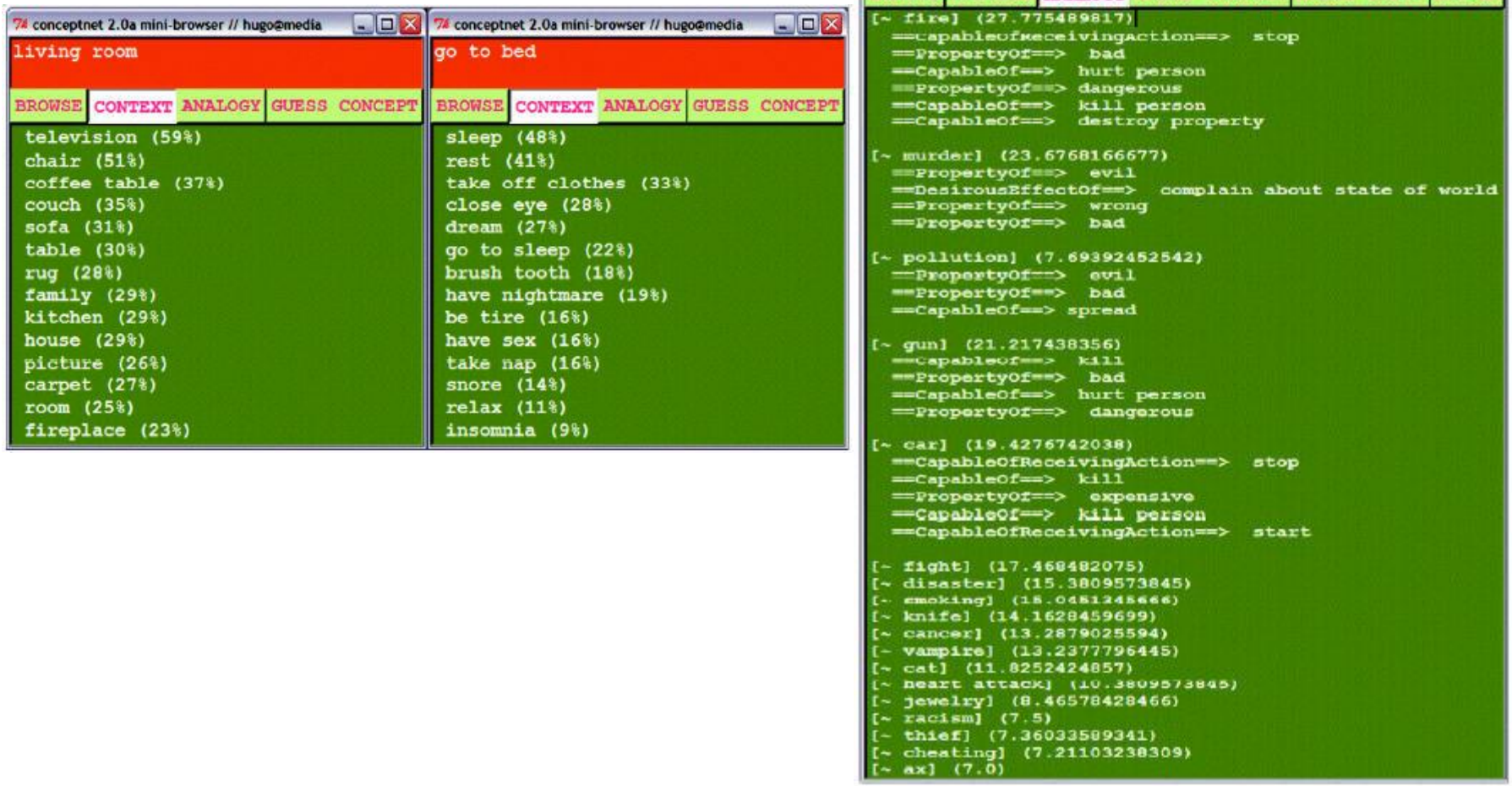
MontyLingua NLP engine

- Textual information management
- Written in Python, also available in Java
- Liberates ConceptNet from normalization of text
- Can take running paras and sentences
- API (2003) — http://web.media.mit.edu/~hugo/montylingua


Teach Open Mind - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://commonsense.media.mit.edu/cgi-bin/process_frame.cgi?form_type=Fillin_action
Go Links
Please tell your friends about us! Our url is http://www.openmind.org/commonsense
Welcome Peter! You have entered 11 items
Search:
Open Mind
Other Activities! Information Preferences Logout
Cause and effect
We can all predict the effects of our ordinary actions and of ordinary events in the world, like eating will make you less hungry, or that things fall when they are not supported, but no computer can do this. Please teach Open Mind more about such cause and effect.
would make you want to maintain good health
Teach Open Mind!
Give me a new template


emacs@E147819
File Edit Options Buffers Tools Help
(CapableOf "accident" "be consider bad situation")
(CapableOf "accident" "be expensive")
(CapableOf "accident" "cause injury and death")
(CapableOf "accident" "cause injury")
(CapableOf "accident" "happen anytime")
(CapableOf "accident" "happen if you be not careful")
(CapableOf "accident" "happen to anyone")
(CapableOf "accident" "happen to clumsy person")
(CapableOf "accident" "happen to people")
(CapableOf "accident" "happen to someone who be n't careful")
(CapableOf "accident" "happen to someone")
(CapableOf "accident" "hold up traffic")
(CapableOf "accident" "kill people")
(CapableOf "accident" "occur by coincidence")
(CapableOf "accident" "occur through carelesssness")
(CapableOf "accident" "occur through recklessness")
(CapableOf "accident" "occur when test")
(CapableOf "accident" "result in fatality")
(CapableOf "accident" "slow down traffic")
-1(Unix)-- predicates.txt (Text)--L20-- 1%

Snapshot of ConceptNet



ConceptNet Application[1]

- Commonsense ARIA
  - Observes a user writing an e-mail and proactively suggests photos relevant to the user's story
  - Bridges semantic gaps between annotations and the user's story
- GOOSE
  - A goal-oriented search engine for novice users
  - Generate the search query

ConceptNet Application[2]

- Makebelieve
  - Story-generator that allows a person to interactively invent a story with the system
  - Generate causal projection chains to create storylines
- GloBuddy: A dynamic foreign language phrasebook
- AAA: Rrecommends products from Amazon.com by using ConceptNet to reason about a person's goals and desires,creating a profile of their predicted tastes.

HowNet

Acknowledgement: Zhendon Dong, GWC 2006 presentation

What is HowNet?

- HowNet: on-line knowledge system for the computation of meaning in HLT.
- Inter-concept relations and inter-attribute relations of the concepts as connoted in its Chinese-English lexicon
- Released in 1999

Statistics - general

Chinese word & expression	84102
English word & expression	80250
Chinese meaning	98530
English meaning	100071
Definition	25295
Record	161743

A record in HowNet dictionary

NO.=076856

W_C=买主

G_C=N [mai3 zhu3]

E_C=

W_E=buyer

G_E=N

E_E=

DEF={human|人:domain={commerce|商业},{buy|买:
agent={~}}}

Statistics - semantic

	Chinese	English
Thing	58153	58096
Component	7025	7023
Time	2238	2244
Space	1071	1071
Attribute	3776	4045
Atttibute-value	9089	8478
Event	12634	10076

Statistics – main syntactic categories

	Chinese	English
ADJ	11705	9576
ADV	1516	2084
VERB	25929	21017
NOUN	46867	48342
PRON	112	71
NUM	225	242
PREP	128	113
AUX	77	49
CLA	424	0

Statistics – part of relations

Chinese synset: Set = 13463

Word Form = 54312

antonym: Set = 12777

converse: Set = 6753

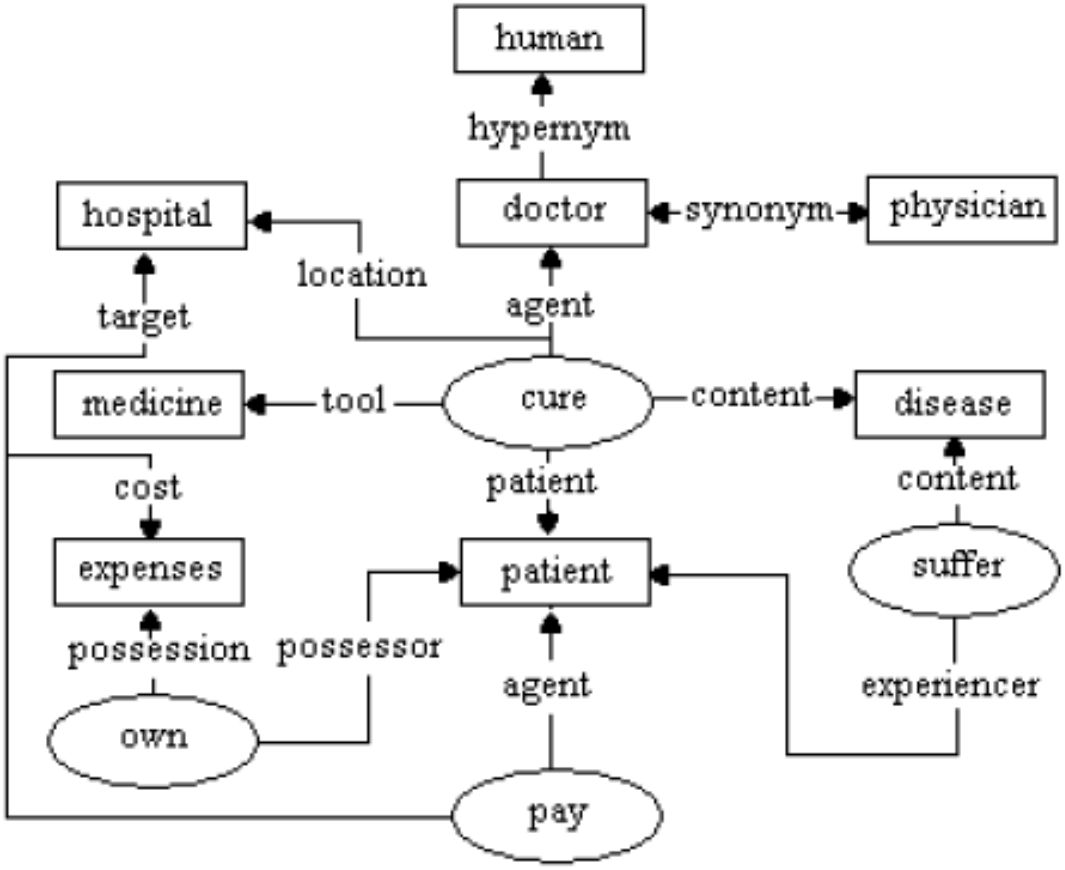
English synset: Set = 18575

Word Form = 58488

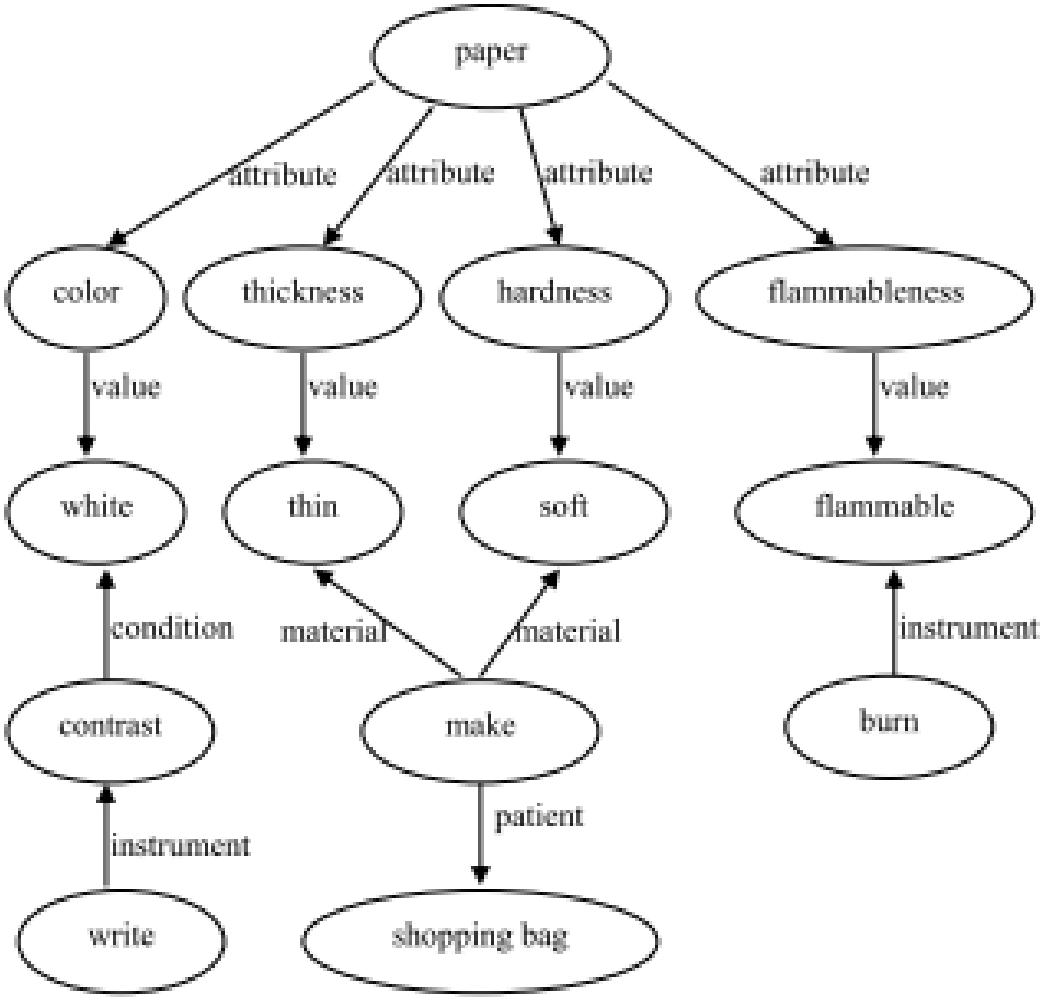
antonym: Set = 12032

converse: Set = 6442

Conceptual Relation Network (CRN) of "Doctor"



Attribute Relation Network (ARN) of "Paper"



4 unique features of HowNet: (1) **Sememe**

- Sememes are regarded as the basic unit of the meaning
- E.g., “paper”, “doctor”

4 unique features of HowNet: (2) **Definition**

- Each concept in HowNet lexicon is defined in a language, called Knowledge Database Markup Language (KDML)
- The KDML is mainly composed of sememes and semantic roles

4 unique features of HowNet: **Definition** ***(contd)***

- Knowledge Database Mark-up Language uses 2089 sememes, 128 secondary features and 94 semantic roles as its vocabulary
- Adopts extended BNF

4 unique features of HowNet: **Definition** ***(contd):*** **"doctor"**

$$DEF=\{human|人:HostOf=\{Occupation|职位\},$$
$$domain=\{medical|医\},\{doctor|医治:agent=\{\sim\}\}$$

4 unique features of HowNet: (3) self sufficiency and (4) language independence

- Systematic integration of hierarchical taxonomies, axiomatic inference, KDML-defined concepts
- (HowNet claim) "Only with the HowNet's shared definitions can we achieve a shared ontology for all languages"

Defining concepts (1)

W_E=doctor
G_E=V
DEF={doctor|医治}

W_E=doctor
G_E=N
DEF={human|人:HostOf={Occupation|职位},domain={medical|医},
{doctor|医治:agent={~}}}

W_E=doctor
G_E=N
E_E=
DEF={human|人:{own|有:possession={Status|身分:
domain={education|教育},modifier={HighRank|高等:
degree={most|最}}},possessor={~}}}

Defining concepts (2)

W_E=buy

G_E=V

DEF={buy|买}

cf. (WordNet) obtain by purchase; acquire by means of finacial transaction

W_E=buy

G_E=V

DEF={GiveAsGift|赠:manner={guilty|有罪},

purpose={entice|勾引}}

cf. (WordNet) make illegal payments to in exchange for favors or influence

Relations – the soul of HowNet

- Meaning is represented by relations
- Computation of meaning is based on relations

Axiomatic Relations & Role Shifting - 1

{buy|买} <----> {obtain|得到} [consequence];
agent OF {buy|买}=possessor OF {obtain|得到};
possession OF {buy|买}=possession OF {obtain|得到}.

{buy|买} <----> {obtain|得到} [consequence];
beneficiary OF {buy|买}=possessor OF {obtain|得到};
possession OF {buy|买}=possession OF {obtain|得到}.

{buy|买} <----> {obtain|得到} [consequence];
source OF {buy|买}=source OF {obtain|得到};
possession OF {buy|买}=possession OF {obtain|得到}.

Axiomatic Relations & Role Shifting - 2

{buy|买} [entailment] <----> {choose|选择};
agent OF {buy|买}=agent OF {choose|选择};
possession OF {buy|买}=content OF {choose|选择};
source OF {buy|买}=location OF {choose|选择}.

{buy|买} [entailment] <----> {pay|付};
agent OF {buy|买}=agent OF {pay|付};
cost OF {buy|买}=possession OF {pay|付};
source OF {buy|买}=taget OF {pay|付}.

Axiomatic Relations & Role Shifting - 3

{buy|买} (X) <----> {sell|卖} (Y) [mutual implication];

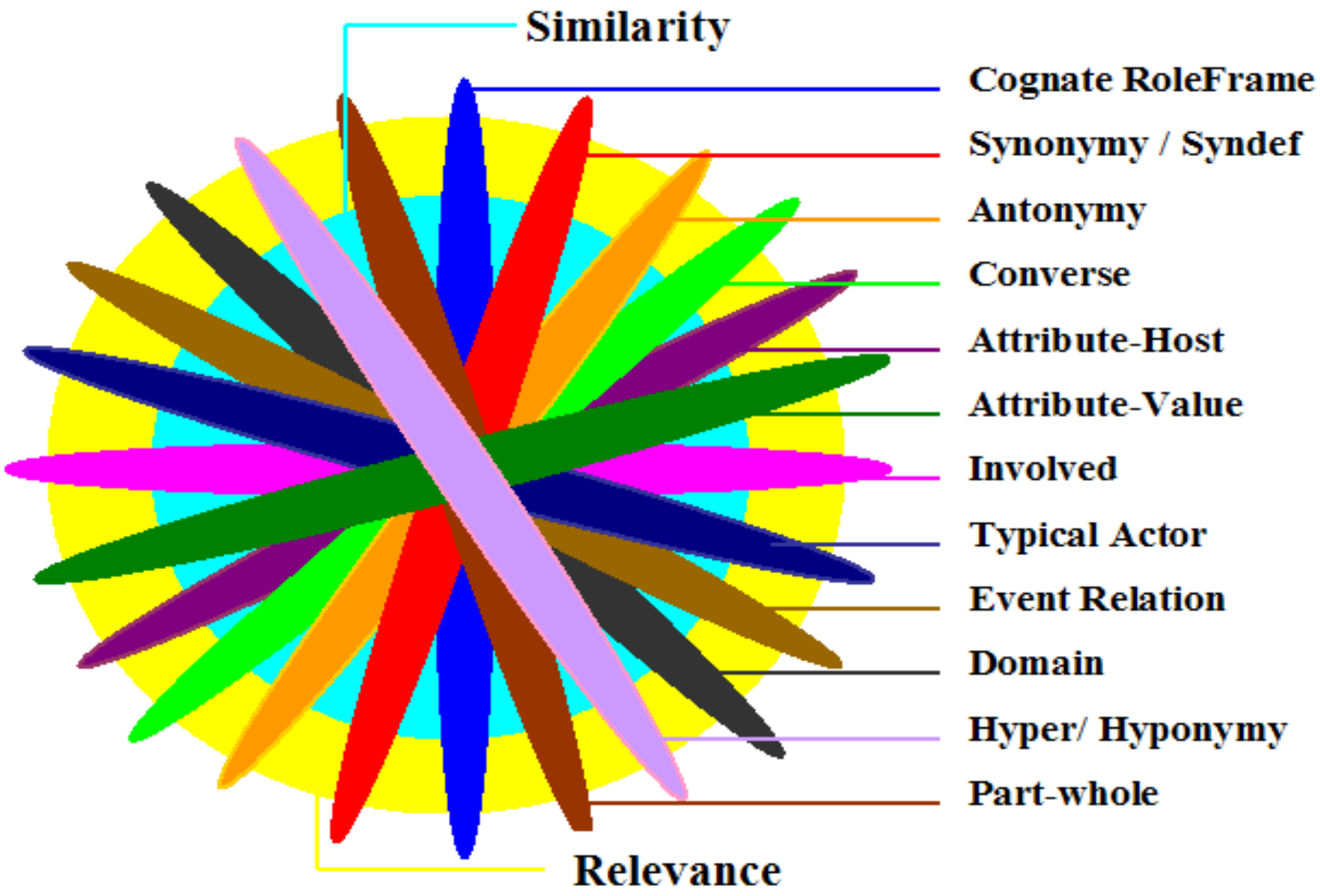
agent OF {buy|买}=target OF {sell|卖};

source OF {buy|买}=agent OF {sell|卖};

possession OF {buy|买}=possession OF {sell|卖};

cost OF {buy|买}=cost OF {sell|卖}.

Types of relations



Some observations on HowNet

- Rich and complex structure
- Relations are computed rather than explicitly stored
- Tries to capture both syntagmatic and paradignatic relationships

PrepNet

Aim of PrepNet

- Describing the syntax and the semantics of prepositions, in a way similar to
  - Verbs as in FrameNet (www.icsi.berkeley.edu/framenet), or VerbNet (www.cis.upenn.edu/verbnet) and
  - Nouns as in WordNet and EuroWordNet)

Criticality of Prepositions

- Prepositions very useful category in a number of applications such as *indexing, knowledge extraction, textual entailment and question answering*
- They convey basic meanings of much interest like *instruments, means, comparisons, amounts, approximations, localizations*, etc.

Examples of Prepositions' use

- Instrument: *eat **with** spoon*
- Localizaton: source: *came **from** Paris*; destination: *went **to** Grenoble*
- Temporality: duration: *works **for** 5 hours*; start: *worked **since** 3AM*; end: *finished **at** 10PM*
- Means: *succeeded **through** hard work*
- Comparison: *Taller **by** 5 inches*

All these preps are ambiguous!

- *With*: eat with spoon, sit with a friend, succeeded with hard work, hotel with a pool
- *At*: came at 9PM, stayed at Grenoble, stared at the face
- *Through*: succeeded through hard work, passed through Lyon, looked through the window, worked through the winter
- *To*: went to Paris, wanted to eat, stands to reason (idiomatic)

The PrepNet Project

- PrepNet (http://www.irit.fr/recherches/ILPL/Site-Equipe/ILPL.htm under revision)
- Framework that aims at constructing a repository of preposition syntactic and semantic behaviors
- Patrick Saint-Dizier: IJCNLP 2006

Structured at 2 levels

- **Abstract notion level**: global, language independent, characterization of preposition senses in an abstract way, where frames represent some generic semantic aspects of these notions
- **the language realization levels**: that deal with realizations for various languages, using a variety of marks (postpositions, affixes, compounds, etc.)

Postpositions: Indian Languages

- Come after nouns
- Hindi: *Eats with a spoon→ chammach (spoon) se (with) khaata_hai (eats)*
- Bengali: *Eats with a spoon→ chaamach (spoon) diye (with) khaay (eats)*
- Marathi: *Eats with a spoon→ chamchaani (with spoon) khaatoy (eats)*

Complex prepositions and postpositions

- English: *in spite of, in accordance to*
- Hindi: *ke saath (with), ke baavajud bhii (in spite of)*
- Multiword detection problem

Main features of PrepNet

- 196 preposition senses
- Lexical Conceptual Structure (LCS) framework (Dorr, 1995)
- 65 primitives
- Multilingual:
  - Themes and approximations (French, Spanish, Catalan, English, Thai) and
  - Instruments (German, Italian, Spanish, French, Arabic, Thai, Bahasa Malaysia, Hindi, Urdu, Kashmiri, Bengali, and Filipino)

Methodology of construction

- Studied bilingual dictionaries and corpora of English, French, Spanish and German
- Indian languages: Hindi, Urdu, Bengali
- Middle East Languages; Arabic
- Every preposition lexemes: constraints and distributions
- 1.5 man years
- Record structure

General architecture (1): categorizing preposition senses

→ Preposition categorization on 3 levels:

- **Family** (roughly thematic roles): localization, manner, quantity, etc.
- **Facets**: e.g., for localization: source, position, destination, etc.
- **Modalities**.

→ Facets viewed as abstract notions on which PrepNet is based

→ **12 families defined**

Families and Facets

Localization with facets: -source, -destination, -via/passage, -fixed position.

From an ontological point of view, all of theses facets can, a priori, apply to spatial, temporal or to more abstract arguments.

- **Quantity with facets**: - numerical or referential quantity, -frequency and iterativity, -proportion or ratio.

Abstract notions in PrepNet[1]

- **Localization with facets**: -source, -destination, -via/passage, -fixed position
  - *Went **from** Grenoble **to** Paris **via** Lyon*
- **Quantity with facets**: - numerical or referential quantity, -frequency and iterativity, -proportion or ratio
  - *Will come **in** 3 hours, divide **into** two parts*

Abstract notions in PrepNet[2]

- **Manner with facets:** -manners and attitudes, - means (instrument or abstract), -imitation, agreement or analogy.
  - Imitation: *he walks **like** a robot; agreement: he behaves **according to** the law*
- **Accompaniment with facets:** -adjunction, -simultaneity of events (co-events), -inclusion, - exclusion
  - *Adjunction : steak **with** French fries Exclusion: they all came **except** Paul.*

Abstract notions in PrepNet[3]

- **Choice and exchange with facets:** -exchange, -choice or alternative, -substitution
  - Substitution : *sign* ***for*** *the head in his absence*, Choice: ***among*** *all my friends, he is the funniest one*
- **Causality with facets:** *-cause, -goal or consequence, -intention, -purpose.*
  - Cause*: the rock fell* ***under*** *the action of frost*

Abstract notions in PrepNet[4]

- **Opposition with two ontological distinctions:** physical opposition and psychological or epistemic opposition
  - Opposition*: to act **contrary to** one's interests*
- **Ordering with facets:** -priority, -subordination, -hierarchy, -ranking, -degree of importance
  - Ranking : *at school, she is **ahead** of me*

Abstract notions in PrepNet[5]

- **Instrument:** complex case with interesting multilingual aspects (discussed later in separate slides)
- **Other groups:** -Theme, -in spite of, -comparison
  - Theme: *a book **on** dinosaurs*

Instrument as an abstract notion

- **Instrument:** -artifact**, -**means, -method
  - Artifact: *eats **with** spoon,* means: *went **by** horse carriage,* method: *solved **with** programming*

Instrument: Hindi[1]

- Default synset:
  - {*se, me, ke karaan, ke dwaraa, kar, karaan, dwara*}
- Syntactic frame is postposition based
  - case marks and the SOV form: *[X(subject, ergative), Y(object, accusative), Z(adjunct), postposition, Verb(action)]*

Instrument: Hindi[2]

- Interesting usage restrictions:
  - instrument type: concrete: *se*
  - instrument type: abstract: *dwaara*
  - instrument type: means of transportation: *me*
  - involvement of instrument: agentive: *ke dwaara*
  - Causal: *ke kaaran*
  - instrumental path: *me, se*

Instrument: French

- Default synset:
  - {*avec, par, au moyen de, grˆace `a, `a l'aide de, `a travers*}
- some syntactic structures:
  - *[X(subj) Verb(action) Y(obj) preposition Z(np or S)]*
  - *['utiliser' Z 'pour' Verb(action, infinitive) Y], etc.*

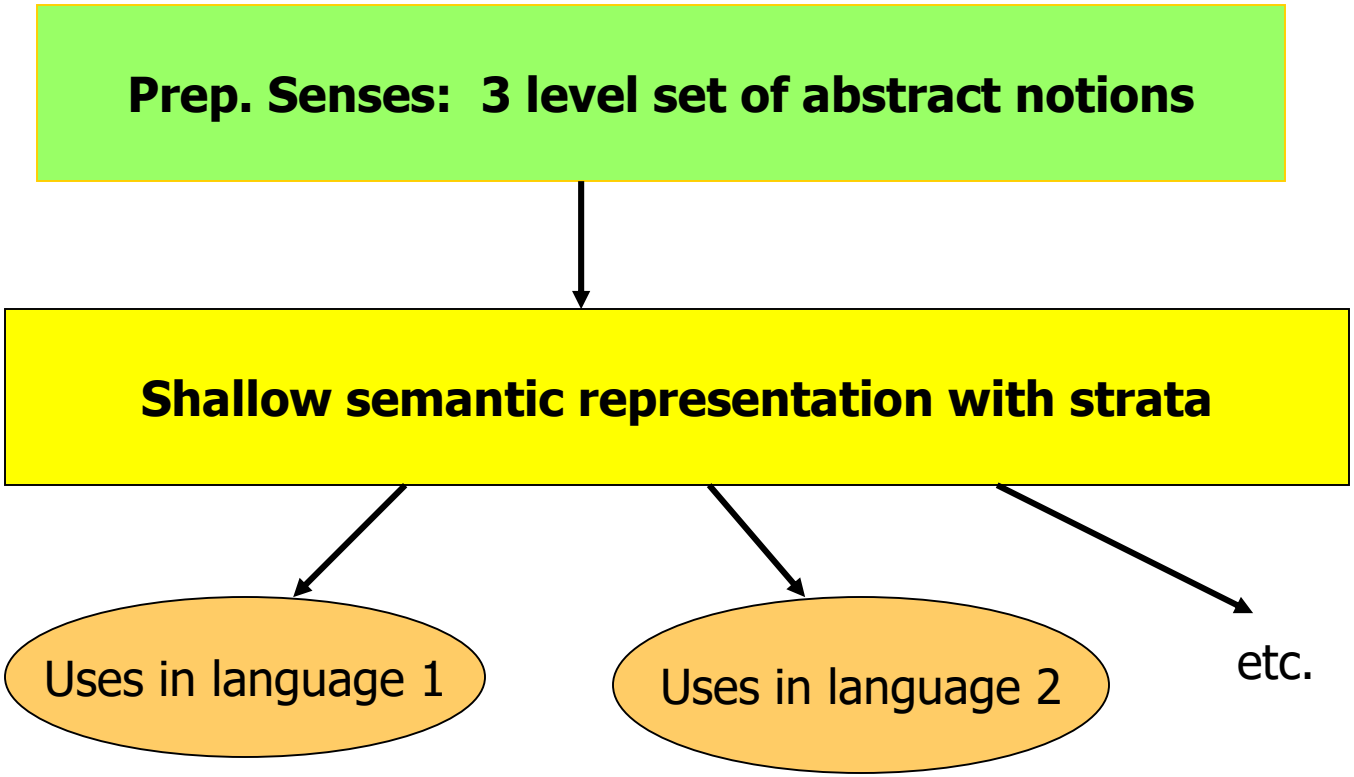
Instrument: German

- Default synset:
  - {*mit, mit Hilfe von, mittels, durch, anhand, kraft, dank, per*}
- Domain and pragmatic restrictions:
  - domain: judicial, psychological: *kraft, anhand*
  - formal usage: *mittels*
  - focus: *mittels, mit Hilfe von*
  - instrumental manner: *durch.*

Instrument: Malay

- Three ways to introduce instruments:
  - preposition + NP, affixes and compounding.
- Affixing:
  - Prefix*: beR- (e.g.* from *"kuda", "horse", "berkuda", on horseback)*
  - Prefix + Suffix: *meN- + -kan (e.g.* from *"paku", "nail", "memakukan", "to fasten with nails")*

Why is PrepNet a 'NET'



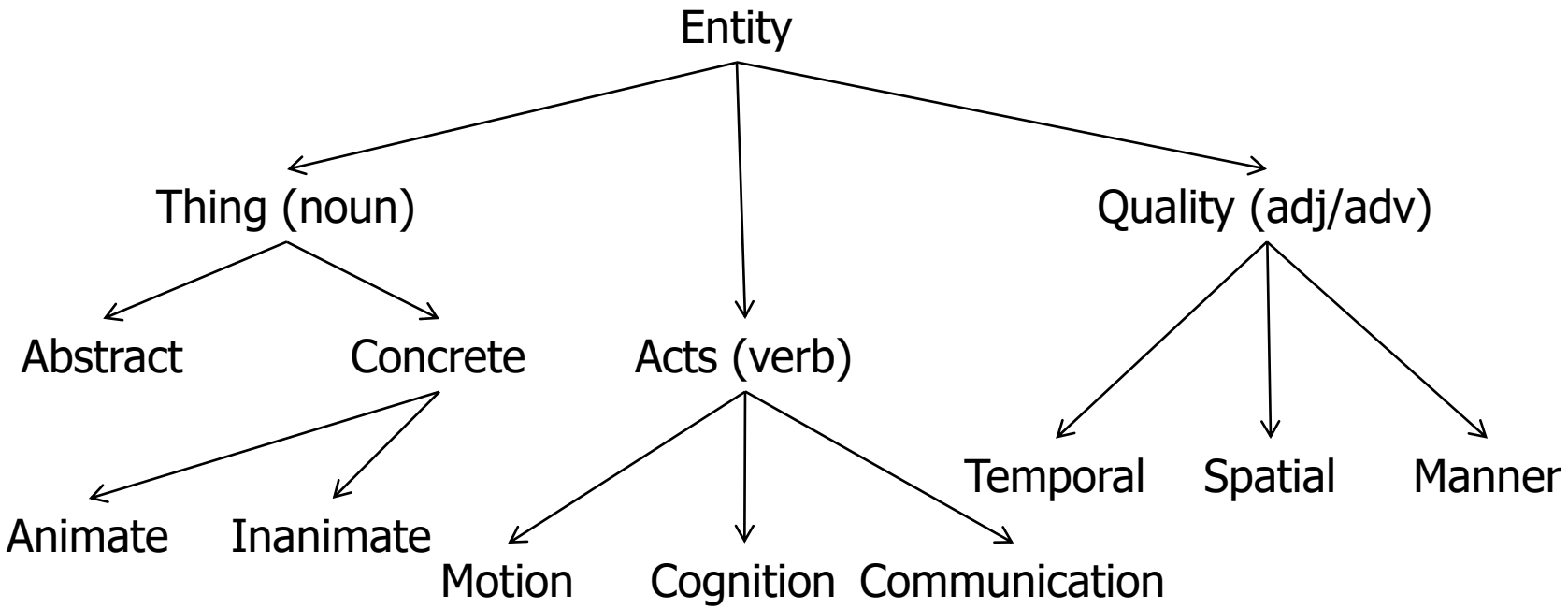
Important observations on PrepNet

- Explicit ontology (unlike WN): the abstract notions
- Nodes are like synsets: e.g. in French {*avec, par, au moyen de, grˆace `a, `a l'aide de, `a travers*}
- No explicit notion of semantic relations
- But a hierarchy in terms of *families* and *facets*

VerbNet and PropBank

Acknowledgement: Martha Palmer presentation in Columbia , 2004

Reminder: Fundamental ontology (starting part)



Based on Levin classes *(Levin, 1993)*

- Verb class hierarchy: 3100 verbs, 47 top level classes, 193

- Each class has a syntactic signature based on alternations.

  John broke the jar. / The jar broke. / Jars break easily.

  change-of-state

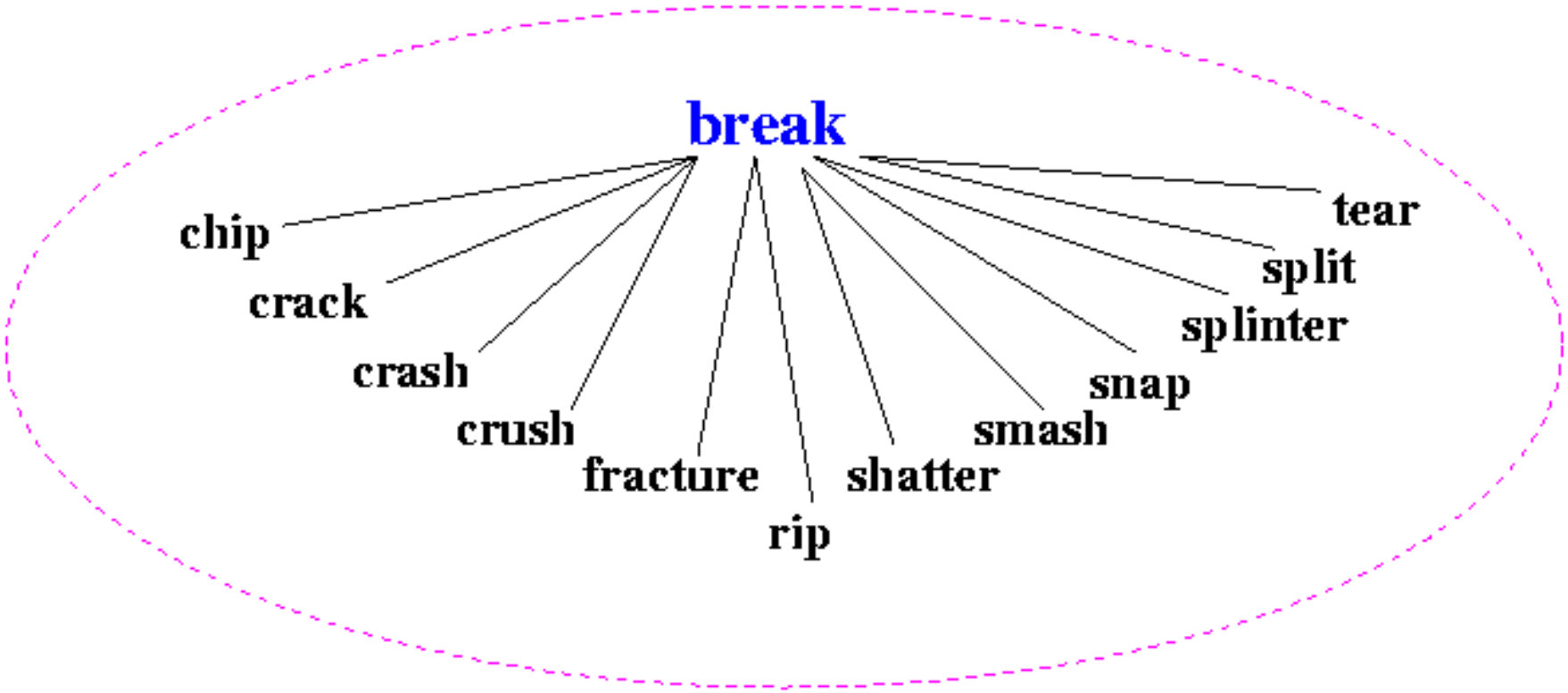
  *John cut the bread. / *The bread cut. / Bread cuts easily.*

  change-of-state, recognizable action, sharp instrument

  *John hit the wall. / *The wall hit. / *Walls hit easily.*

  contact, exertion of force

Break Levin class - *Change-of-state*



VerbNet

- Class entries:
  - Capture generalizations about verb behavior
  - Organized hierarchically
  - Members have common semantic elements, semantic roles and syntactic frames
    - John loves mangoes
    - John is fond of mangoes
- Verb entries:
  - Refer to a set of classes (different senses)
  - each class member linked to WN synset(s) (not all WN senses are covered)

VerbNet class example

-<VNCLASS ID="**obtain-13.5.2**"…>
-<MEMBERS>-<MEMBER name="**accept**" wn="**accept%2:40:00**" grouping="**accept.01**"/>
-<MEMBER name="**accrue**" wn="**accrue%2:30:00** grouping="**accrue.01…>**
…<MEMBERS>
-<THEMROLES>
-<THEMROLE type="**Agent**">
-<SELRESTRS logic="**or**">
<SELRESTR Value="**+**" type="**animate**" />
<SELRESTR Value="**+**" type="**organization**" />
</SELRESTRS>
</THEMROLE>
-<THEMROLE type="**Theme**">…</THEMROLE>
-<THEMROLE type="**Source**">… </THEMROLE>
</THEMROLE>
<FRAMES>-<FRAME>
<DESCRIPTION descriptionNumber="**0.2**" primary="**NP V NP**" secondary="**Basic Transitive**" xtag="**0.2**" />
-<EXAMPLES>
<EXAMPLE>**Carmen obtained the spare part.**</EXAMPLE>
</EXAMPLES>
-<SYNTAX>
-<NP value="**Agent**"> <SYNRESTRS /> </NP>
<VERB />
<NP value="**Theme**"> <SYNRESTRS /> </NP>

VerbNet class example

```
<SEMANTICS>
    -<PRED value="has_possession">
        -<ARGS>
         <ARG type="Event" value="start(E)" />
         <ARG type="ThemRole" value="?Source" />
         <ARG type="ThemRole" value="Theme" />
          </ARGS>
     </PRED>
    -<PRED value="transfer">
        -<ARGS>
         <ARG type="Event" value="during(E)" />
                    <ARG type="ThemRole" value="Theme" />
         </ARGS>
    </PRED>
    -<PRED value="has_possession">
        -<ARGS>
         <ARG type="Event" value="end(E)" />
         <ARG type="ThemRole" value="Agent" />
         <ARG type="ThemRole" value="Theme" />
        </ARGS>
     </PRED>
    -<PRED value="cause">
        -<ARGS>
         <ARG type="ThemRole" value="Agent" />
         <ARG type="Event" value="E" />
         </ARGS>
      </PRED>
</SEMANTICS>
```

VerbNet class example

```
                    </FRAME>
                    -<FRAME>
                    ….
                    </FRAME>
          </FRAMES>
          <SUBCLASSES>
                    -<VNSUBCLASS ID="obtain-13.5.2-1">
                      -<MEMBERS>
                        -<MEMBER name="acquire" wn="acquire%2:40:00"
    grouping="acquire.03" />
                          ….
                       </MEMBERS>
                    -<THEMROLES>
                                  …
                    </THEMROLES>
                    <FRAMES>
                              -<FRAME> ….</FRAME>
                    </FRAMES>
          <SUBCLASSES />
    </VNSUBCLASS>
    </SUBCLASSES>
</VNCLASS>
```

Semantic role labels:

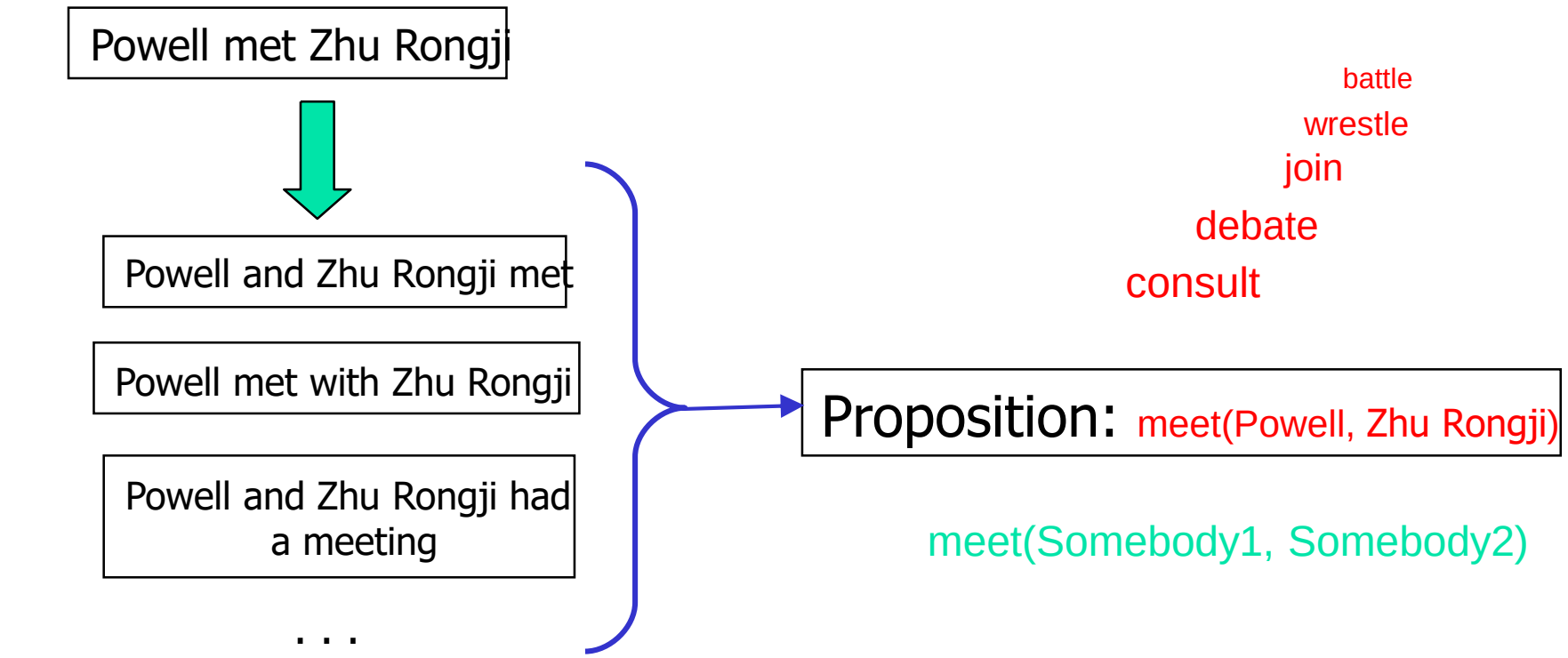
Julia broke the LCD projector.

break (agent(Julia), patient(LCD-projector))

cause(agent(Julia), broken(LCD-projector))

agent(A) -> intentional(A), sentient(A),
causer(A), affector(A)
patient(P) -> affected(P), change(P),…

Proposition Bank:
From Sentences to Propositions



When Powell met Zhu Rongji on Thursday they discussed the return of the spy plane.

meet(Powell, Zhu) discuss([Powell, Zhu], return(X, plane))

Capturing semantic roles*

SUBJ

- Owen broke [ARG1 the laser pointer.]

SUBJ

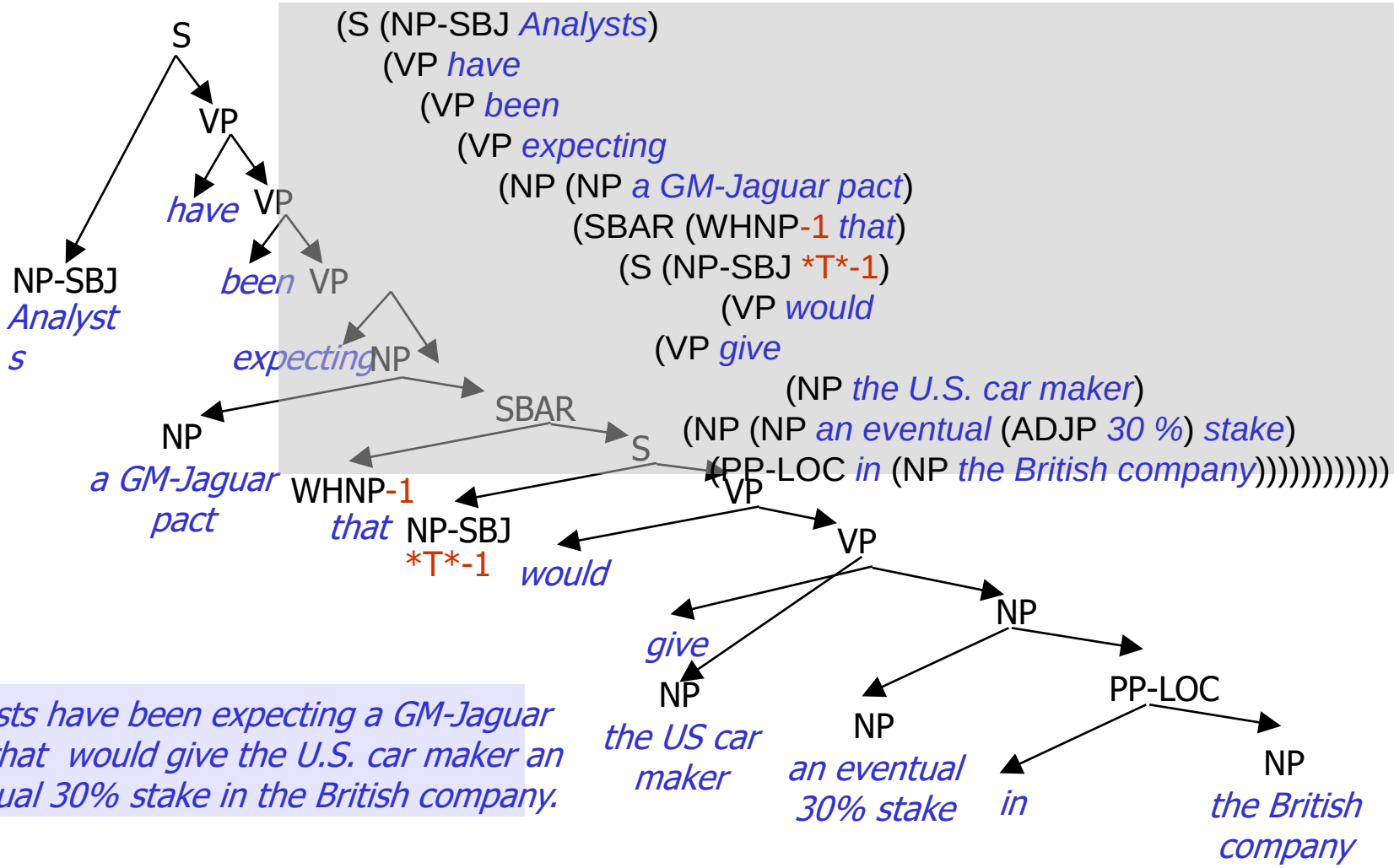
- [ARG1 The windows] were broken by the hurricane.

SUBJ

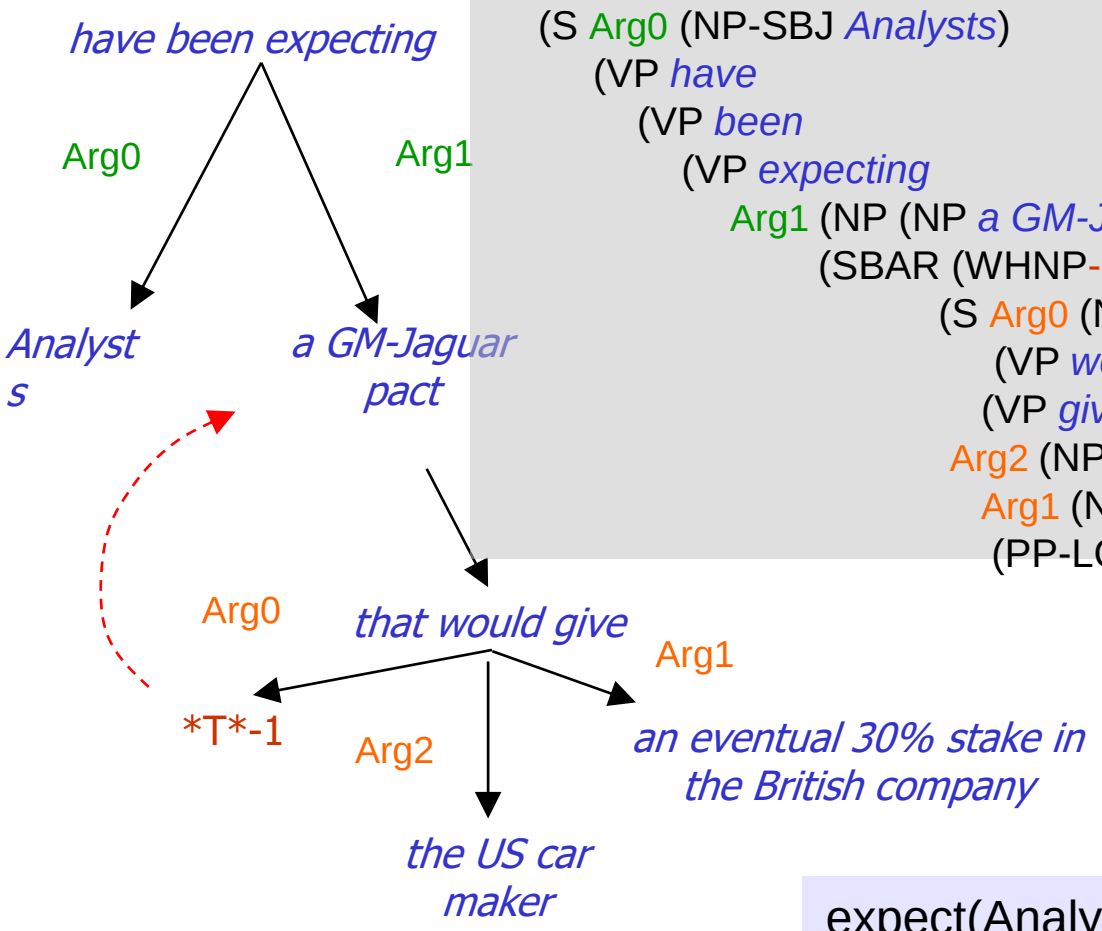
- [ARG1 The vase] broke into pieces when it toppled over.

*See also Framenet, http://www.icsi.berkeley.edu/~framenet/

A TreeBanked Sentence



The same sentence, PropBanked




```
(S Arg0 (NP-SBJ Analysts)
    (VP have
      (VP been
        (VP expecting
          Arg1 (NP (NP a GM-Jaguar pact)
                   (SBAR (WHNP-1 that)
                         (S Arg0 (NP-SBJ *T*-1)
                           (VP would
                             (VP give
                               Arg2 (NP the U.S. car maker)
                                 Arg1 (NP (NP an eventual (ADJP 30 %) stake)
                                  (PP-LOC in (NP the British company)))))))))))
```

expect(Analysts, GM-J pact)
give(GM-J pact, US car maker, 30% stake)

Frames File Example: *expect*

Roles:

Arg0: expecter

Arg1: thing expected

Example: Transitive, active:

Portfolio managers ***expect*** *further declines in interest rates.*

Arg0:	*Portfolio managers*
REL:	*expect*
Arg1:	*further declines in interest rates*

Frames File example: *give*

Roles:

Arg0: giver
Arg1: thing given
Arg2: entity given to

Example: double object

The executives gave the chefs a standing ovation.

Arg0:	*The executives*
REL:	*gave*
Arg2:	*the chefs*
Arg1:	*a standing ovation*

Word Senses in PropBank

- 700+ verbs
  - *Mary left the room*
  - *Mary left her daughter-in-law her pearls in her will*

Frameset **leave.01** "move away from":
Arg0: entity leaving
Arg1: place left

Frameset **leave.02** "give":
Arg0: giver
Arg1: thing given
Arg2: beneficiary

How do these relate to traditional word senses in VerbNet and WordNet?

Observations on VerbNet-Propbank

- Deeper hierarchy of verbs
- Semantic roles recorded
- Selectional preferences too along with sentence frames

Sentence frame in WordNet

- For the verb *Give:*
- Sense 2: yield, give, afford -- (be the cause or source of; "He gave me a lot of trouble"; "Our meeting afforded much interesting information")

  *> Something ----s something
- Sense 3: give -- (transfer possession of something concrete or abstract to somebody; "I gave her my money"; "can you give me lessons?"; "She gave the children lots of love and tender loving care")

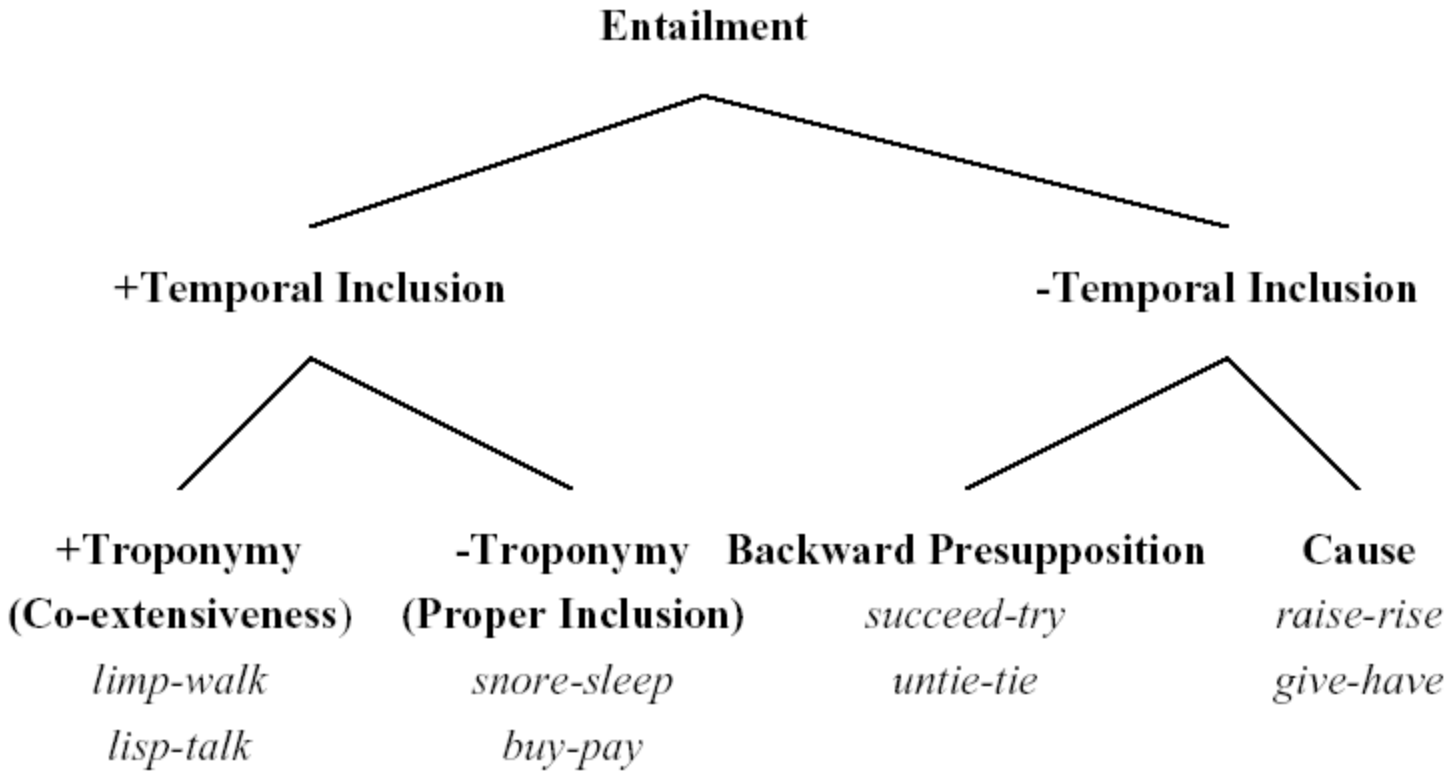
  *> Somebody ----s somebody something

  *> Somebody ----s something to somebody

VerbOcean

(Timothy Chklovski and Patrick Pantel , 2008)

Reminding: WordNet Verb Hierarchy



New verb relationships: proposed by *VerbOcean*

SEMANTIC RELATION	*EXAMPLE*	*Alignment with WordNet*	*Symmetric*
similarity	transform :: integrate	synonyms or siblings	Y
strength	wound :: kill	synonyms or siblings	N
antonymy	open :: close	antonymy	Y
enablement	fight :: win	cause	N
happens-before	buy :: sell; marry :: divorce	cause entailment, no temporal inclusion	N

Particulars

- VERBOCEAN is available for download at http://semantics.isi.edu/ocean/.
- The 1.5GB corpus consists of San Jose Mercury, Wall Street Journal and AP Newswire articles from the TREC-9 collection.
- Mined from the web using the DIRT (Discovery of Inference Rules from Text) algorithm

Comparisons - I

- Domain
- Principles of construction:
  - Constructive: HowNet
  - Differentiative: WordNet
  - Definition oriented: MindNet
  - Generative: FrameNet
  - Uncentralised: MindNet, ConceptNet
- Method of Construction:
  - Manual (expert): WordNet, HowNet, PrepNet
  - Crowdsourced: ConceptNet

Comparisons - II

- Representation:
  - Record-oriented: HowNet, PrepNet, FrameNet
  - Synsets: WordNet
  - Assertions: ConceptNet
- Coverage and Quality:
  - High: WordNet, HowNet, FrameNet
  - Medium: ConceptNet

Conclusions

- Lexical structures can be applied to NLP, IR tasks
- Increasing in coverage and utility
- Key issues: Method of construction, quality of database, coverage
- Key tradeoff: Quality vs. speed of collection