CS626: Speech, Natural Language Processing and the Web

Wordnet and Word Sense Disambiguation (continued)

Pushpak Bhattacharyya

Computer Science and Engineering Department

IIT Bombay

Week 4 of 16th August, 2021

Wordnet

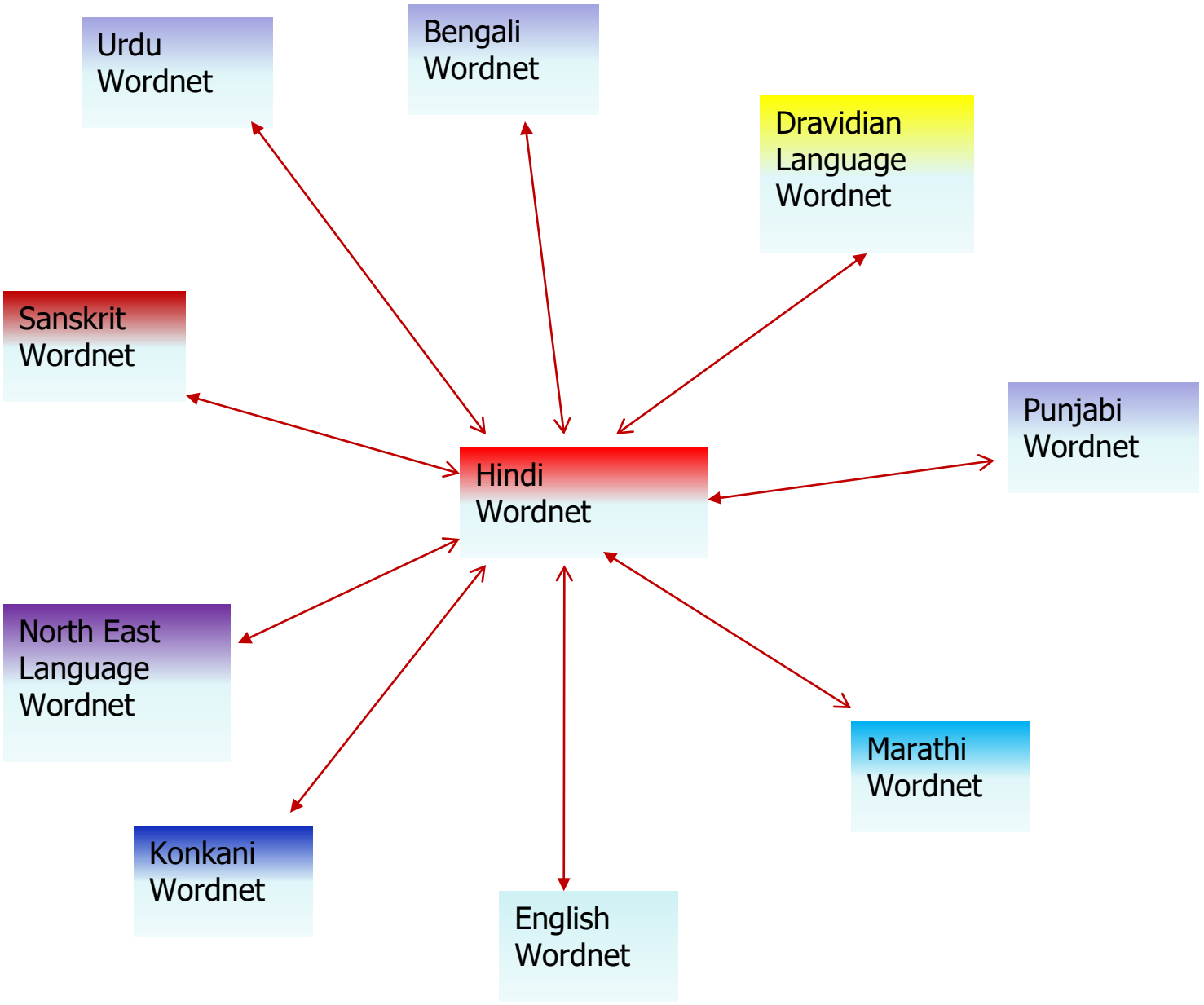
Wordnet: main purpose

- Disambiguation: **Sense Disambiguation**
- Main instrument: Relational Semantics
- Disambiguate a *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*

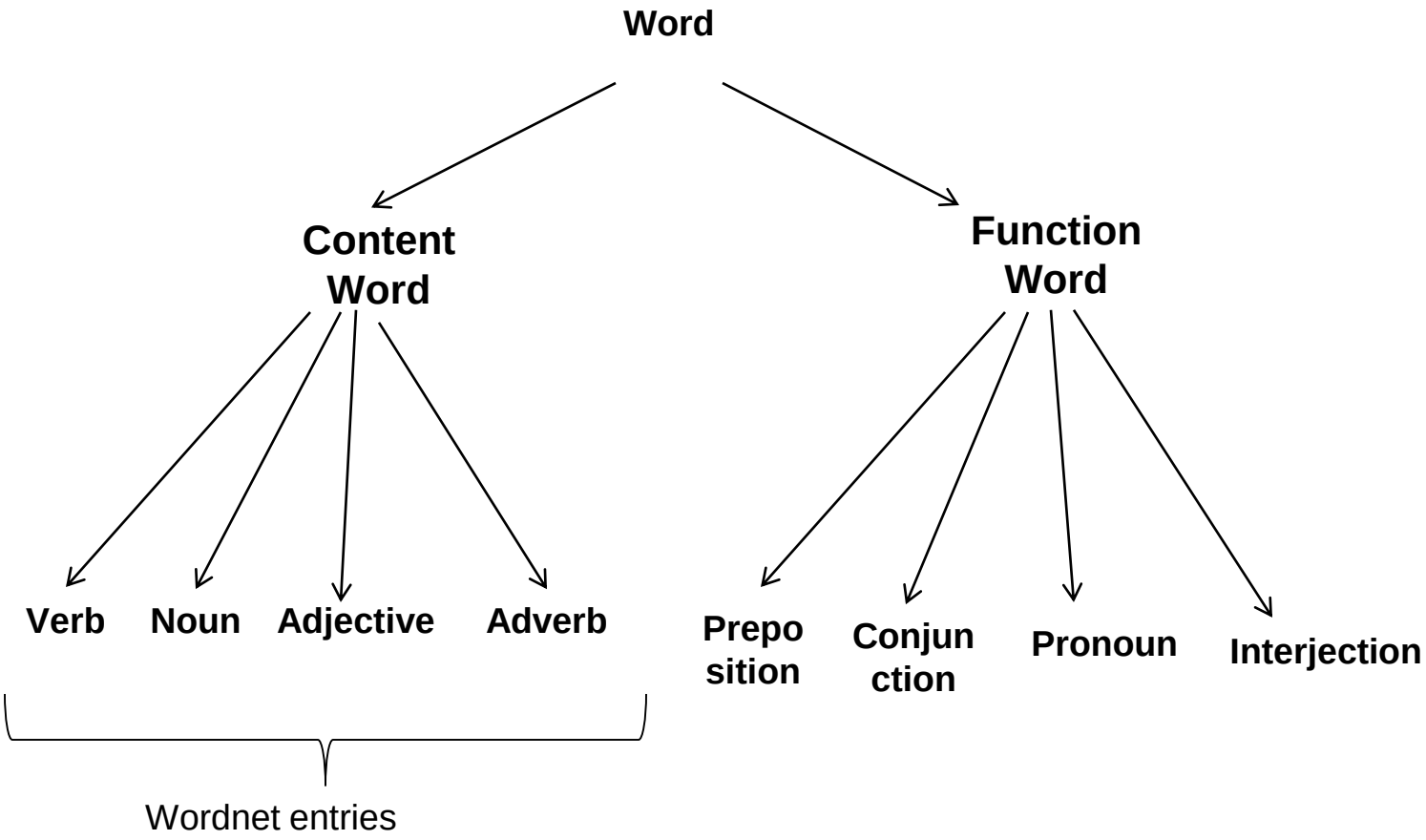
Wordnet - Lexical Matrix (with examples)

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					$E_{m,n}$

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")

Sense marked Data Statistics of Indowordnet

Corpus	No. of documents	Total no. of words	Total no. of words tagged	% of tagged words
Hindi Health	72	366230	74531	20.35%
Hindi Tourism	152	408439	178217	43.63%
Hindi BBC News paper	257	934482	382429	40.92%
Bengali	9	155035	50494	32.56%
Gujarati	101	337094	112884	33.49%
Kashmiri	350	1,00,000	42290	40%
Konkani	623	207828	107125	51.55%
Punjabi	174	322155	135122	41.94%
Oriya	159	197935	101758	51.38%.
Urdu	240	110000	50171	45.61%
Marathi	24	5608	2583	46%

Reasons for low percentage in Sense-tagged data

- Function Words
- Name Entity- names of persons, places, institutes, etc.
- English Transliterated words and their variations
- Spelling errors
- Abbreviations
- Numbers – 34567890
- Word not found in WN

Basic Principle

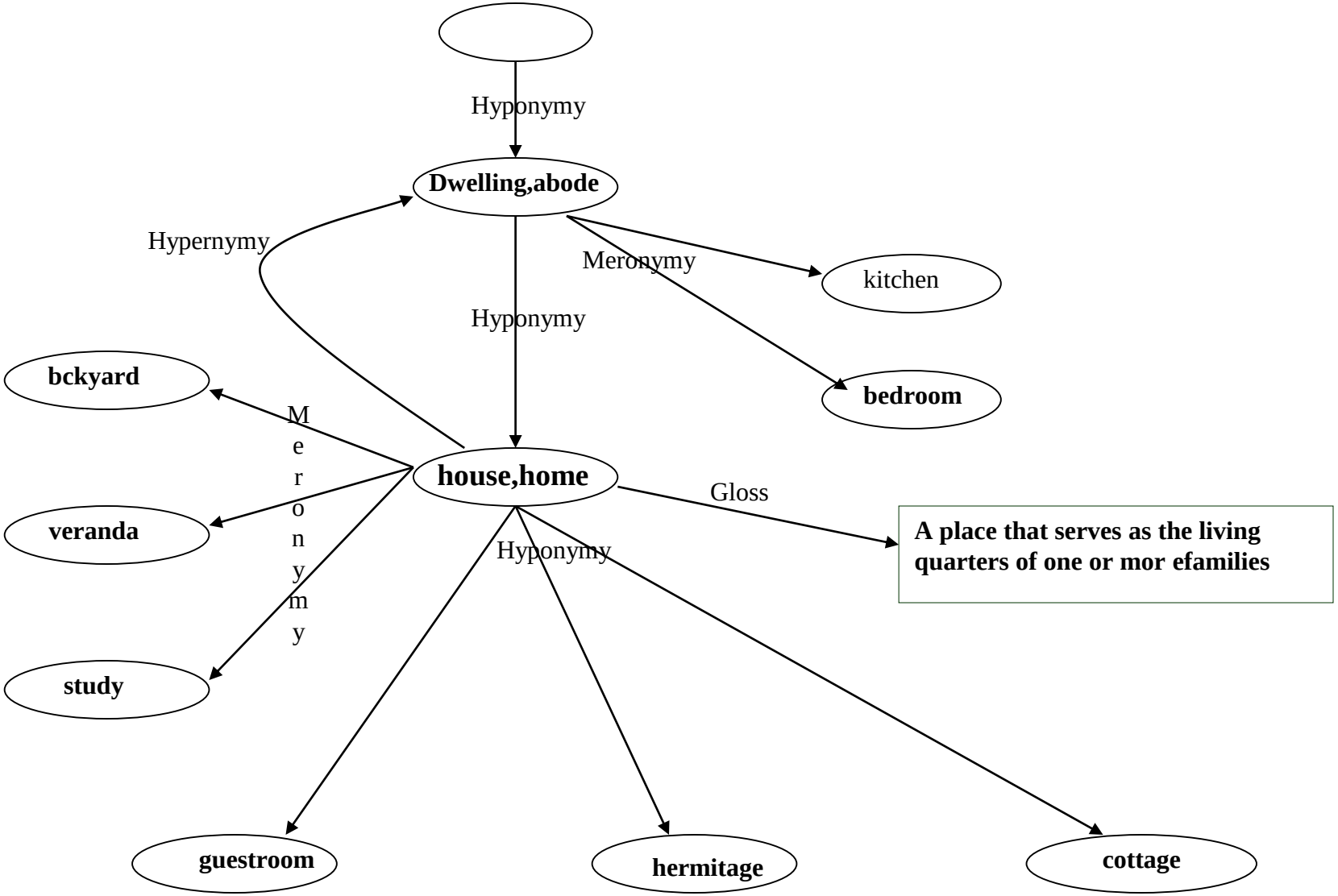
- Words in natural languages are polysemous.
- However, when synonymous words are put together, a unique meaning often emerges.
- Use is made of *Relational Semantics.*

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).*

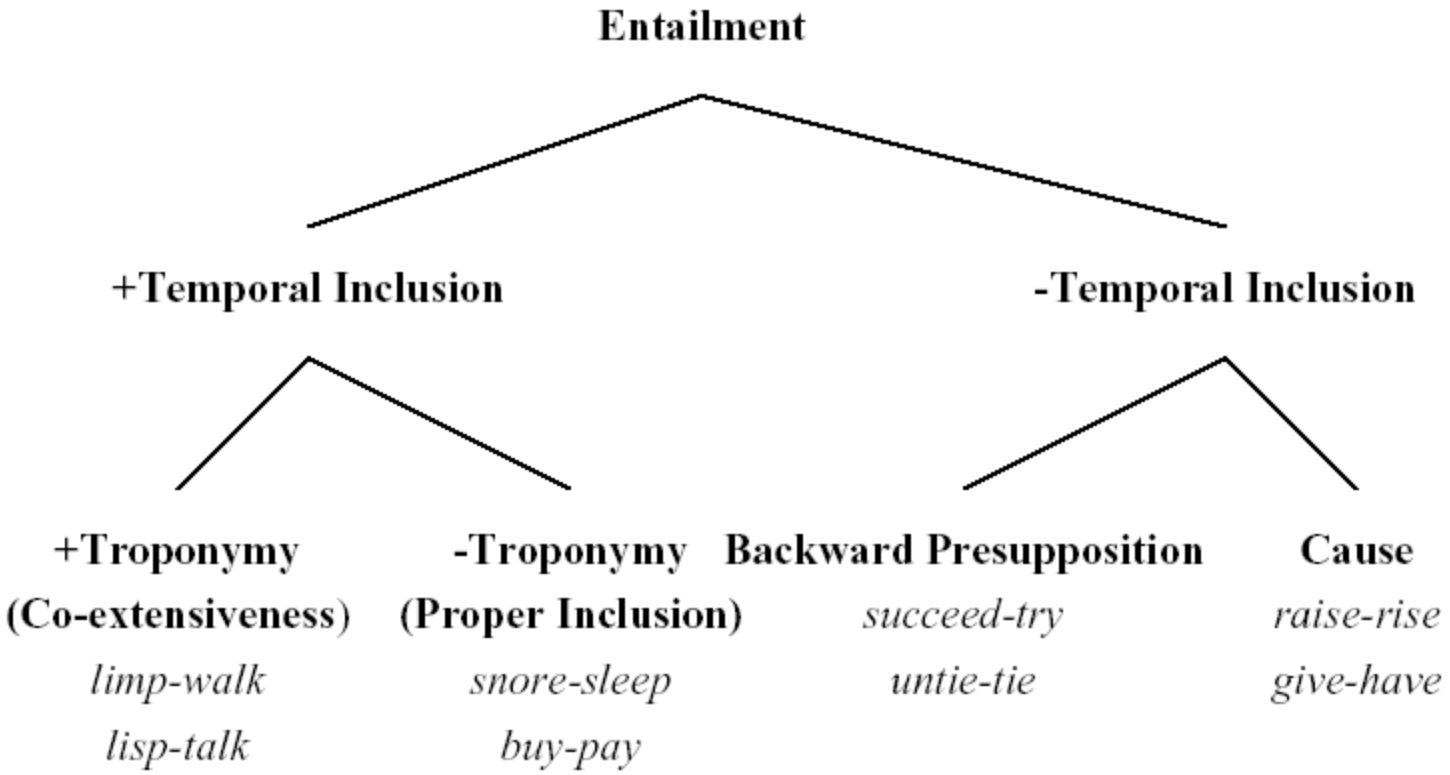
WordNet Sub-Graph



Vector representation of a synset

- Vector of a synset: < *Hypernymy id, Meronymy id, Hyponymy id, Representation for the gloss, Representation for example sentence, and so on* >

- Hypernymy id – Id of the synset which is linked by hypernymy to the given node


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

Principles behind creation of Synsets

Three principles:

- Minimality
- Coverage
- Replacability

Synset creation: from first principles

From first principles

- Pick all the senses from good standard dictionaries.
- Obtain synonyms for each sense.
- Needs hard and long hours of work.

Dictionary making→ Lexical Exploration

- Dictionary makers are like explorers (Columbus discovering America, e.g.)
- Field Linguists are like explorers
  - pick words: first level
  - Then for a *word* pick the senses: 2nd level

Synset creation: Expansion approach

From the wordnet of another language preferably in the **same family**

- Pick the synset and obtain the sense from the gloss.
- Get the words of the target language.
- Often same words can be used- especially for words with the same etymology borrowed from the parent language in the typology.
- Translation, Insertion and deletion.

Illustration of expansion approach with noun[1]

English

- bank (sloping land (especially the slope beside a body of water)) "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"

French (*wrong!*)

- ~~banque~~ (les terrains en pente (en particulier la pente à côté d'un plan d'eau)) "ils ont tiré le canot sur la rive», «il était assis sur le bord de la rivière et j'ai vu les courants"

Illustration of expansion approach with noun[2]

No hypernymy in the synset

English

- bank (sloping land (especially the slope beside a body of water)) "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the

French

- {rive, rivage, bord} (les terrains en pente (en particulier la pente à côté d'un plan d'eau)) "ils ont tiré le canot sur la rive», «il était assis sur le bord de la rivière et j'ai vu les courants"

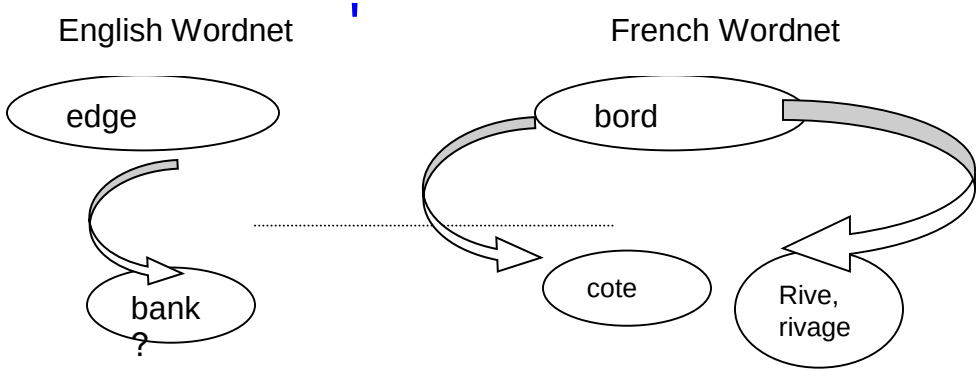


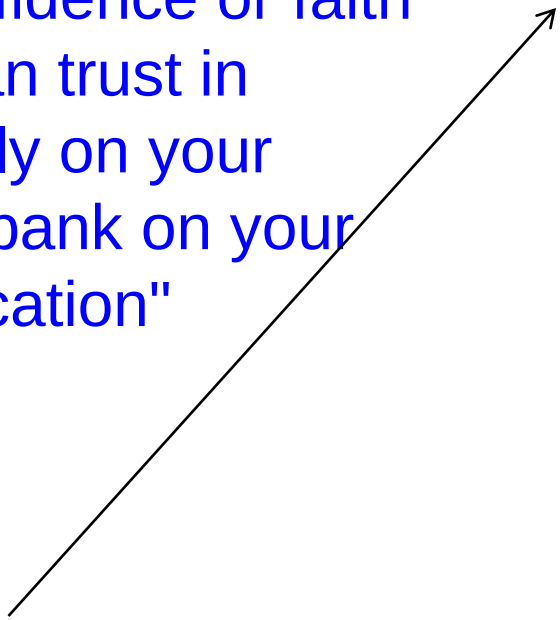
Illustration of expansion approach with verb[3]

English

- trust, swear, rely, bank (have confidence or faith in) "We can trust in God"; "Rely on your friends"; "bank on your good education"

French

- compter_sur, avoir_confiance_en, se_fier_a ’, faire_confiance_a’ (avoir confiance ou foi en) "Nous pouvons faire confiance en Dieu»,«Fiez-vous à vos amis",



Ordered by frequency

Lexical Relation

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

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**

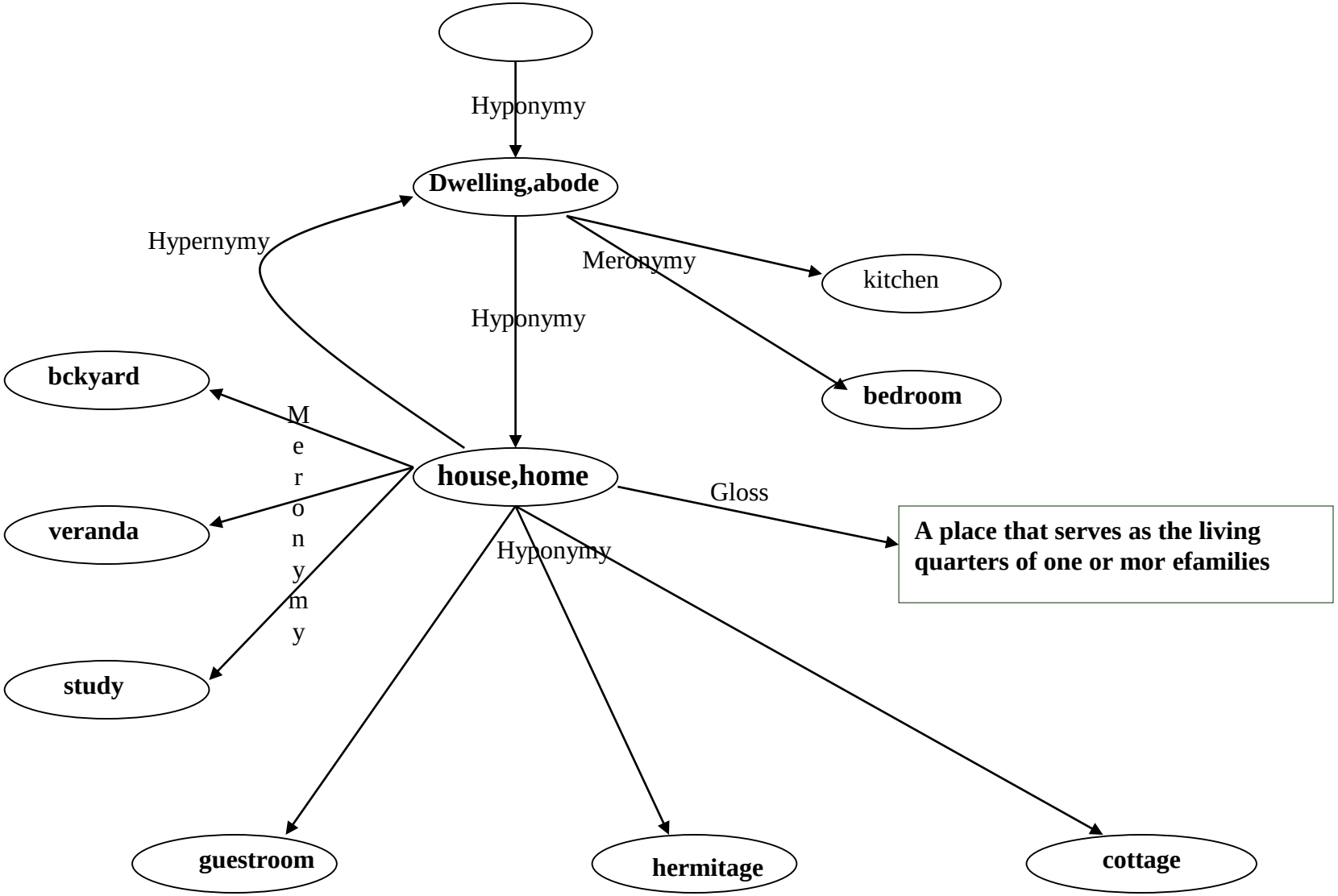
Kinds of Meronymy

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



Metonymy

- Associated with *Metaphors* which are epitomes of semantics
- Oxford Advanced Learners Dictionary definition: “The use of a word or phrase to mean something different from the literal meaning”
- *Does it mean Careless Usage?!*

Categories of Synsets (1/2)

•**Universal**: Synsets which have an indigenous lexeme in all the languages *(e.g. Sun ,Earth).*

•**Pan Indian**: Synsets which have indigenous lexeme in all the Indian languages but no English equivalent *(e.g. Paapad).*

•**In-Family:** Synsets which have indigenous lexeme in the particular language family (*e.g.* the term for *Bhatija* in Dravidian languages).

Categories of Synsets (2/2)

- **Language specific**: Synsets which are unique to a language (*e.g. Bihu* in Assamese language)

- **Rare**: Synsets which express technical terms (*e.g. ngram*).

- **Synthesized:** Synsets created in the language due to influence of another language (*e.g. Pizza*).

Need for categorization

- To bring systematicity in the way the wordnet synsets are linked
  - Universal→Pan Indian→Language Family→Language→Synthesised→Rare
- All members have finished the Universal and Pan Indian synsets

Categorization methodology

- 34378 Hindi synsets were sent to all Indo-wordnet groups in the tool, in which they had these options to categorize:
  - **Yes**
  - **No**

- **Universal synsets:-** The synsets which were categorized Yes and also have equivalent English words or synsets.
- **Pan-Indian :-** The synsets which were categorized Yes and did not have equivalent English words or synsets.

Well known Data: Brown Corpus

- 1,014,312 words of running text of edited English prose printed in the United States
- 500 samples of 2000+ words each
- Facilitate automatic or semi-automatic syntactic analysis

Tag repository and probability

- Where do tags come from?
  - Tag set
- How to get probability values i.e. P(.)?
  - Annotated corpora

After modeling of the problem, emphasis should be on the corpus

Computing P(.) values

Let us suppose annotated corpus has the following sentence

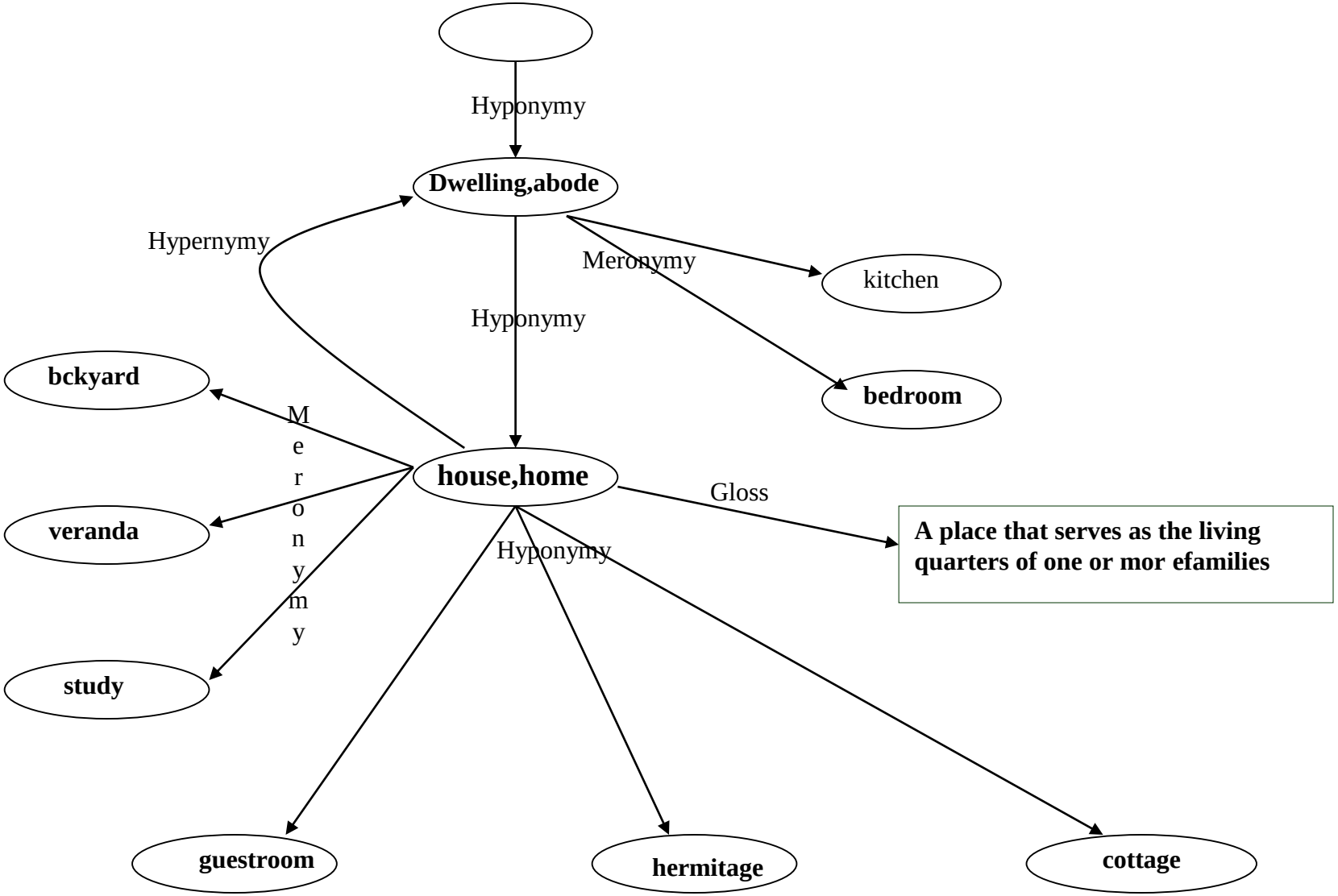
I	have	a	brown	bag	.
PRN	VB	DT	JJ	NN	.

$$P(NN \mid JJ) = \frac{Number\_of\_times\_JJ\_followed\_by\_NN}{Number\_of\_times\_JJ\_appeared}$$

$$P(Brown \mid JJ) = \frac{Number\_of\_times\_Brown\_tagged\_as\_JJ}{Number\_of\_times\_JJ\_appeared}$$

WSD

WordNet Sub-Graph



What is WSD (1/2)

- The task of Word Sense Disambiguation (WSD) consists of associating words in context with their most suitable entry in a pre-defined sense inventory.
- The de-facto sense inventory for English in WSD is WordNet. For example, given the word “mouse” and the following sentence:

What is WSD (2/2)

- For example, given the word “mouse” and the following sentence:
  - “A mouse consists of an object held in one’s hand, with one or more buttons.”
  - we would assign “mouse” with its electronic device sense (the 4th sense in the WordNet sense inventory).

Training Data for WSD

- The most widely used training corpus used is SemCor, with 226,036 sense annotations from 352 documents manually annotated.
- Some supervised methods, particularly neural architectures, usually employ the SemEval 2007 dataset.
- The most usual baseline is the Most Frequent Sense (MFS) heuristic, which selects for each target word the most frequent sense in the training data.

WSD: State of Art (1/2)

Supervised:

Model	Senseval 2	Senseval 3	SemEval 2007	SemEval 2013	SemEval 2015
MFS baseline	65.6	66.0	54.5	63.8	67.1
Bi-LSTM$_{att+LEX}$	72.0	69.4	63.7*	66.4	72.4
Bi-LSTM$_{att+LEX+POS}$	72.0	69.1	64.8*	66.9	71.5
context2vec	71.8	69.1	61.3	65.6	71.9
ELMo	71.6	69.6	62.2	66.2	71.3
GAS (Linear)	72.0	70.0	–*	66.7	71.6
GAS (Concatenation)	72.1	70.2	–*	67	71.8
GAS$_{ext}$ (Linear)	72.4	70.1	–*	67.1	72.1
GAS$_{ext}$ (Concatenation)	72.2	70.5	–*	67.2	72.6
supWSD	71.3	68.8	60.2	65.8	70.0
supWSD$_{emb}$	72.7	70.6	63.1	66.8	71.8
BERT (nearest neighbor)	73.8	71.6	63.3	69.2	74.4
BERT (linear projection)	75.5	73.6	68.1	71.1	76.2
GlossBERT	77.7	75.2	72.5	76.1	80.4
SemCor+WNGC, hypernyms	79.7	77.8	73.4	78.7	82.6
BEM	79.4	77.4	74.5	79.7	81.7
EWISER	78.9	78.4	71.0	78.9	79.3
EWISER+WNGC	80.8	79.0	75.2	80.7	81.8

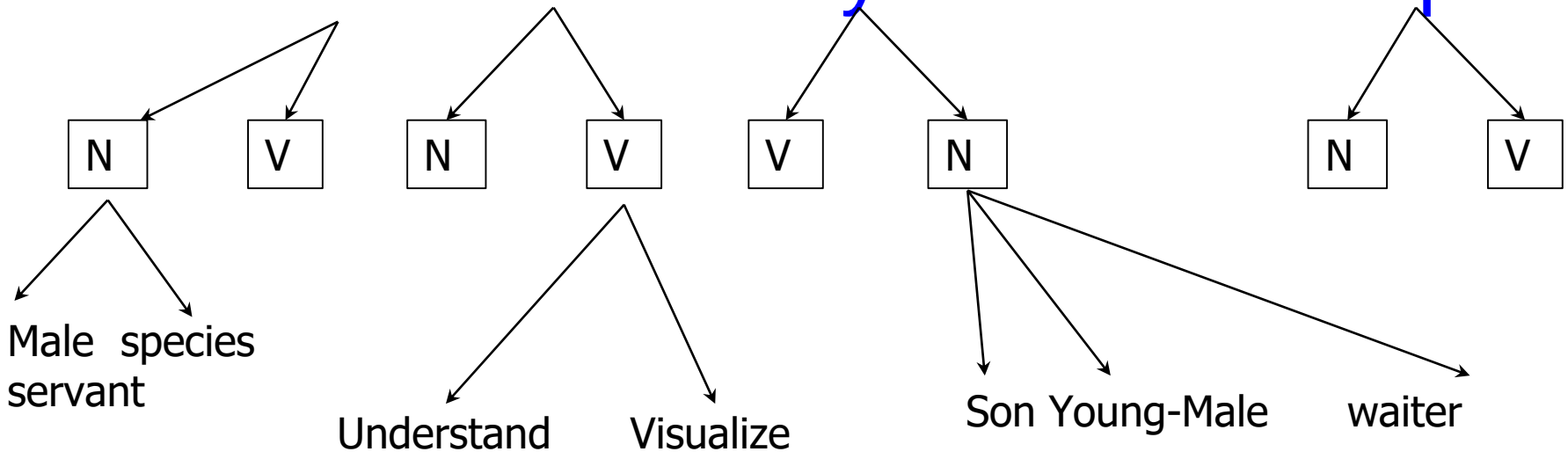
WSD: SOTA (2/2)

Knowledge-based:

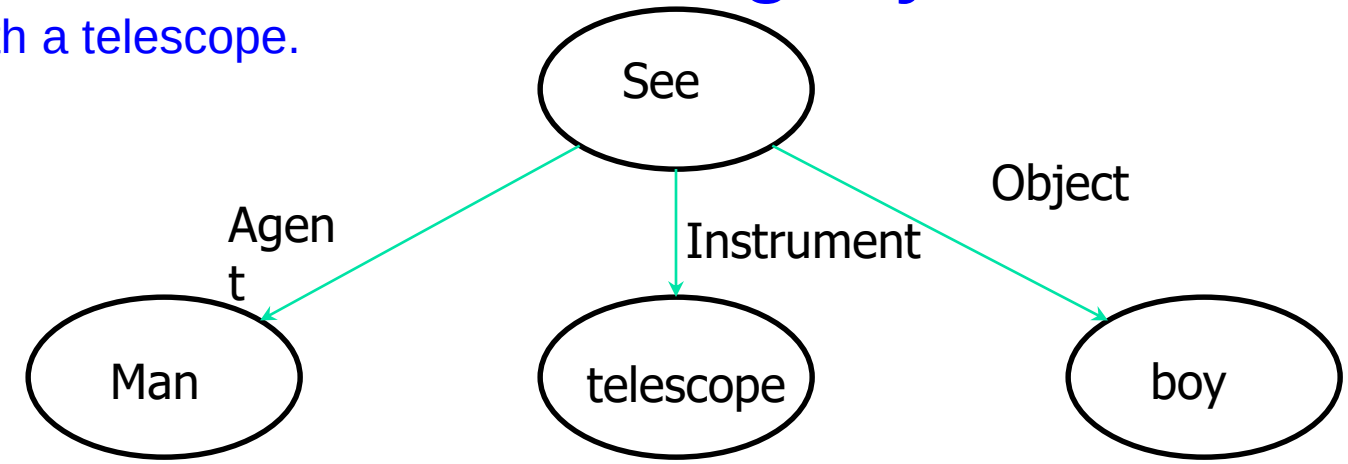
Model	All	Senseval 2	Senseval 3	SemEval 2007	SemEval 2013	SemEval 2015
WN 1st sense baseline	65.2	66.8	66.2	55.2	63.0	67.8
Babelfy	65.5	67.0	63.5	51.6	66.4	70.3
$UKB_{ppr\_w2w\text{-}nf}$	57.5	64.2	54.8	40.0	64.5	64.5
UKB_{ppr_w2w}	67.3	68.8	66.1	53.0	**68.8**	70.3
WSD-TM	66.9	69.0	**66.9**	55.6	65.3	69.6
KEF	**68.0**	**69.6**	66.1	**56.9**	68.4	**72.3**

WSD (multiple options for word meaning)

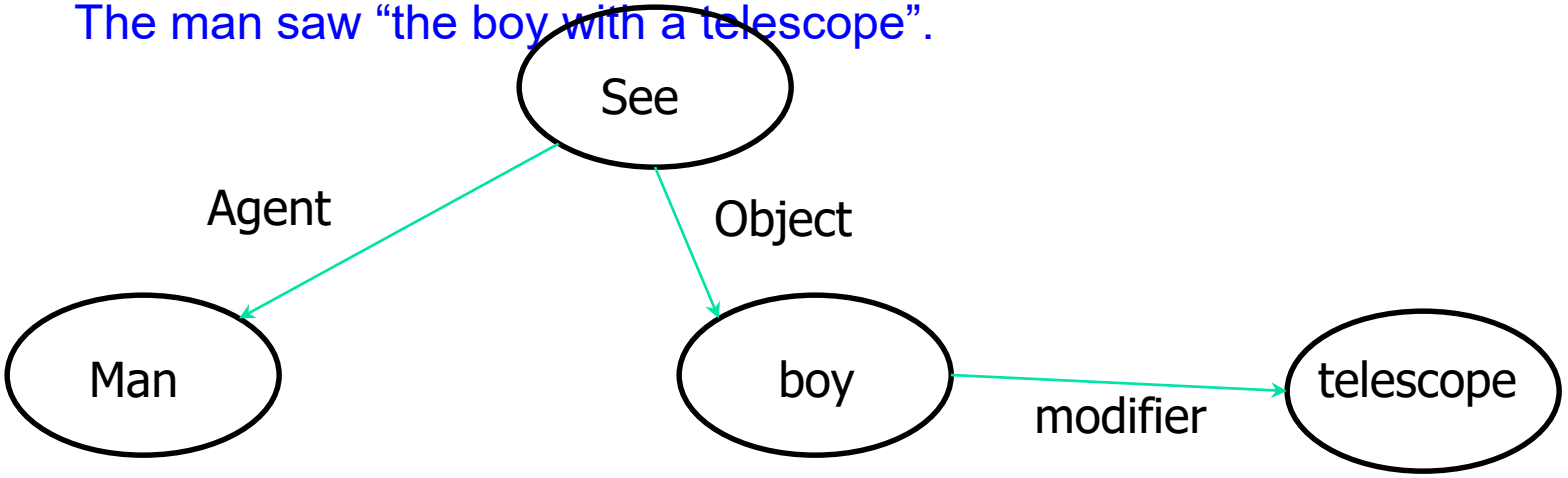
The man saw the boy with a telescope.



Semantic Role Ambiguity: The man saw “the boy” with a telescope.



The man saw “the boy with a telescope”.



Ambiguity

- Referential ambiguity
  - The dog chased a cat and it bit it.
- Preposition ambiguity
  - Ram ate some rice **with** vegetables.
  - Ram ate some rice **with** soon.
  - Ram ate some rice **with** Shayam.

Origin of a polysemous word

- How a word gets multiple sense in a language?
  - Metaphoric use of a word (‘fruit fell’ vs. ‘kingdom fell’)
  - Language contact (‘bank’ from German ‘Banque’)

Modeling of WSD- sense *S* given word *W* and context *C*

$$S^* = \arg\max_{S} P(S \mid w, C) \qquad w \in C$$

$$P(S \mid w, C) = \frac{\#(w\_tagged\_as\_S\_in\_context\ C)}{\#(w\_in\_context\ C)}$$

Isolate "*prior*" probability

$$P(S \mid w, C)$$

$$= \frac{P(S, w, C)}{P(w, C)}$$

$$= \frac{P(w)P(S, C \mid w)}{P(w)P(C \mid w)}$$

$$= \frac{P(S, C \mid w)}{P(C \mid w)}$$

$$= \frac{P(S \mid w)P(C \mid S, w)}{P(C \mid w)}$$

Constant in *argmax* calculation

$$S^* = \arg\max_S(P(S \mid w, C)) = \arg\max_S(P(S \mid w)P(C \mid S, w))$$

Prior

$$P(S \mid w) = \frac{\#(w\_tagged\_as\_S)}{\#w}$$

Likelihood

Let $W^S = W$ in sense S

Apply chain rule and make Markov Bi-gram assumption

$$P(C \mid w^S) = \prod_{i=1}^{K} P(c_i \mid w^S)$$

K=#words in context C, leaving out w

Example: modelling of WSD (1/3)

- Sentence - *He has Jupiter in the seventh house of his horoscope,* w: house, C: All words other than house
  - (*He, has, Jupiter, in, the, seventh, of, his, horoscope*)
- Word house has 3 senses (Astrological, Family, Dwelling)
- $S^* = \arg\max_S P(S|w, C)$, where $w \in C$

$$= \arg\max_S P(S|w, c) = P(S|w) * P(C|S, w)$$

Example: Modelling of WSD (2/3)

- Let S = Sense expressed by the synset id for particular sense(ex: Astrological)
- Prior : $P(S|w) =$ $\frac{\text{number of times word house tagged in astrological sense}}{\text{number of times house appears in corpus}}$
- Likelihood :

$P(C|S,w)$ = P(He, has, Jupiter, in, the, seventh, of, his, horoscope | word house in astrological sense)

Example: Modelling of WSD (3/3)

- W^s = Word w in sense S (here S = Astrological)
- Apply chain rule
  - P(he | W^s) * P(has | he,W^s)…..P(horoscope | He,has,Jupiter,in,the,seventh,of,his, W^s)
- Make Markov assumption (Bi-gram)
  - P(he| W^s)*P(has|He, W^s)……P(horoscope|his, W^s)

Two sentences

- “The navy performed a successful operation, the doctor said” (A)
- “The doctor performed a successful operation, the navy officer said” (B)
- Without a scoring mechanism, the system will not know what sense to output for operation, *medical/military/mathematics/cse*
- The scoring mechanism is probability
- Probability makes use of proportion of association counts (how many times doctor/navy is associated with operation)
- Even with scoring mechanism, system will go wrong in one of the examples; *A correct* *B wrong* and *vice versa*

Long-distance dependency

- I went to the $bank^{Finance}$ to draw money.
- I went to the $bank^{River}$ to draw water.

$$P(I \mid bank^{Finance}).P(went \mid I, bank^{Finance}) \square\ P(money \mid draw, bank^{Finance})$$

$$P(I \mid bank^{River}).P(went \mid I, bank^{River}) \square\ P(water \mid draw, bank^{River})$$

$$P(I, bank^{Finance}) = \frac{\#< I, bank^{Finance} >}{\#< bank^{Finance} >} \qquad P(went \mid I, bank^{Finance}) = \frac{\#< I, went, bank^{Finance} >}{\#< T, bank^{Finance} >}$$

$$P(I, bank^{River}) = \frac{\#< I, bank^{River} >}{\#< bank^{River} >} \qquad P(went \mid I, bank^{River}) = \frac{\#< I, went, bank^{River} >}{\#< T, bank^{River} >}$$

Annotation

- We need sense-tagged corpus
- WordNET [https://wordnet.princeton.edu/]
  - Bank
    - 10 Noun senses
      - Bank_883: Financial institution
      - Bank_99: Sloppy land
      - …
    - 8 Verb senses

Rule based approach

- If “water” appears in the context of “bank” then
  - “bank” most likely has the sense “River bank”
- If “money” appears in the context of “bank” then
  - “bank” most likely has the sense “Financial bank”

Decision List

- Decision list are clusters of rules

```
ContextOf(Bank)
     case "water" : □  River bank
     case "sand"  : □  River bank
     case "money" : □  Financial
   bank
```

. . .

How many such rules ?

Assume English has 100K words and on an average there are 3 senses per word.

#Rules > 300K

Robustness of rule-based systems

- Decision lists to be “complete”- Is that enough for WSD?

- No!
  - Need MA- If morphological form of “water” appears e.g. “watering”
  - Need POS tagging- POS tag: E.g. “He waters plants in front of a bank every week.”
  - Unrelated reference- “I went to bank to draw money and heard someone shouting ‘water’.”

Machine Learning for WSD

- HMM– Captures local context only (not adequate for WSD)
- Condition Random Field (CRF)– Captures long distance dependency
- Neural Network- Captures long distance dependency

Feature Abstraction in Deep learning

What is the intuition behind deep learning?

Deep Learning **discovers** “water” as important features for disambiguation, instead of being ***given*** that feature

1. Closeness between “water” and “river” in word embeddings
2. Using clustering formation:
   - Some neurons will produce a cluster whose members share a common feature which represents ‘water’
   - However, the neurons will not be able to put a name to the feature

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।

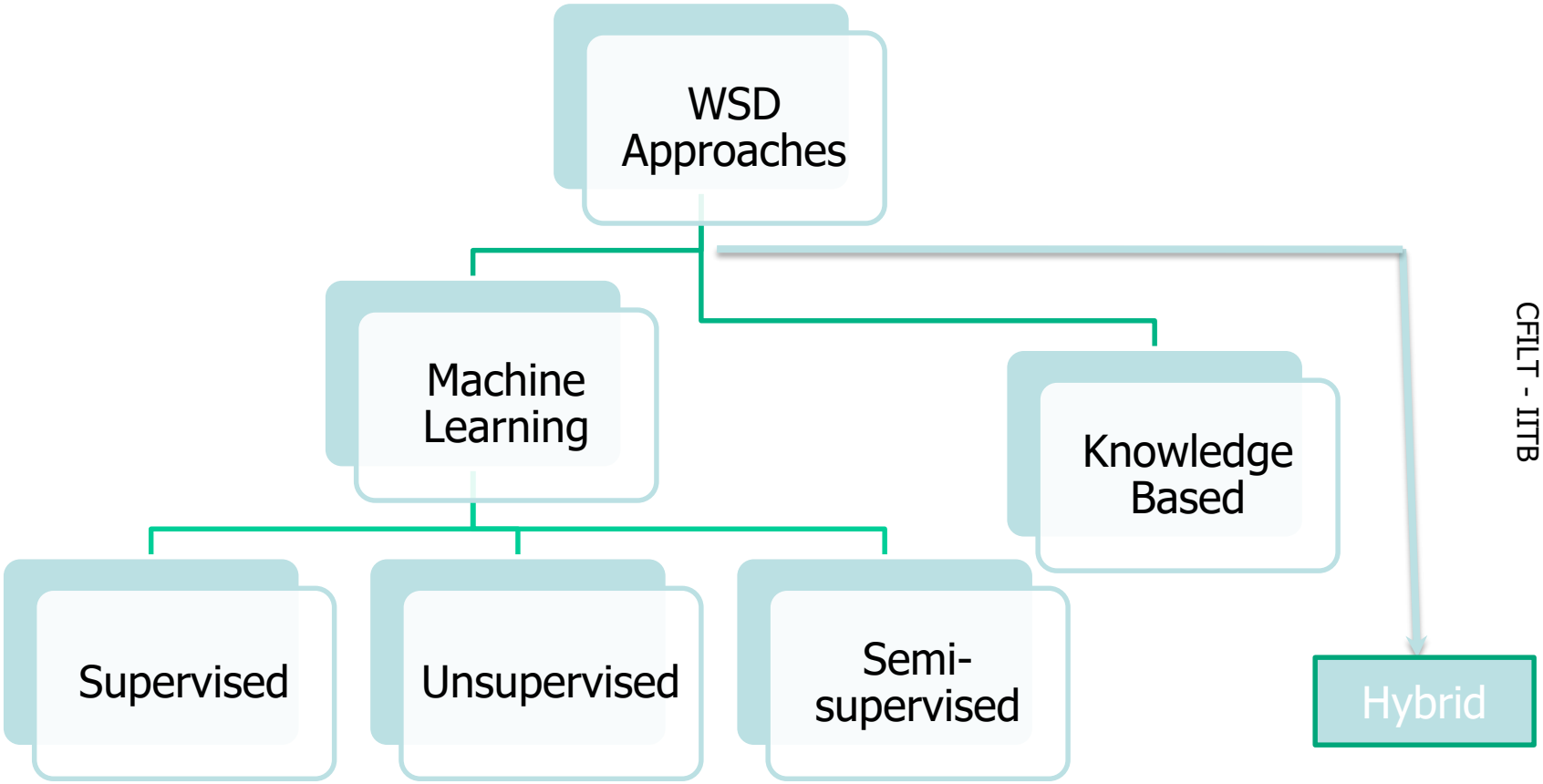
(According to a new research, those people who have a busy social life, have larger space in a part of their brain).

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



What is expected of *Course Project*

Activities in the paper

- You will take a problem
- Implement ONE paper (may be allied papers will support)
- Prepare the demo well
- DO result presentation and error analysis

Example

- Say, POS tagging for Malayalam
- Understand complexities
- Take a paper that gives the best results for Malayalam POS tagger
- Implement and see results