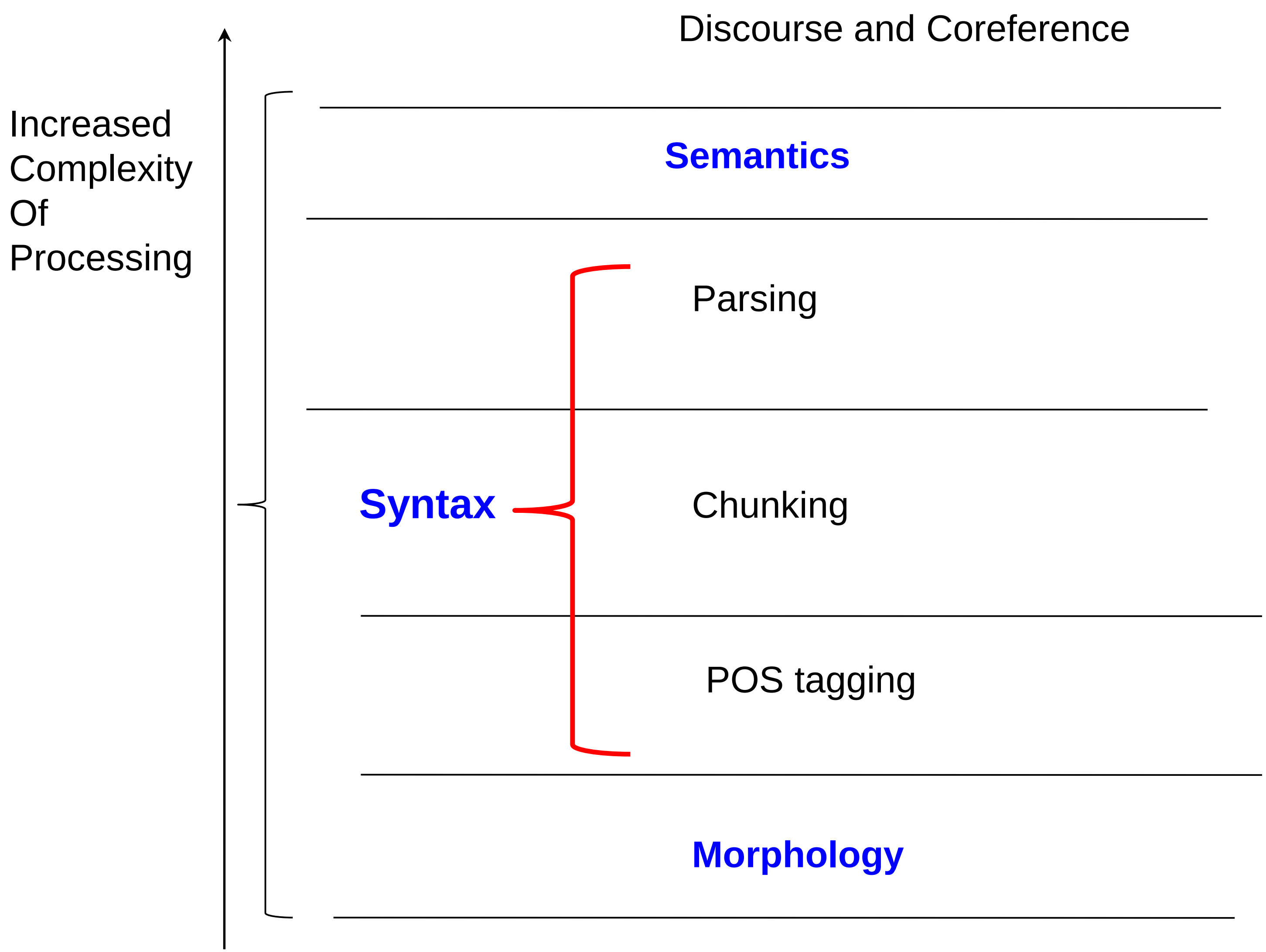
CS626: Speech, Natural Language Processing and the Web

Semantics, WN, FFNN, BP

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

Week 10 of 26th September, 2022

NLP Layer and Linguistics



Sound-Structure-Meaning continuum

Sound:
Phonetics, Phonology

Structure:
Morphology, Syntax

Meaning:
Semantic, Pragmatics

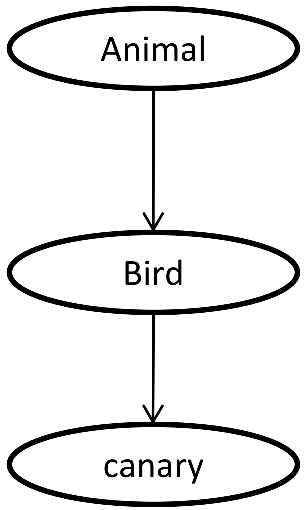
Syntagmatic and Paradigmatic Relations

- Syntagmatic and paradigmatic relations
  - Lexico-semantic relations: synonymy, antonymy, hypernymy, mernymy, troponymy etc. **CAT is-a ANIMAL**
  - Co-ccurence: **CATS MEW**
- Resources to capture semantics:
  - Wordnet: primarily paradigmatic relations
  - ConceptNet: primarily Syntagmatic Relations
- Interesting observation: for English, whenever a word is uttered, automatically words are pulled by association of which ~50% are syntagmatic and ~50% paradigmatic

Representing Word Meaning: ***Wordnet***

Psycholinguistic Evidence

- Human lexical memory for nouns as a hierarchy.
  - *Can canary sing? - Pretty fast response.*
  - *Can canary fly? - Slower response.*
  - *Does canary have skin? – Slowest response.*



(can move, has skin)

(can fly)

(can sing)

Wordnet- a lexical reference system based on psycholinguistic theories of human lexical memory.

Fundamental Device- 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*) $E_{2,\dots}$	
M_3		*(bank)* $E_{3,2}$	$E_{3,3}$		
…				…	
M_m					$E_{m,n}$

Wordnet: History

- **Princeton Wordnet** for English developed over 15 years. Released 1992.

- **Eurowordnet**- linked structure of European language wordnets built in 1998 over 3 years.

- **IndoWordnet** completed in 2010; effort of 10 years.

Basic Principle

- Words in natural languages are polysemous- meaning has many (‘poly’) meanings (‘sems’)

- However, when synonymous words are put together, a unique meaning often emerges.

- Use is made of *Relational Semantics.*
- Competing scheme: *Componential Semantics*, where a word is represented by features, e.g.,
  - Features: <Large?, Domesticable?, carnivorous?, furry?>
  - Tiger: <1, 0, 1, 1>, Cat: <0, 1, 1, 1>, Cow: <1, 1, 0, 0>

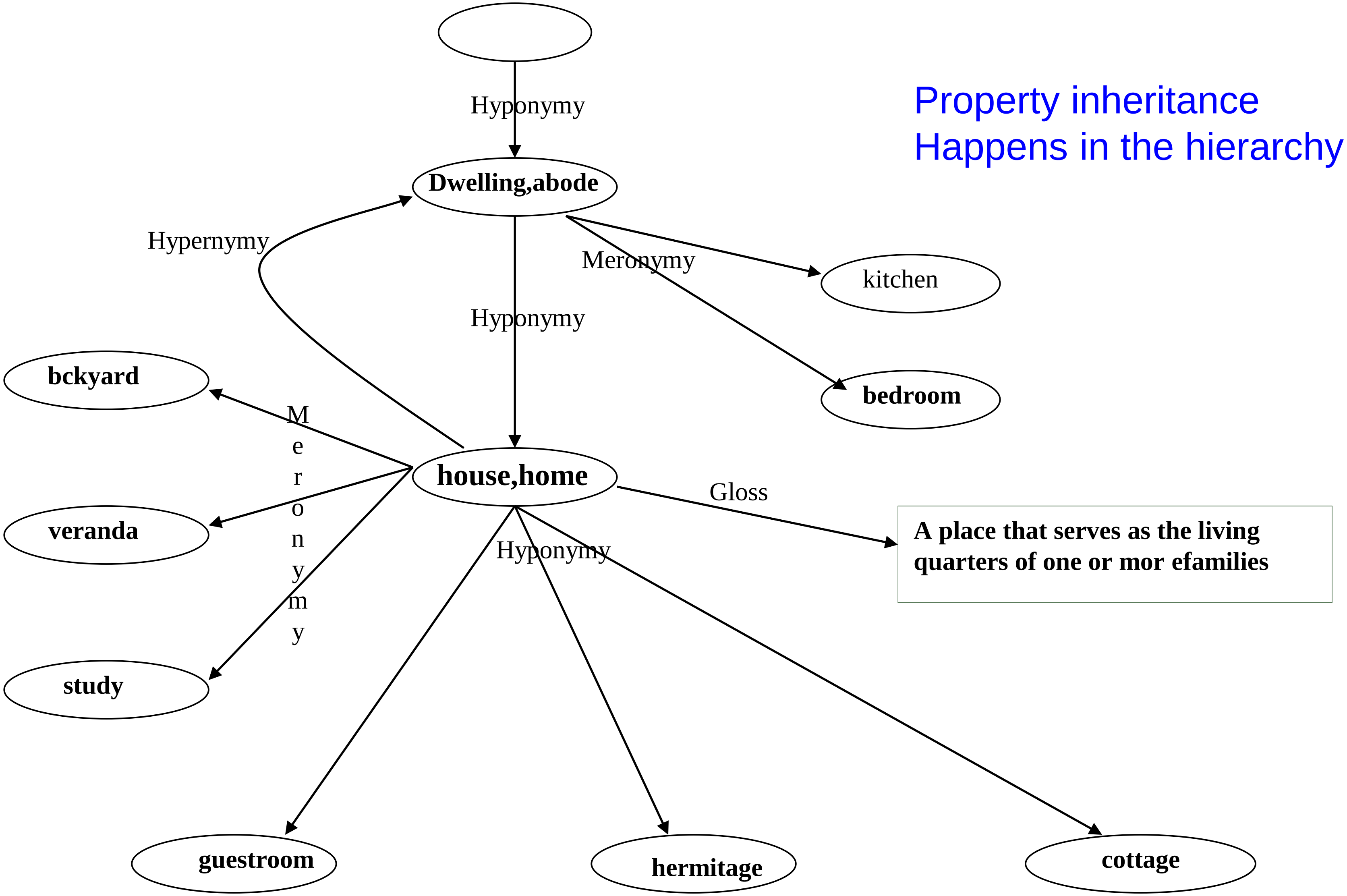
Lexical and Semantic relations in wordnet

1. Synonymy
2. Hypernymy / Hyponymy *(kind-of)*
3. Antonymy
4. Meronymy / Holonymy *(part of)*
5. Gradation
6. Entailment
7. Troponymy (*manner of)*

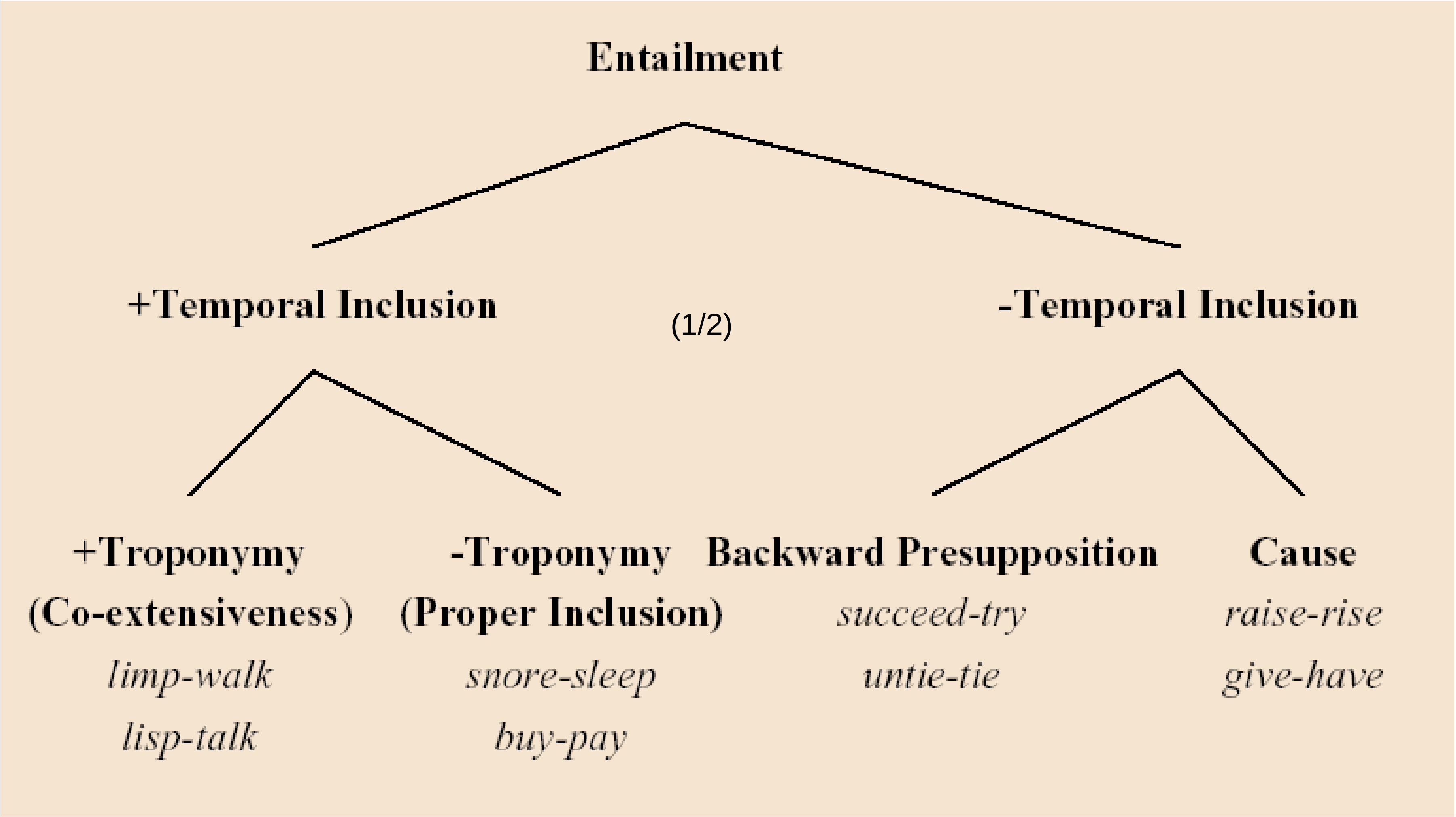
1, 3 and 5 are lexical (*word to word)*, rest are semantic (*synset to synset).*

WordNet Sub-Graph

Property inheritance
Happens in the hierarchy



Entailment: fundamental meaning relation linking verbs



Principles behind creation of Synsets

Three principles:

Minimality: (first decide the exact synonyms that are minimally needed to make the meaning unique)

Coverage: for that sense include ALL the words in the synset

Replacability: at least the first few words should be able to replace one anothere

Wordnet Engineering

Three Principles of Synset creation

- Minimality
- Coverage
- Replacability

Synset creation: example

Home

John's home was decorated with lights on the occasion of Christmas.

Having worked for many years abroad, John Returned home.

House

John's house was decorated with lights on the occasion of Christmas.

Mercury is situated in the eighth house of John's horoscope.

Synsets (continued)

{house} is ambiguous.

{house, home} has the sense of *a social unit living together;*

Is this the minimal unit?

{family, house} will make the unit completely unambiguous.

For coverage:

{family, household, house} ordered according to frequency.

Replacability of the most frequent words is a requirement which is satisfied

Representation using syntagmatic relations: Co-occurrence Matrix

Corpora: I enjoy cricket. I like music. I like deep learning

	I	enjoy	cricket	like	music	deep	learning
I	-	1	1	2	1	1	1
enjoy	1	-	1	0	0	0	0
cricket	1	1	-	0	0	0	0
like	2	0	0	-	1	1	1
music	1	0	0	1	-	0	0
deep	1	0	0	1	0	-	1
learning	1	0	0	1	0	1	-

Co-occurence Matrix

Fundamental to NLP

Also called **Lexical Semantic Association (LSA)**

Very sparse, many 0s in each row

Apply Principal Component Analysis (PCA) or Singular Value Decomposition (SVD)

Do Dimensionality Reduction; merge columns with high internal affinity (e.g., *cricket* and *bat*)

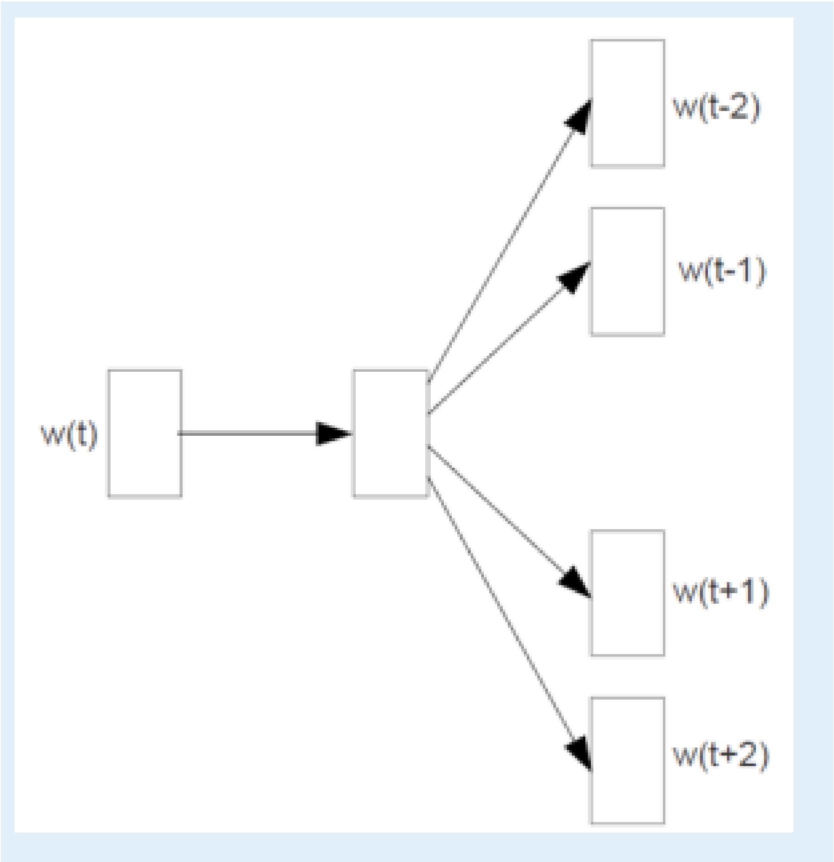
Compression achieves better semantics capture

Linguistic foundation of word representation by vectors

“Linguistics is the eye”: Harris Distributional Hypothesis

- Words with similar distributional properties have similar meanings. (Harris 1970)
- 1950s: Firth- “A word is known by the company its keeps”
- Model **differences** in meaning rather than the proper meaning itself

“Computation is the body”: Skip gram- predict context from word



For CBOW:

Just reverse the Input-Ouput

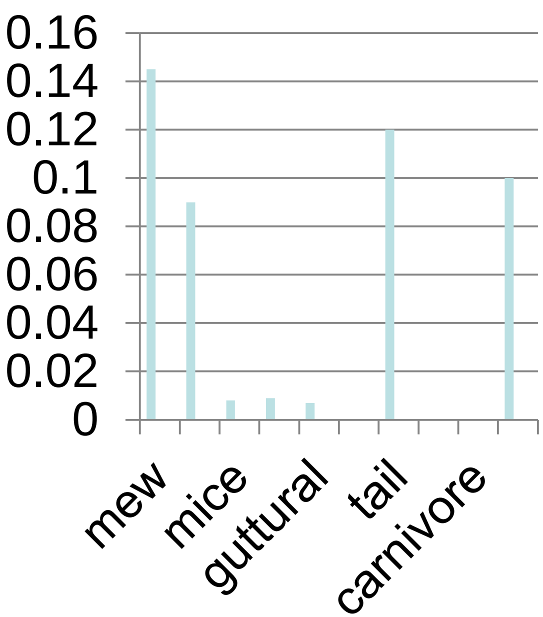
Dog – Cat - Lamp

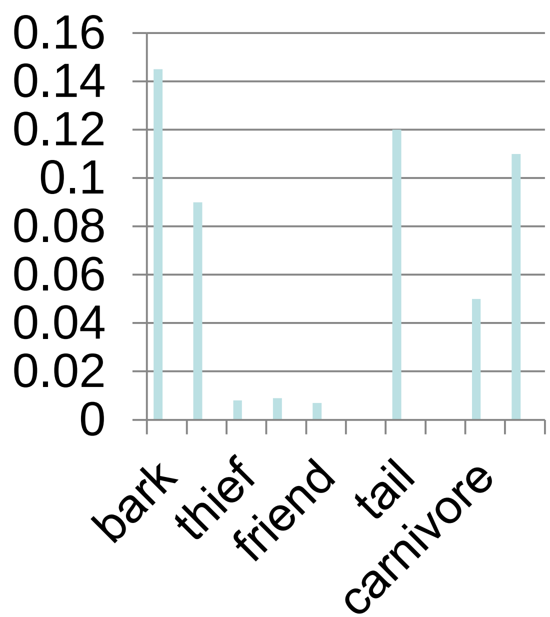
{bark, police, thief, vigilance, faithful, friend, animal, milk, carnivore)

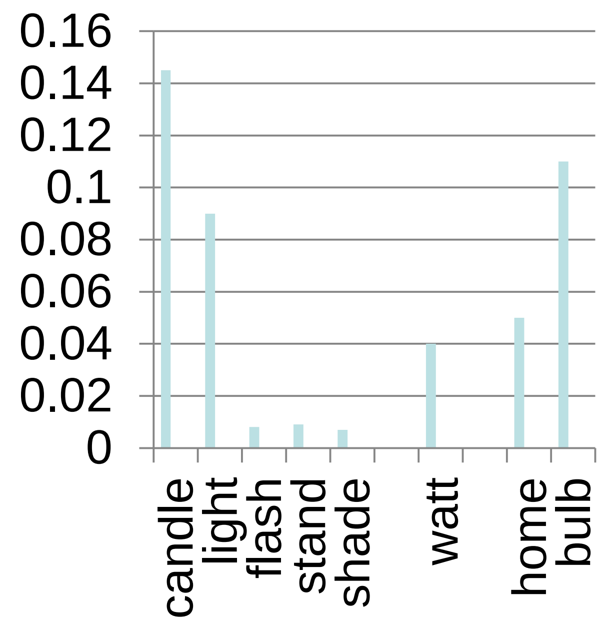
{mew, comfort, mice, furry, guttural, purr, carnivore, milk}

{candle, light, flash, stand, shade, Halogen}

Probability distributions of context words
CE(dog, lamp) > CE(dog, cat)







Test of representation

- **Similarity**
  - ‘Dog’ more similar to ‘Cat’ than ‘Lamp’, because
  - Input- vector(‘dog’), output- vectors of associated words
  - More similar to output from vector(‘cat’) than from vector(‘lamp’)

“Linguistics is the eye, Computation is the body”

The encode-decoder deep learning network is nothing but

the ***implementation*** of

Harris’s Distributional Hypothesis

Fine point in Harris Distributional Hypothesis

- Words with similar distributional properties have similar meanings. (Harris 1970)
- Harris does mentions that distributional approaches can model differences in meaning rather than the proper meaning itself

Representation Learning

Basics

- What is a good representation? At what granularity: words, n-grams, phrases, sentences
- Sentence is important- (a) *I <u>bank</u> with SBI; (b) I took a stroll on the river <u>bank</u>; (c) this <u>bank</u> sanctions loans quickly*
- Each ‘bank’ should have a different representation
- We have to LEARN these representations

Principle behind representation

- Proverb: “A man is known by the company he keeps”
- Similarly: “A word is known/represented by the company it keeps”
- “Company” → Distributional Similarity

Starting point: 1-hot representation

- Arrange the words in lexicographic order
- Define a vector *V* of size |*L*|, where *L* is the lexicon
- For word w_i in the i^{th} position, set the ith bit to 1, all other bits being 0.
- Problem: cosine similarity of ANY pair is 0; wrong picture!!

Representation: to learn or not learn?

- 1-hot representation does not capture many nuances, e.g., semantic similarity
  - But is a good starting point
- Co-occurences also do not fully capture all the facets
  - But is a good starting point

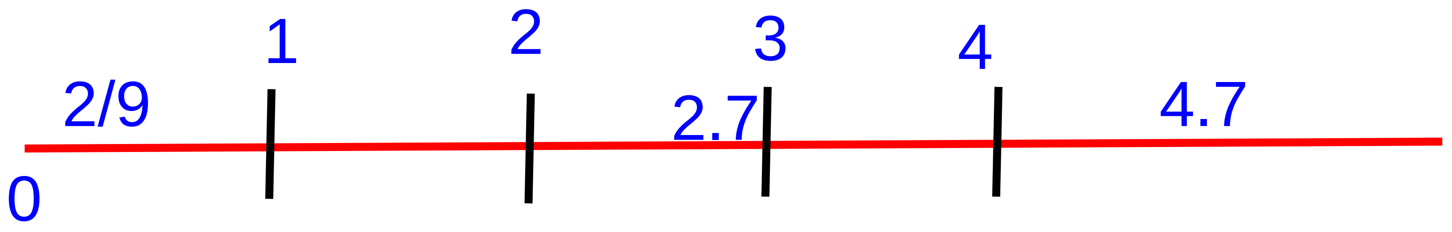
So learn the representation…

- Learning Objective

- ***MAXIMIZE CONTEXT PROBABILITY***

Foundations-1: Embedding

- Way of taking a discrete entity to a continuous space
- E.g., 1, 2, 3, 2.7, 2/9, $22^{1/2}$, … are numerical symbols
- But they are points on the real line
- Natural embedding
- Words' embedding not so intuitive!



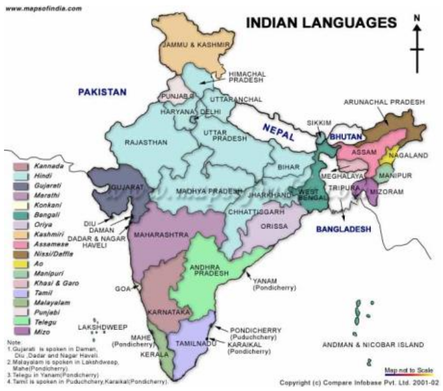
Foundations-2: Purpose of Embedding

- Enter geometric space
- Take advantage of “distance measures”- Euclidean distance, Riemannian distance and so on
- “Distance” gives a way of computing similarity

Foundations-3: Similarity and difference

- Recognizing similarity and difference- foundation of intelligence
- Lot of Pattern Recognition is devoted to this task (Duda, Hart, Stork, 2nd Edition, 2000)
- Lot of NLP is based on Text Similarity
- Words, phrases, sentences, paras and so on (verticals)
- Lexical, Syntactic, Semantic, Pragmatic (Horizontal)

Similarity study in MT



English:

This blanket is very soft

Hindi:

yaha kambal bahut naram hai

Bangla:

ei kambal ti khub naram <null>

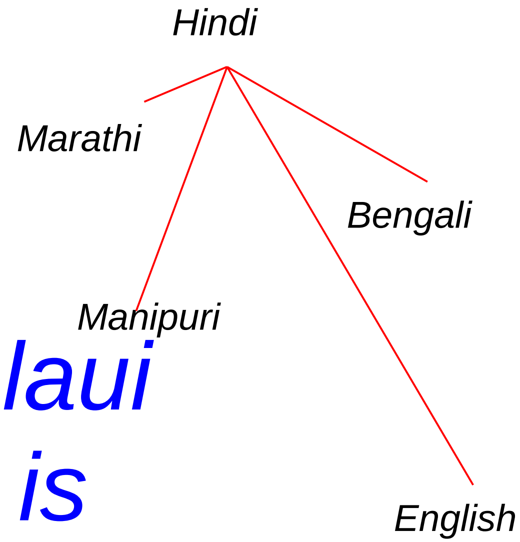
Marathi:

haa kambal khup naram aahe

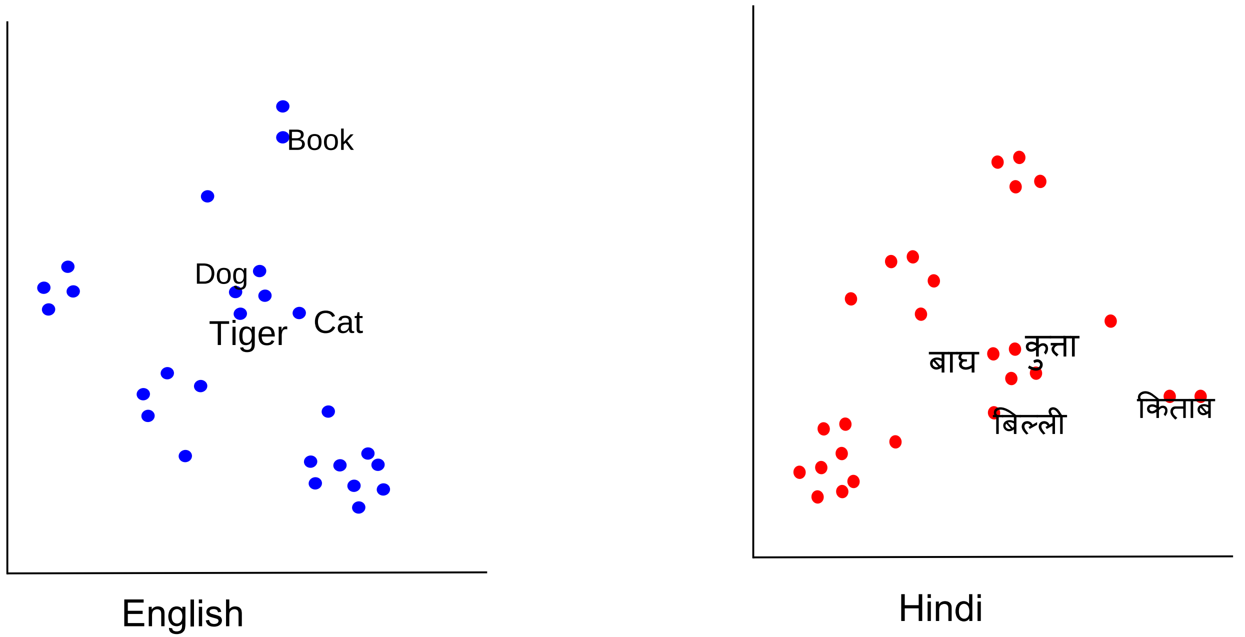
Manipuri:

kampor asi mon mon laui

blanket this soft soft is



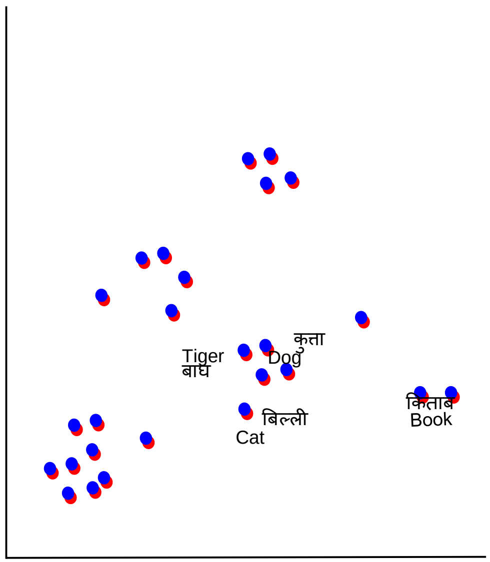
ISO-Metricity



Across Cross-lingual Mapping

This involves strong assumption that embedding spaces across languages are isomorphic, which is not true specifically for distance languages (Søgaard et al. 2018). However, without this assumption unsupervised NMT is not possible.

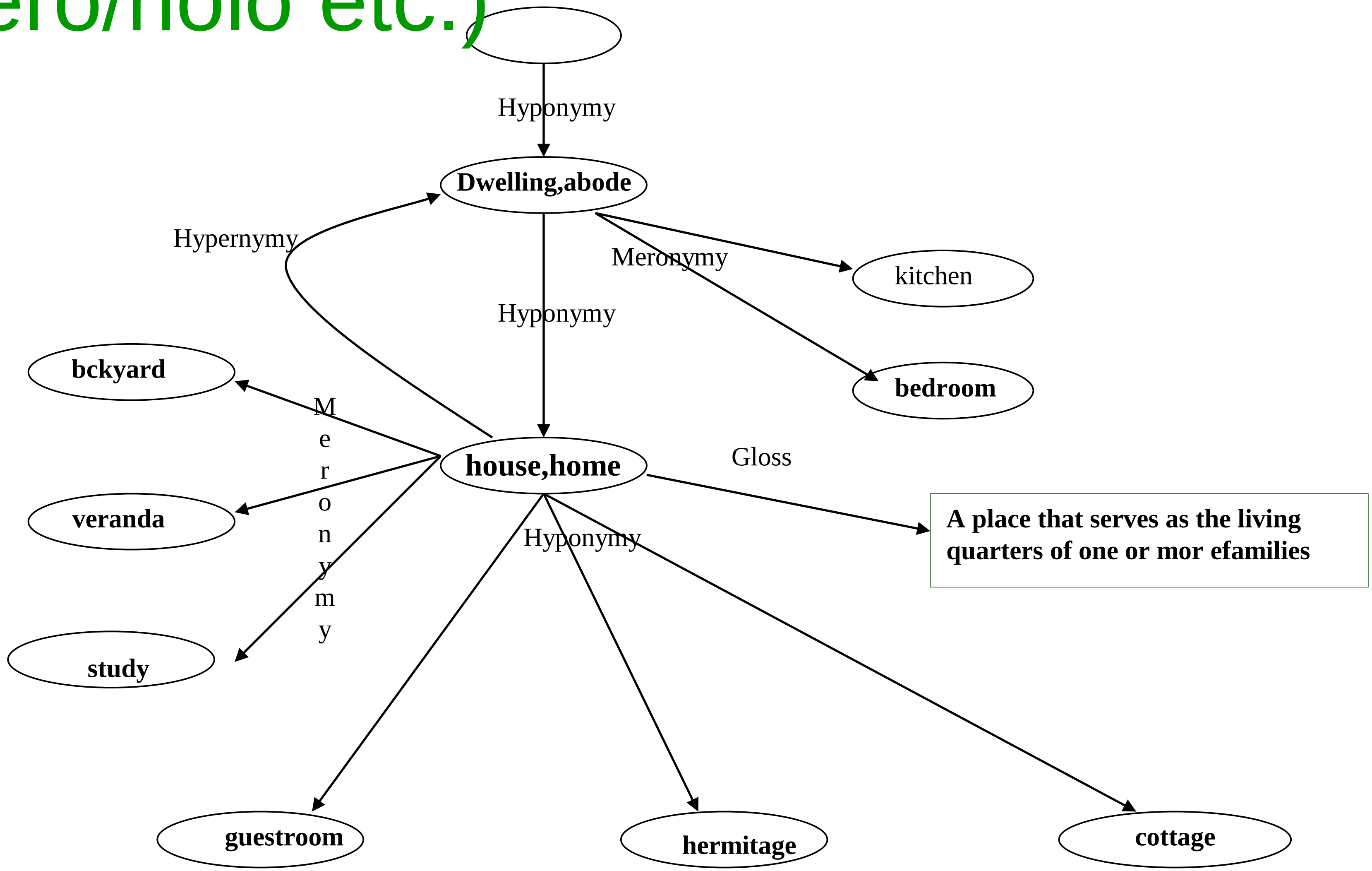
Søgaard, Anders, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. ACL



Foundations-4: Syntagmatic and Paradigmatic Relations

- Syntagmatic and paradigmatic relations
  - Lexico-semantic relations: synonymy, antonymy, hypernymy, mernymy, troponymy etc. **CAT is-a ANIMAL**
  - Coccurence: **CATS MEW**
- Wordnet: primarily paradigmatic relations
- ConceptNet: primarily Syntagmatic Relations

WordNet Sub-Graph with lexico-semantic relations (hyper/hypo, mero/holo etc.)



Lexical and Semantic relations in wordnet

1. Synonymy (e.g., *house, home*)
2. Hypernymy / Hyponymy (kind-of, e.g., *cat* ←→ *animal*)
3. Antonymy (e.g., *white and black*)
4. Meronymy / Holonymy (part of, e.g., *cat and tail*)
5. Gradation (e.g., *sleep*→*doze*→*wake up*)
6. Entailment (e.g., *snoring* → *sleeping*)
7. Troponymy (manner of, e.g., *whispering and talking*)

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

‘Paradigmatic Relations’ and ‘Substitutability’

- Words in paradigmatic relations can substitute each other in the sentential context
- E.g., ‘The cat is drinking milk’ → ‘The animal is drinking milk’
- Substitutability is a foundational concept in linguistics and NLP

Foundations-5: Learning and Learning Objective

- Probability of getting the context words given the target should be maximized (skip gram)

- Probability of getting the target given context words should be maximized (CBOW)

Learning objective (skip gram)

$$J'(\theta) = \frac{1}{T} \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \ne 0}} p(w_{t+j} \mid w_t; \theta)$$

$$J(\theta) = -\frac{1}{T} \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \ne 0}} p(w_{t+j} \mid w_t; \theta)$$

$$Minimize \quad L = -\sum_{t=1}^{T} \sum_{\substack{-m \le j \le m \\ j \ne 0}} \log[p(w_{t+j} \mid w_t; \theta)]$$

Modelling *P(context word|input word)* *(1/2)*

- We want, say, *P('bark'|'dog')*
- Take the weight vector **FROM** 'dog' neuron **TO** projection layer (call this u_{dog})
- Take the weight vector **TO** 'bark' neuron **FROM** projection layer (call this v_{bark})
- When initialized u_{dog} and v_{bark} give the initial estimates of word vectors of 'dog' and 'bark'
- The weights and therefore the word vectors get fixed by back propagation

Modelling *P(context word|input word)* *(2/2)*

- To model the probability, first compute dot product of u_{dog} and v_{bark}
- Exponentiate the dot product
- Take softmax over all dot products over the whole vocabulary

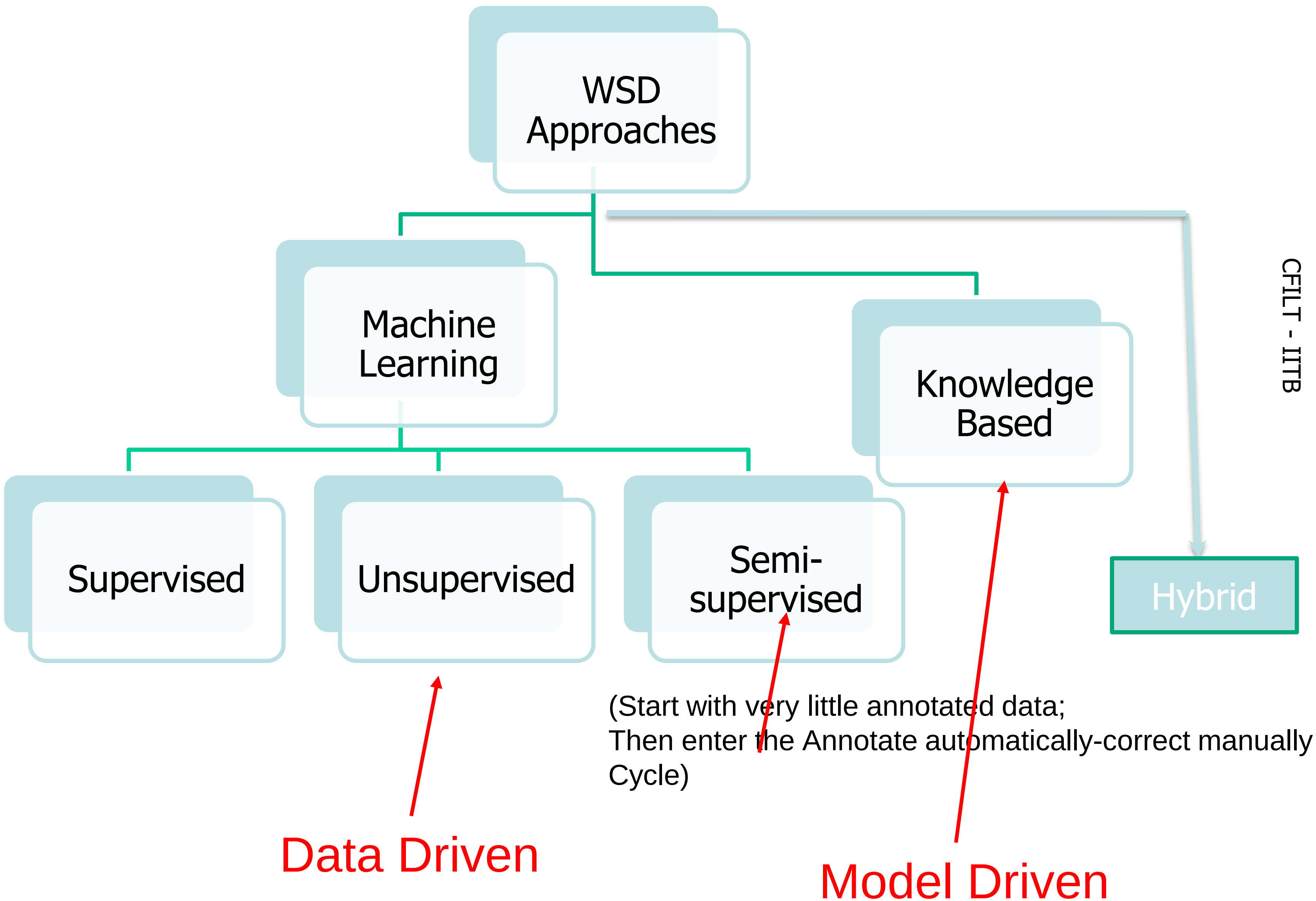
$$P('bark'|'dog') = \frac{\exp(u_{dog}^T v_{bark})}{\sum_{v_k \varepsilon Vocabulary} \exp(u_{dog}^T v_k)}$$

Exercise

- Why cannot you model *P('bark'|'dog')* as the ratio of counts of <bark, dog> and <dog> in the corpus?
- Why this way of modelling probability through dot product of weight vectors of input and output words, exponentiation and soft-maxing works?

Word Sense Disambiguation

Bird's eye view of WSD techniques



Wordnet - Lexical Matrix (with examples)

Word Meanings (IDs)	Word				
	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*) $E_{2,\ldots}$	
M_3		*(bank)* $E_{3,2}$	$E_{3,3}$		
…				…	
M_m					$E_{m,n}$

Sense tagged corpora (task: sentiment analysis)

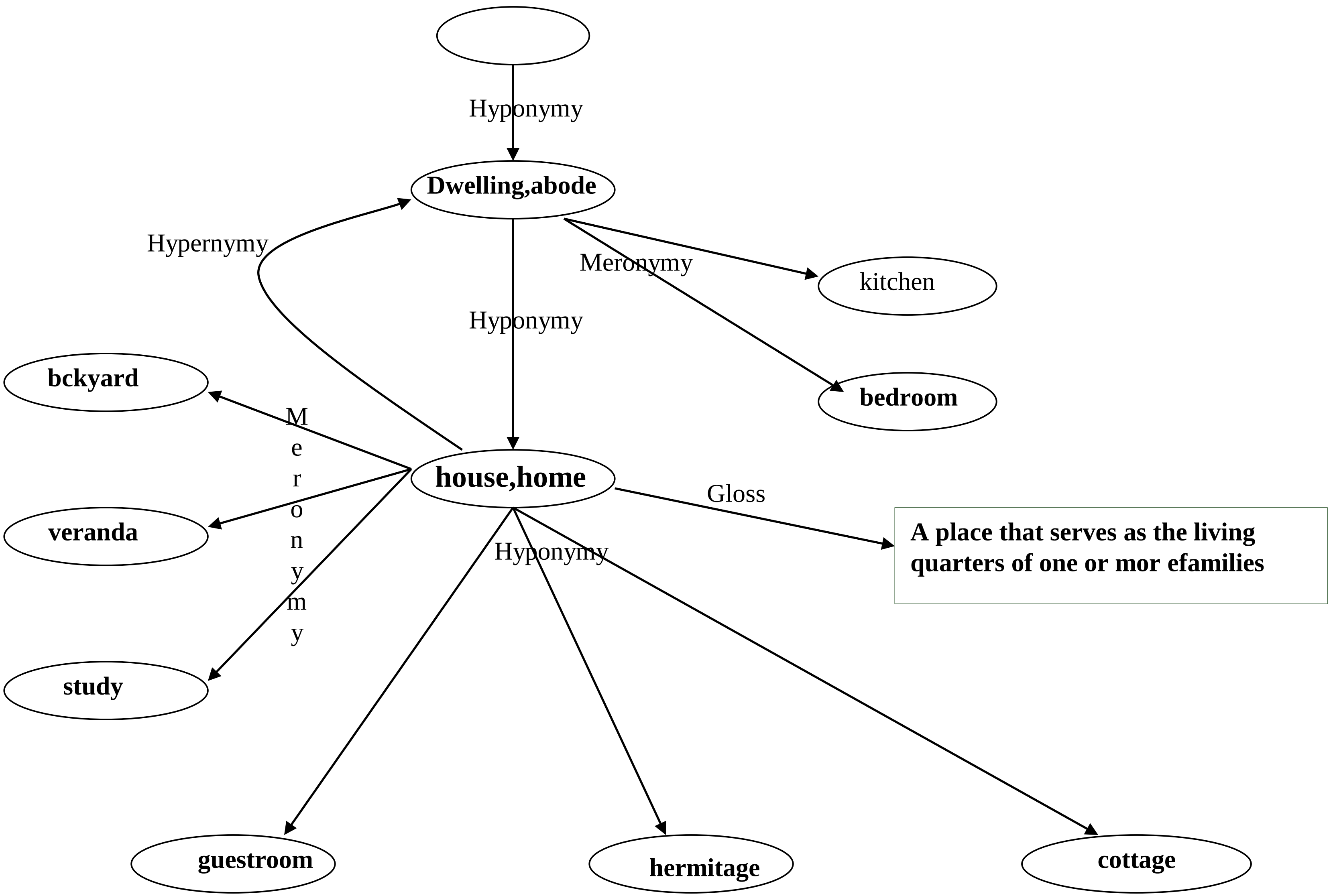
- I have enjoyed_21803158 #LA#_18933620 every_42346474 time_17209466 I have been_22629830 there_3110157 , regardless_3119663 if it was for work_1578942 or pleasure_11057430.
- I usually_3107782 fly_21922384 into #LA#_18933620, but this time_17209466 we decided_2689493 to drive_21912201 .
- Interesting_41394947,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

Senses of “pleasure”

The noun pleasure has 5 senses, 4 of which are shown below:

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")

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

Definition of WSD

- 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 sentences:
  - (a) the mouse ran away, (b) my mouse is not working
  - The senses are “animal” and “computer accessory”

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
$\text{UKB}_{\text{ppr\_w2w-nf}}$	57.5	64.2	54.8	40.0	64.5	64.5
$\text{UKB}_{\text{ppr\_w2w}}$	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**

A note on baselines: MFS and WFS

- Most frequent sense (MFS) is obtained from sense annotated data
- MFS algo is that given a new target word and its context, output that sense which is most frequent in the corpora
- This is a very difficult to beat baseline
- Wordnet first sense (WFS) is the sense that is given the first position in the ranked order of senses, as per frequency, often a linguistic judgement
- WFS algo is simply output the first sense of the target word.
- Both MFS and WFS algo are context insensitive

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.

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

$$
\begin{aligned}
&P(S \mid w, C) \\
&= \frac{P(S, w, C)}{P(w, C)} \\
&= \frac{P(w)P(S, C \mid w)}{P(w)P(C \mid w)} \\
&= \frac{P(S, C \mid w)}{P(C \mid w)} \\
&= \frac{P(S \mid w)P(C \mid S, w)}{P(C \mid w)}
\end{aligned}
$$

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 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)

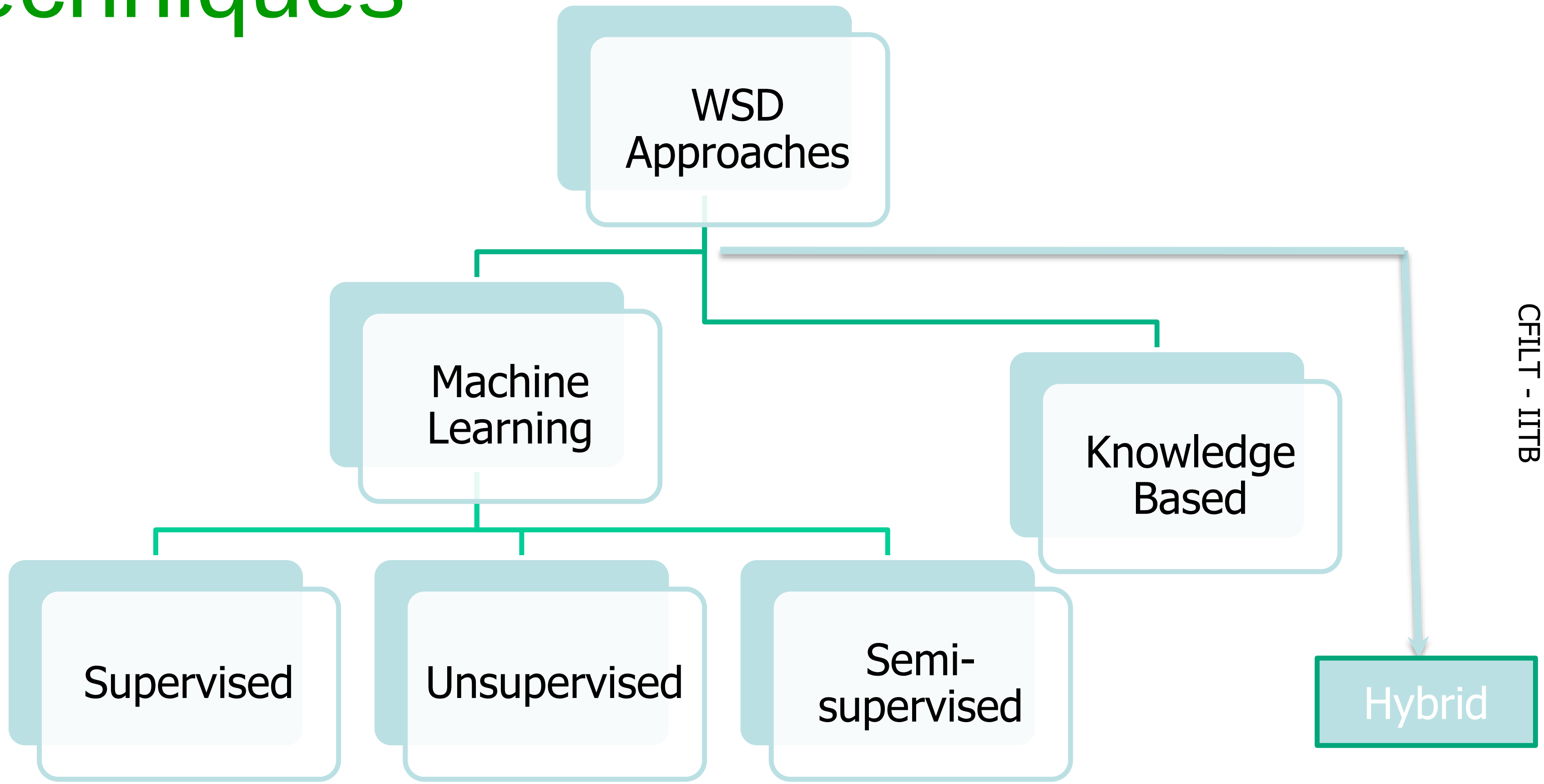
Example: Modelling of WSD (3/3)

- 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 Naive Bayes assumption (Bi-gram)
  - P(he| W^s)*P(has| W^s)......P(horoscope| W^s)

Observations on parameters

- The word ‘horoscope’ is a very strong signal for astrology sense; high *P(‘horoscope’|house*$^{astrology\text{-}sense}$*)*
- *P(‘horoscope’|house*$^{dwelling\text{-}sense}$*)* will be a weak signal
- Similarly for *P(‘horoscope’|house*$^{family\text{-}sense}$*)*
- Words like ‘he’, ‘his’ etc. are non-discriminative

Revisit: Bird's eye view of WSD techniques



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 context.*

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)

LESK'S ALGORITHM (contd..)

- Note the **importance of lower layer tasks** in NLP stack for a higher layer task like **Word Sense Disambiguation**
  - **Morphological Analysis**: Comparing the root words while finding overlap could be useful
  - Ex: 'burned' and 'burning' have the same root word in the previous example
  - **POS Tagging:** Identifying the POS tag of a word would reduce the search space while finding its sense
  - Ex: Finding out POS of 'ash' as noun reduces the

CRITIQUE

- Many times there may not be any overlap: sparsity problem
  - The ash from the combustion
- Overlap may be spurious leading to “drift”
  - *The ash tree was burned*
- Proper nouns as as strong disambiguators, but not present in WN

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

  Sachin Tendulkar plays **cricket.**

- Typical Accuracy
  - 50% when tested on 10 highly polysemous English words.

Extended Lesk's algorithm

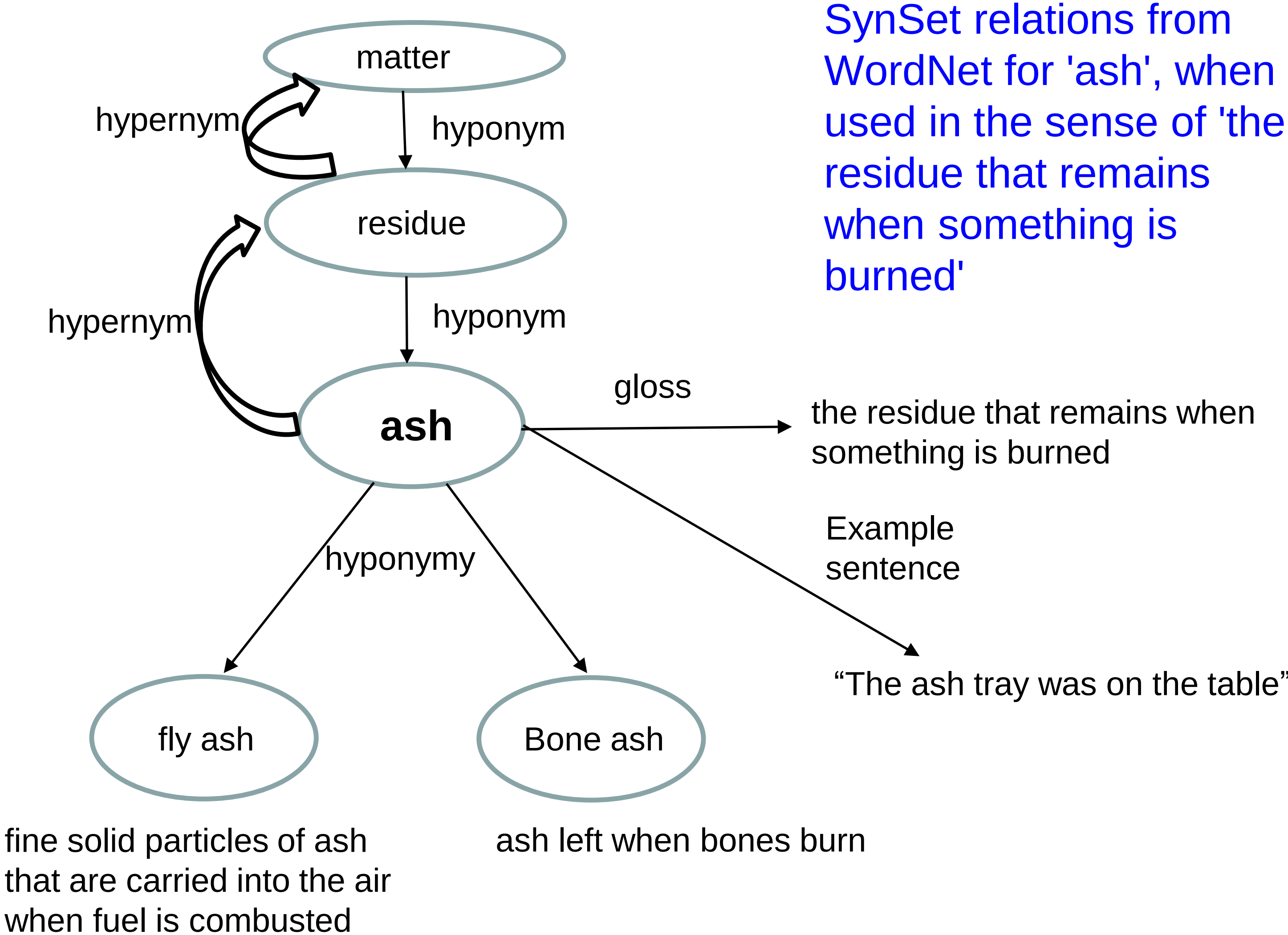
- Extension includes glosses of semantically related senses from WordNet (e.g. *hypernyms*, *hyponyms*, etc.).

- The scoring function now computes the overlap of context bag with not only the words local to the synset but also words occurring in neighhjboring synsets

- Vide next slide

WordNet Sub-graph

SynSet relations from WordNet for 'ash', when used in the sense of 'the residue that remains when something is burned'



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

  - E.g. for "there were some bones under the ash tree"→ Spurious overlap with bone under "bone ash"

What if overlaps tie?

- There is “tree” also in the context
- Both “bone” and “tree” will contribute equally to overlap
- Then we will invoke other factors like PROXIMITY which is also called SANNIDHI in Indian linguistic tradition (SANNIDHI means “proximity”)
- AKANGJSHA (desire), YOGYATA (suitability) and SANNIDHI (proximity) are fundamental disambiguators
- Since “tree” is *CLOSER* to “ash”, ash tree will be the winner sense

Argument Frame Selection Preference

- “eat” and “rice”

- Eat needs an object→ akangksha (argument)

- Object should be edible, rice is edible→ yogyata (selectional preference)

WSD using Sense Embedding

- We will create the **sense embedding** by averaging the word vector for each word in the Gloss.

  E.g. “On burning *coal* we get ***ash***.”

- We have three senses from Wordnet

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

2. 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)

- sense_emb = sum of word vector of each word in Gloss /# of words in Gloss

- context_emb = sum of word vector of each word in input /# of words in input

WSD using Sense Embedding (cont'd…)

- sense_emb = sum of word vector of each word in Gloss /# of words in Gloss
- context_emb = sum of word vector of each word in input /# of words in input
- Compute the cosine similarity between each sense embedding and context embedding:

  *similarity_with_sense_1 = cosine_similarity(sense_emb_1, context_emb)=**0.4675***

  similarity_with_sense_2 = cosine_similarity(sense_emb_2, context_emb) =0.4315

  similarity_with_sense_3 = cosine_similarity(sense_emb_3, context_emb)=0.4019

- The sense having the maximum cosine similarity will be the disambiguated sense for the given context word.

$$best\_sense \quad = \; argmax \; (\; similarity\_with\_sense\_i \;) \quad \forall \, i$$

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

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

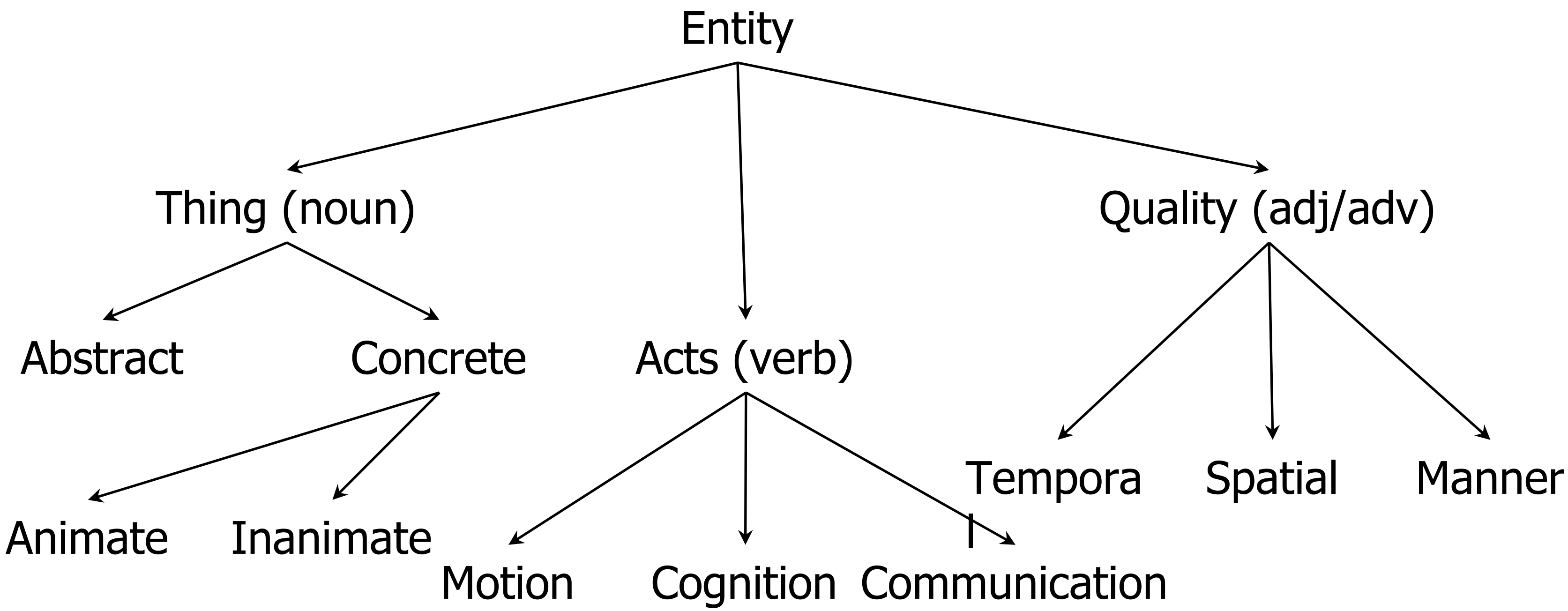
Walker Algo cntd.

- Thesaurus is a systematic organization of concepts
- “bank”, “interest”, “annum” etc. appear in the finance domain and contribute to each others count in the walker algo
- Lesk insists on local exact symbol match
- Extended lesk on inside and outside synset matches
- Walker insists on domain (concept category) matching

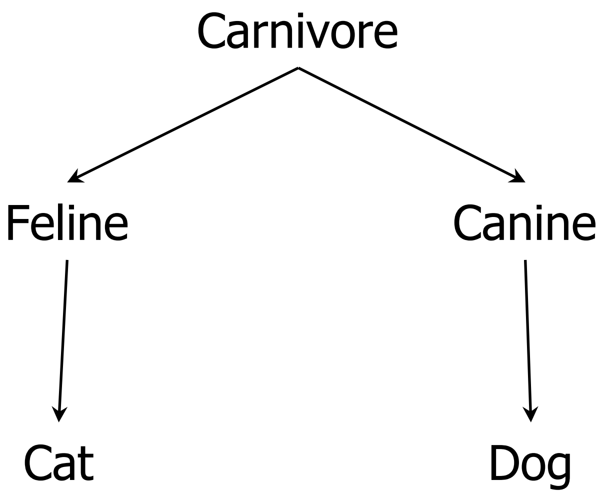
WSD USING CONCEPTUAL DENSITY *(Agirre and Rigau, 1996)*

- 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 (more heterogeneity)
- Semantic similarity will reduce at higher levels

Relevance in the era of DL-NLP

- The notion of conceptual density is important for DL-NLP too
- Similarity in DL-NLP is computed by cosine similarity of word vectors
- Word vectors are created exploiting SYNTAGMATIC relations (coming from corpus)
- Ontology based similarity is computed using PARADIGMATIC relations

Conceptual Density to be cntd.