Computational Linguistics

Lecture 9

Dr. Dina Khattab

dina.khattab@cis.asu.edu.eg

LEXICAL SEMANTICS

- The lexicon as a component of human language.
- Theoretical models of the lexicon.
- Computational models of the lexicon (WordNet).
- The lexicon in Natural Language Processing: Similarity and word sense disambiguation.

The Lexicon

Definition: Collection of all words of a language.

- Very large (> 40000 words, dictionaries).
- Cannot enumerate its entries.
- Words come and go.
- Words change their meaning, usage, pronunciation.

Components of Lexicon

- ➢ Sound (pronunciation).
- ➢ Morphology (e.g., plural formation: woman → women, house → houses)
- ➢ Syntax: sub-categorization/ selectional restrictions

Selectional Restriction Example:

- Direct objects of word "eat" can only be things that are considered food.

Selectional Preference Example:

- "strong tea", not "powerful tea"

Lexicon's Representation of Meaning

- Most important: meaning/concept behind the word.
- Assumption: given the large number of words and concepts, there must be an organizing system.

Concepts Connecting Words to Meaning

Synonymy: Different words with the same meaning.

Polysemy/Homonymy

Same word with multiple meanings.

- Polysemy: Slightly different meanings.
- Homonomy: Completely different meanings.

Synonymy

One meaning/concept is expressed by several different word forms:

- beat, hit, strike, shut, close
- car, motorcar, auto, automobile
- big, large, difficult, hard

Homonomy

- Homonymy: multiple unrelated meanings

- E.x: bank (money institution/elevated land)

Polysemy

- One word form expresses multiple meanings

- E.x: table: tabular array, piece of furniture

  newspaper: (paper copy/institution/building)

Polysemy / Homonomy

Polysemy Test

- He left Rome and later the country.
- He left the bills and an apple on the table.

Metonymy: conflating a part and the whole

- The “White House” issued a statement.
- The “office” isn't answering the phone.

Metaphor:

- “Time is Money” (conventional)
- My surgeon is a butcher (unconventional)
- This restaurant is a zoo on weekends (unconventional)

Computational Models of Word Meaning

- PMI: Measuring word association (e.g., selectional preference)
- Computational Lexicons (WordNet)

Computational Measures of Association

- Pointwise Mutual Information (PMI): Measures how strongly two words are with each other in a text.
- Compares joint probability of two words probability of observing each word independently

$$PMI_{x,y} = \log \frac{p(x,y)}{p(x)p(y)}$$

- Where p(x) is the probability of a word appearing in a sentence
- Interpretation: the number of bits of information obtained about probability of x given you’ve observed y (or vice-versa)

PMI Examples

- If PMI close to 0, then x and y are independent.
- If PMI is positive, then seeing one tells you that you’re more likely to see the other.

Videocassette …. recorder	**15.94**
Unsalted ….. butter	**15.19**
Time ….. last	**0.29**

Word Sense Disambiguation (WSD)

- Given a polysemous word in context, which meaning is correct?
- Necessary if we want a computerized, unambiguous representation of meaning.
- Current state of the art: 70% precision

What’s a Meaning?

- To answer this problem, we need an inventory of meanings.
- Most common solution: WordNet

History of WordNet

- Electronic dictionary of words and meaning.
- George Miller and Christiane Fellbaum (Princeton)

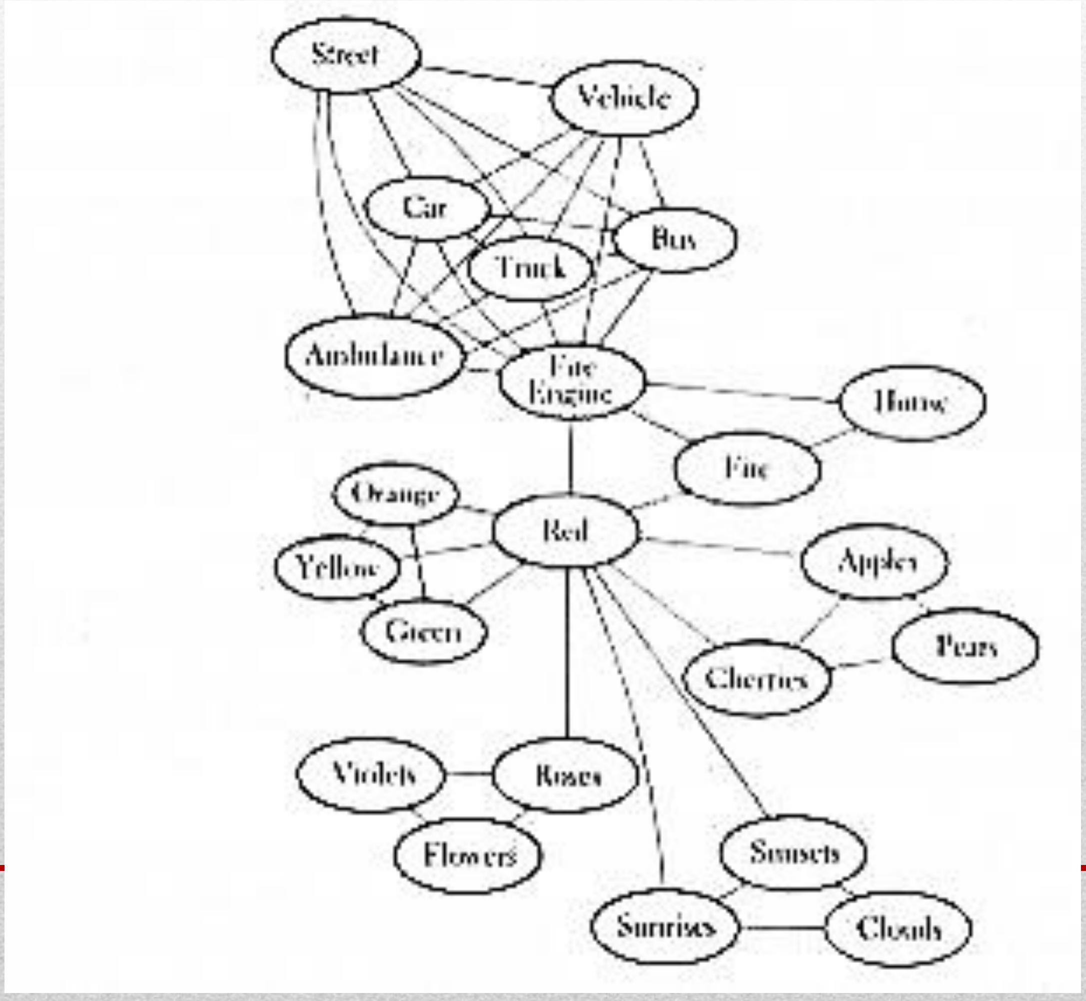
  http://wordnet.princeton.edu
- Includes most English nouns, verbs, adjectives and adverbs.
- Inspired WordNets in many other languages (70).

Organization

- Existing (western) dictionaries: organize by sound.
- Syntax-based organization
  - Levin (1993): syntactic properties of English
  - VerbNet (Kipper and Palmer)
- The WordNet model
  - Semantics-based.
  - Words are interconnected by meaning relations.

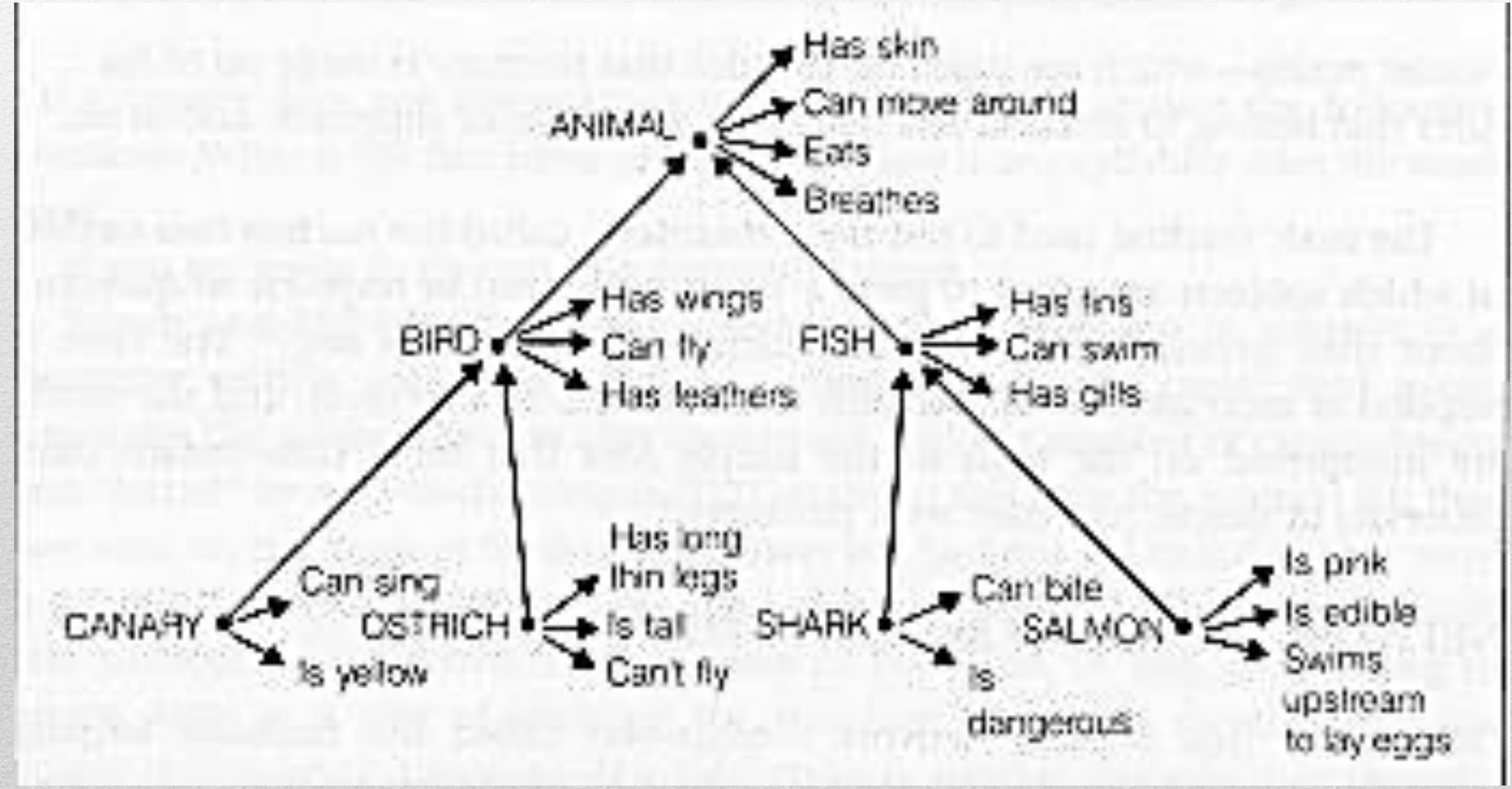
Semantic Network (Psychological Foundation)

- Semantic network representation (Collins and Quillian 1969, 1970, 1972)



Semantic Network (Psychological Foundation)

- Semantic network representation (Collins and Quillian 1969, 1970, 1972)



Semantic Network

- Collins & Quillian (1969) measured reaction times to statements involving knowledge distributed across different “levels”.
- Responses to statements like
  - Do birds move?
  - Do canaries move?
  - Do canaries have feathers?
  - Are canaries yellow?
- Reaction times varied depending on how many nodes had to be traversed to access the information.

What are the Connections in WordNet?

- If the (English) lexicon can be represented as a semantic network (a graph), what are the links that connect the nodes?
- WN distinguishes two kinds of links
  - Links among nodes (concepts) are conceptually semantic (e.g., bird-feather)
    - Hyponomy
    - Meronomy
    - Synonomy
  - Links among specific word forms are lexical (e.g., feather-feathery)

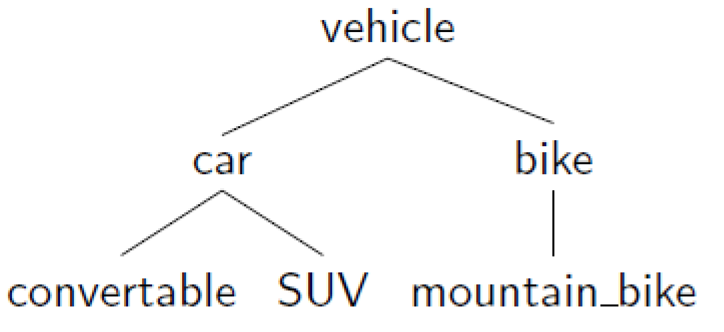
WordNet Statistics

- Synsets: are interconnected Bi-directional arcs express semantic relations

Part of speech	Word forms	Synsets
noun	117,798	82,115
verb	11,529	13,767
adjective	21,479	18,156
adverb	4,481	3,621
Total	155,287	117,659

Hypo/hypernym

- ➢Hypo-/hypernymy relates noun synsets.
- ➢ Relates more/less general concepts.
- ➢ Creates hierarchies, or "trees".



- ➢"A car is a kind of vehicle" – "Is a" relation.
- ➢Noun hierarchies have up to 16 levels.

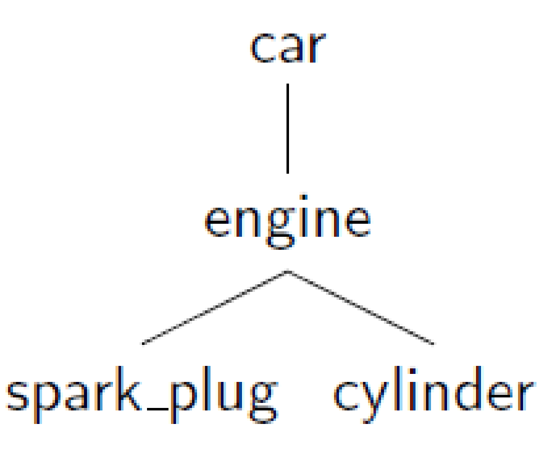
Meronymy / holonymy

- ➢Meronymy/holonymy: "part-whole" relation
- "An engine has spark plugs"
- "Spark plus and cylinders are parts of an engine"

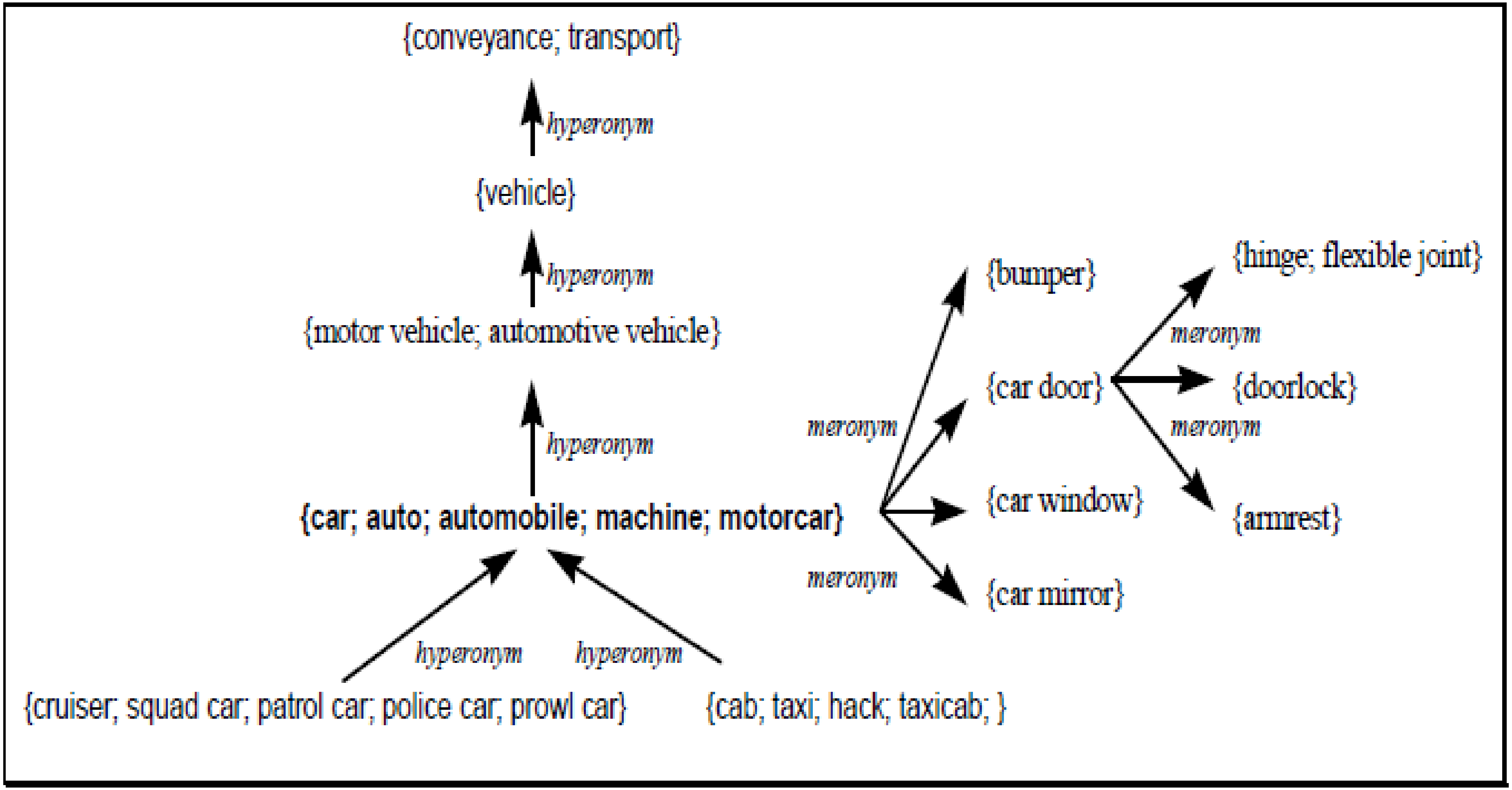
Inheritance:

- ➢A finger is part of a hand
- ➢A hand is part of an arm
- ➢An arm is part of a body

A finger is part of a body



WN Structure



Antonymy & troponymy

- Adjective relations: antonymy

E.x: hot-cold, old-new, high-low, big-small

- Verb relations: troponymy

E.x: move-walk, whisper-talk, smack-hit, gobble-eat

- Other relations among verbs reflect temporal or logical order between two events
  - divorce-marry (backward presupposition)
  - pay-buy (inclusion)
  - kill-die (cause)

WN Similarity: Resnik

- Probability of a concept: need all the words contained in the concept
- E.g.: w(mammal) = { hampster, monkey, horse, . . . }

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

- Given a corpus with N words (double count words with multiple meanings)
- Define the information content of a concept as IC = −log P(c)
  - “entity” concept has 0.0 information content.
  - “dog” has much higher information content.

WN Similarity: Resnik

- Least-Common Subsumer (LCS): Lowest (most-specific) node in WordNet containing a common ancestor
  - lcs(cat, dog) = mammal
  - lcs(wolf, puppy) = canine

  Here LCS is the concept that has the shortest distance from the two concepts compared. For example, animal and mammal are the subsumers of cat and dog but mammal is the lower subsumer than animal for them.

- Resnik similarity $\mathrm{Sim}_{\mathrm{Res}}\ (S_1, S_2) = \mathrm{IC}\ [\mathrm{lcs}(S_1,S_2)]$

Word Sense Disambiguation (WSD)

- Supervised: Tag a bunch of sentences and train a classifier.

- Difficulty: Getting training data
  - People don’t agree (unlike parsing)
  - Requires skilled annotators

70% accuracy