# Natural Language Processing (3)

## Word Senses and Embeddings

Daisuke Kawahara

Department of Communications and Computer Engineering,
Waseda University

# Lecture Plan

1. Overview of Natural Language Processing
2. Formal Language Theory
3. Word Senses and Embeddings
4. Topic Models
5. Collocations, Language Models, and Recurrent Neural Networks
6. Sequence Labeling and Morphological Analysis
7. Parsing (1)
8. Parsing (2)
9. Transfer Learning
10. Knowledge Acquisition
11. Information Retrieval, Question Answering, and Machine Translation
12. Guest Talk (1)
13. Guest Talk (2)
14. Project: Survey or Programming
15. Project Presentation

# Word Sense

- Intension: the ideas, properties, or corresponding signs that are implied or suggested by a concept (or word).
  - A = {x | x is an odd number less than 10}
  - (dictionary definition)
    **plant** a living thing that has leaves and roots and obtains most of its energy from sunlight via photosynthesis
- Extension: the set of things to which a concept (or word) extends or applies.
  - A = {1, 3, 5, 7, 9}

# Metaphor / Metonymy

- Metaphor
  - How can I kill a process? [Martin, 88]
  - My car drinks gasoline. [Wilks, 78]
  - He shot down all of my arguments. [Lakoff & Johnson, 80]
  - He is a big star. 
- Metonymy
  - Washington and Tokyo agree on ...
  - The ham sandwich is waiting for his check. [Lakoff & Johnson, 80]
  - Japanese people often eat nabe in winter.

# Thesaurus

- A kind of dictionary which lists words grouped together according to similarity and shows their generic/specific relations.
  - *Roget's Thesaurus*, by Peter Mark Roget, published in 1852.
  - *WordNet,* compiled in 1990s at Princeton Univ. extended to EuroWordNet, IndoWordNet, Chinese WordNet, Japanese WordNet, …

# WordNet

**motor vehicle, automotive vehicle** (a self-propelled wheeled vehicle that does not run on rails)

hypernym

**car, auto, automobile, machine, motorcar** (a motor vehicle with four wheels)

meronym

**automobile engine** (the engine that propels an automobile)

**car door** (the door of a car)

...

hyponym

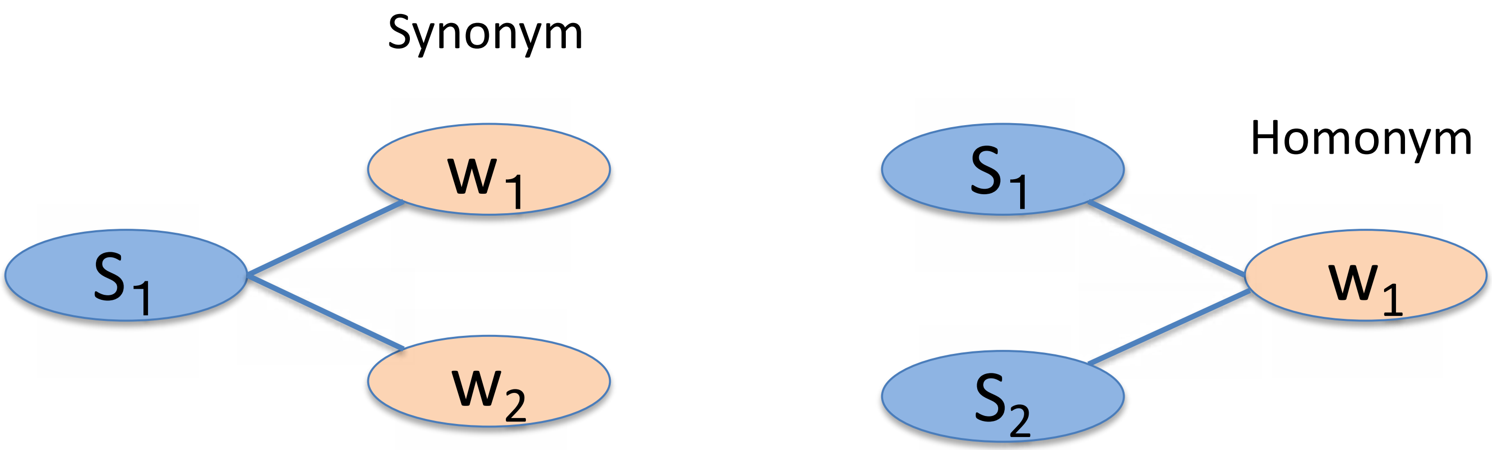
**ambulance** (a vehicle that takes people to and from hospitals)

**bus, jalopy, heap** (a car that is old and unreliable)

**cab, hack, taxi, taxicab** (a car driven by a person whose job is to take passengers where they want to go in exchange for money)

...

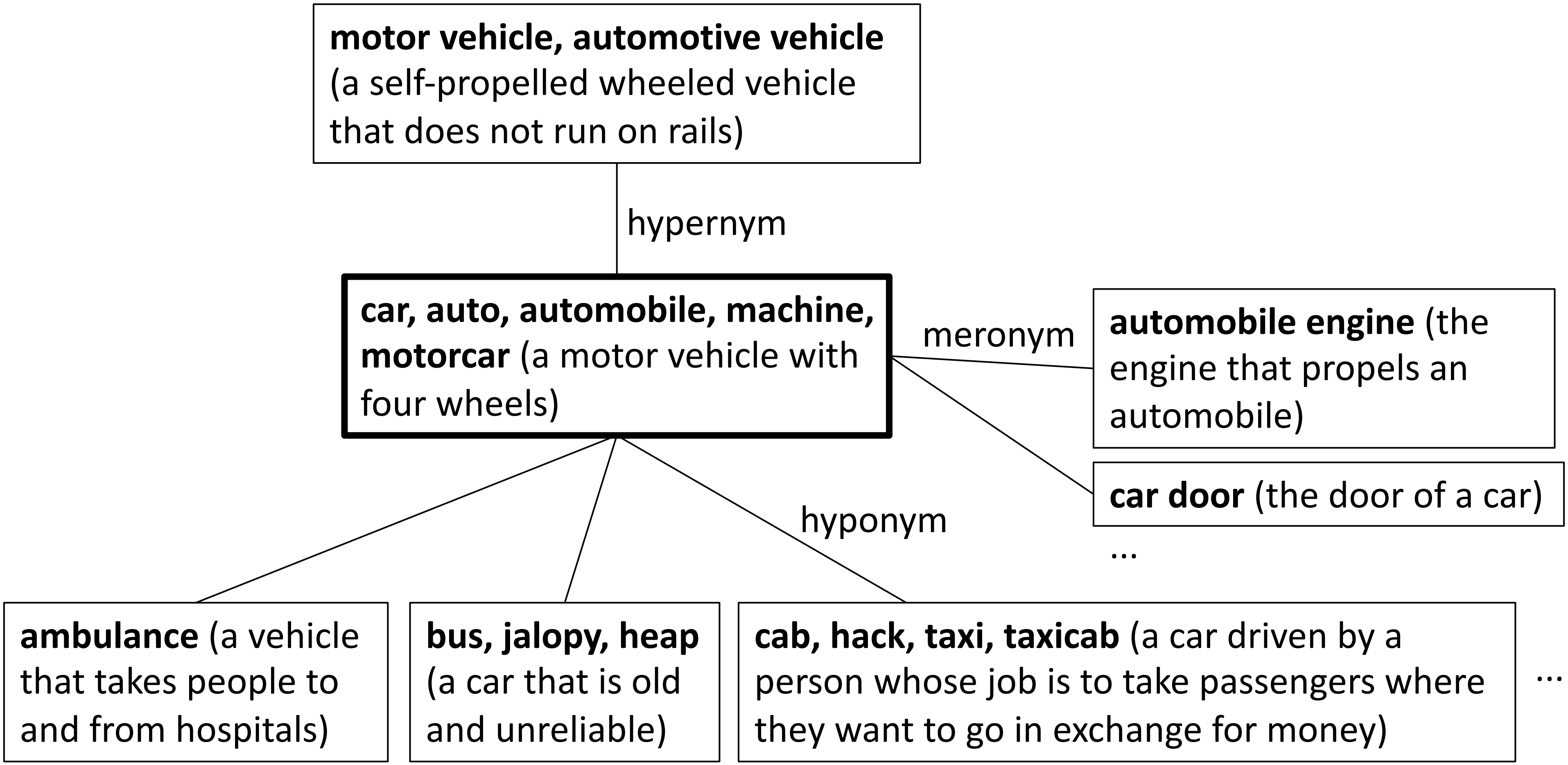
# Synonymy and Homonymy



# Synonyms

- Spelling variations
  - center, centre
  - 林檎, りんご, リンゴ 
- Different words (synonym … near synonym)
  - apple, アップル, 林檎 (translation)
  - NLP, Natural Language Processing (acronym)
  - helium, He;  meeting, mtg (abbreviation)
  - big, large

# WordNet


**motor vehicle, automotive vehicle** (a self-propelled wheeled vehicle that does not run on rails)
hypernym
**car, auto, automobile, machine, motorcar** (a motor vehicle with four wheels)
meronym
**automobile engine** (the engine that propels an automobile)
**car door** (the door of a car)
...
hyponym
**ambulance** (a vehicle that takes people to and from hospitals)
**bus, jalopy, heap** (a car that is old and unreliable)
**cab, hack, taxi, taxicab** (a car driven by a person whose job is to take passengers where they want to go in exchange for money)
...

# Distributional Similarity

- Distributional Hypothesis: words that occur in the same contexts tend to have similar meanings [Harris 1954; Firth 1957]

- Contexts are defined by related words judged by PMI (pointwise mutual information)

$$PMI(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

# Distributional Similarity

- Similarity measure:
  - Jaccard coefficient $$\frac{|X \cap Y|}{|X \cup Y|}$$
  - Simpson coefficient $$\frac{|X \cap Y|}{\min(|X|,|Y|)}$$
  - Dice coefficient $$\frac{2|X \cap Y|}{|X|+|Y|}$$

(*X*: related words for *x*;  *Y*: related words for *y*)

# Distributional Similarity

| | 医師 | 医者 |
|---|---|---|
| ～の診察 (observation of … ) | 8225 | 495 |
| ～に相談 (consult …) | 4374 | 1359 |
| ～の許可 (admission of …) | 1474 | 254 |
| . . | . . | . . |

0.382

| Similar word | Sim. |
|---|---|
| ドクター (doctor) | 0.395 |
| 医者 (doctor) | 0.382 |
| 先生 (teacher) | 0.374 |
| 獣医 (veterinary) | 0.350 |

Similar words with 医師

# Examples of Similar Words

- コンピュータ (computer)
  - 計算機(computer) : 0.44, パソコン(personal computer) : 0.40, Macintosh : 0.39, プリンタ(printer) : 0.32, ノートパソコン(notebook computer) : 0.29
- ゲーム (game)
  - RPG : 0.40, ドラクエ(Dragon Quest) : 0.38, オンラインゲーム(online game) : 0.37, ビリヤード(billiard) : 0.36, FF : 0.32
- メタボ (metabolic syndrome)
  - 花粉症(pollen allergy) : 0.32, 病気(disease) : 0.30, 病(disease) : 0.26, 癌(cancer) : 0.24

Words with red color mean these words are not listed in a thesaurus.

# Homonyms / Polysemic Words

bank


interest

# Homonyms / Polysemic Words

- homonym
  - *bank*: **Different origins (English, Italian)**
    1. The banks of a river, canal, or lake are the raised areas of ground along its edge.
    2. A bank is an institution where people or businesses can keep their money.
- polysemic words
  - *interest*: **Same origins (English, Italian)**
    1. If you have an interest in something, you want to learn or hear more about it.
    2. Interest is extra money that you receive if you have invested a sum of money.

[Collins COBUILD]

# Systematic Polysemy

- “the act of X” and “the people doing X”
  e.g., competition, organization

- “the act of X” and “the result of doing X“
  e.g., deposit

# Word Sense Disambiguation

- Ambiguity
  – Many words have several meanings (senses)

- Methods for disambiguation
  – Dictionary-based disambiguation
  – Unsupervised disambiguation
  – Supervised disambiguation

# Upper and Lower Bounds

- Upper bounds
  - Human agreement
    - over 95% for clearly distinct senses (e.g., bank)
    - 65% to 70% for polysemous words with many related senses (e.g., title, side, way)
- Lower bounds **First sense**
  - Simplest possible algorithm:
    - most frequent sense
    - first sense in a dictionary

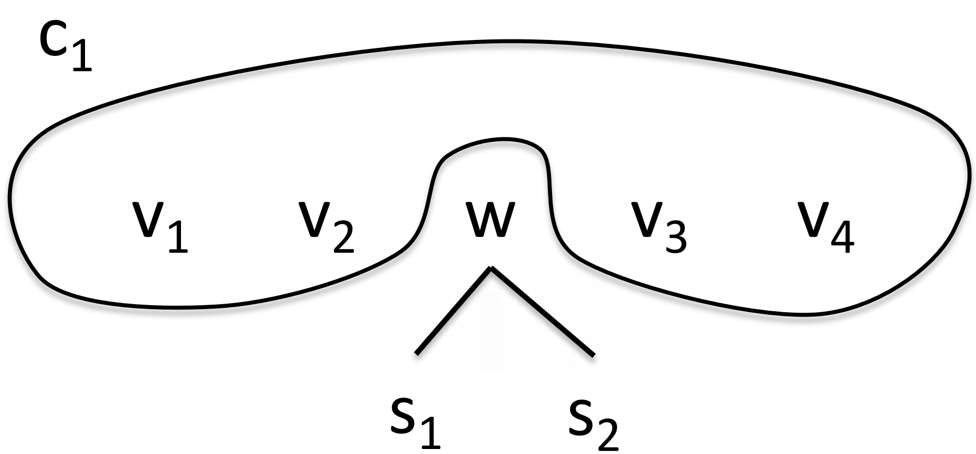
# Notation

- $w$ — an ambiguous word
- $s_1, \ldots, s_k, \ldots s_K$ — senses of the ambiguous word $w$ (sense inventory)
- $c_1, \ldots, c_i, \ldots c_I$ — contexts of $w$ in a corpus
- $v_1, \ldots, v_j, \ldots v_J$ — words used as contextual features for disambiguation

※ Length of context needed for disambiguation

- Verb: local context (argument)
- Noun: broad context

# Notation



# Dictionary-based Disambiguation
## (using sense definitions)

*cone*:

1. a mass of ovule-bearing or pollen-bearing scales or bracts in <u>trees</u> of the pine family or in cycads that are arranged usually on a somewhat elongated axis

2. something that resembles a cone in shape: as … a crisp cone-shaped wafer for holding <u>ice</u> cream

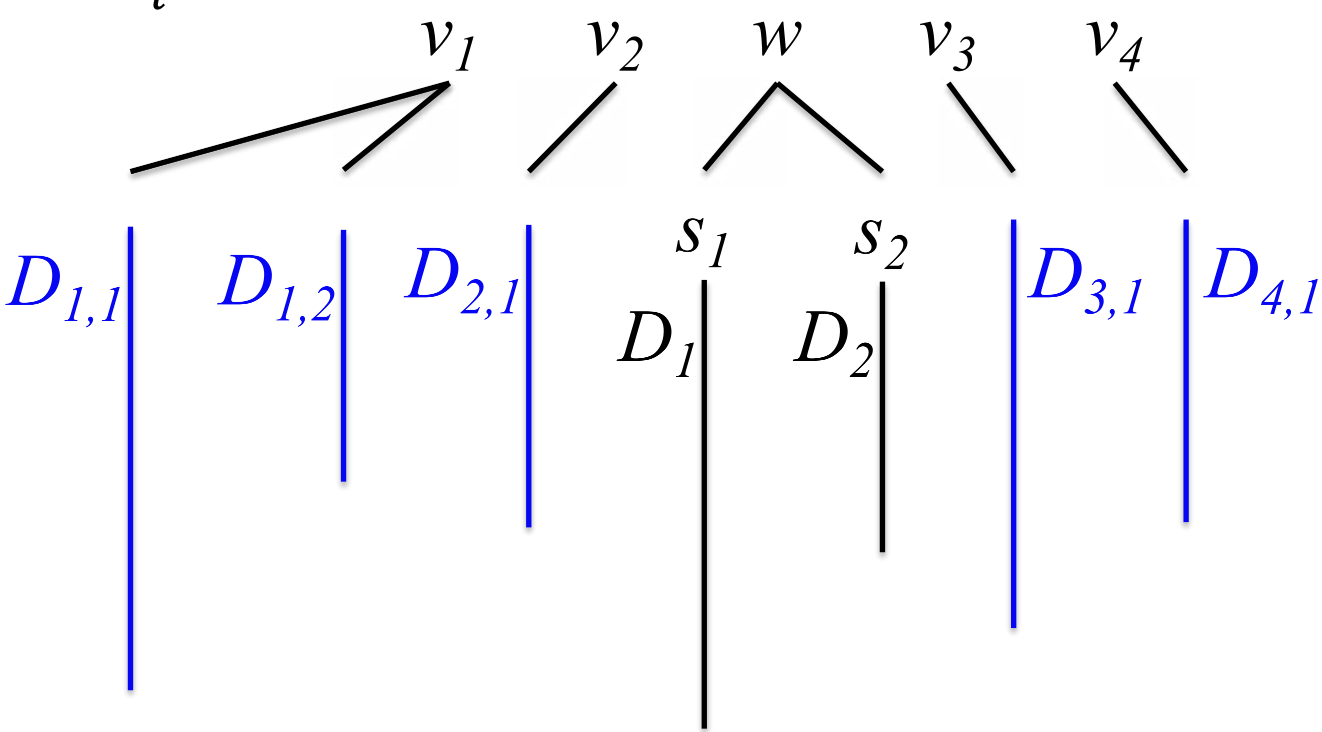
$$s' = \arg\max_{s_k} score(s_k)$$

$$= \arg\max_{s_k} overlap\left(D_k, \bigcup_{v_j \text{ in } c} E_{v_j}\right)$$

$$s' = \arg\max_{s_k} score(s_k)$$

$$= \arg\max_{s_k} overlap\left(D_k, \bigcup_{v_j \text{ in } c} E_{v_j}\right)$$

$$E_{v_j} = \bigcup_i D_{ji}$$



# Dictionary-based Disambiguation
## (using a bilingual dictionary)

- Exploit different translations in other languages
  - German translations of *interest*:
    1. Beteiligung (legal share)
    2. Interesse (attention, concern)
  - "acquire an interest":
    - "erwerben" co-occurs with "Beteiligung"
  - "show interest":
    - "zeigen" co-occurs  with "Interesse"

# One Sense Per Discourse/Collocation

- The dictionary-based methods process each occurrence separately
- However, there are constraints between different occurrences [Yarowsky 1995]
  - One sense per discourse
    - The sense of a target word is highly consistent within any given document
  - One sense per collocation
    - Nearby words provide strong and consistent clues to the sense of a target word

seeds

- $plant_A \rightarrow$ life
- $plant_B \rightarrow$ manufacturing

Tagging (one sense per collocation)

| 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 | ... that divide **life** into *plant* and animal kingdom |
| A | ... many dangers to *plant* and animal **life** ... |
| A | mammals . Animal and *plant* **life** are delicately |
| A | beds too salty to support *plant* **life** . River ... |
| A | heavy seas, damage , and *plant* **life** growing on ... |
| A | ... ... |
| ? | ... vinyl chloride monomer *plant* , which is ... |
| ? | ... molecules found in *plant* and animal tissue |
| ? | ... Nissan car and truck *plant* in Japan is ... |
| ? | ... and Golgi apparatus of *plant* and animal cells ... |
| ? | ... union responses to *plant* closures . ... |
| ? | ... ... |
| ? | ... ... |
| ? | ... cell types found in the *plant* kingdom are ... |
| ? | ... company said the *plant* is still operating ... |
| ? | ... Although thousands of *plant* and animal species |
| ? | ... animal rather than *plant* tissues can be ... |
| ? | ... computer disk drive *plant* located in ... |
| B | ... ... |
| B | automated **manufacturing** *plant* in Fremont ... |
| B | ... vast **manufacturing** *plant* and distribution ... |
| B | chemical **manufacturing** *plant* , producing viscose |
| B | ... keep a **manufacturing** *plant* profitable without |
| 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 ... |

collocation patterns

**Initial decision list for *plant* (abbreviated)**

| LogL | Collocation | Sense |
|---|---|---|
| 8.10 | *plant* **life** | $\Rightarrow$ A |
| 7.58 | **manufacturing** *plant* | $\Rightarrow$ B |
| 7.39 | **life** (within $\pm$2-10 words) | $\Rightarrow$ A |
| 7.20 | **manufacturing** (in $\pm$2-10 words) | $\Rightarrow$ B |
| 6.27 | animal (within $\pm$2-10 words) | $\Rightarrow$ A |
| 4.70 | equipment (within $\pm$2-10 words) | $\Rightarrow$ B |
| 4.39 | employee (within $\pm$2-10 words) | $\Rightarrow$ B |
| 4.30 | assembly *plant* | $\Rightarrow$ B |
| 4.10 | *plant* closure | $\Rightarrow$ B |
| 3.52 | *plant* species | $\Rightarrow$ A |
| 3.48 | automate (within $\pm$2-10 words) | $\Rightarrow$ B |
| 3.45 | microscopic *plant* | $\Rightarrow$ A |
| | ... | |

Assign the majority sense in a document/discourse to ambiguous words (one sense per discourse)

# Unsupervised Disambiguation

- In the case of no knowledge sources
- Clustering

- EM (Expectation Maximization)
  1. Initialization
  2. Expectation (E-step)
  3. Maximization (M-step)

# Observed Data

context words

money ... *bank* ... lend

borrow ... *bank* ... lend

borrow ... *bank* ... money

river ... *bank* ... river

c is
overall context.
constant, can
ignore denom

$$\arg\max_{s_k} P(s_k \mid c) = \arg\max_{s_k} \frac{P(c \mid s_k)}{P(c)} \cdot P(s_k)$$

Naive Bayes
assumption

$$= \arg\max_{s_k} P(c \mid s_k) \cdot P(s_k)$$

$$= \arg\max_{s_k} \prod_{v_j \text{in} c} P(v_j \mid s_k) \cdot P(s_k)$$

# Initialization

$P(s1) = 0.5$

$P(s2) = 0.5$

| | |
|---|---|
| $P(money\|s1)=0.27$ | $P(money\|s2)=0.26$ |
| $P(borrow\|s1)=0.24$ | $P(borrow\|s2)=0.28$ |
| $P(lend\|s1)=0.26$ | $P(lend\|s2)=0.22$ |
| $P(river\|s1)=0.23$ | $P(river\|s2)=0.24$ |

# E-step

- Estimation of complete data

Use Naive Bayes assumption

$$P(s_k \mid c_i) = \frac{P(s_k, c_i)}{P(c_i)} = \frac{P(s_k, c_i)}{\sum_j P(s_j, c_i)} = \frac{P(s_k)P(c_i \mid s_k)}{\sum_j P(s_j)P(c_i \mid s_j)}$$

P(s1| money … *bank* … lend) = ?

P(s2| money … *bank* … lend) = ?

…

# E-step

naive Bayes \Pi P(vj l sk) P (sk)

P(s1 | money … *bank* … lend)

$$= \frac{0.5 \times \boxed{0.27 \times 0.26}}{0.5 \times 0.27 \times 0.26 + 0.5 \times 0.26 \times 0.22} = 0.55$$

P(s2 | money … *bank* … lend) $= 0.45$

# Result of E-step

money … *bank* … lend
$s_1$:0.55, $s_2$:0.45

borrow … *bank* … lend
$s_1$:0.50, $s_2$:0.50

borrow … *bank* … money
$s_1$:0.47, $s_2$:0.53

river … *bank* … river
$s_1$:0.48, $s_2$:0.52

# M-step

- Maximum likelihood estimation from complete data

P(s1) = ?

P(money | s1) = ?

…

# M-step

- Maximum likelihood estimation from complete data

mean of different occurrences of s1

$$P(s1) = \frac{0.55+0.50+0.47+0.48}{4} = 0.50$$

$$P(money \mid s1) = \frac{0.55+0.47}{\underline{2\times(0.55+0.50+0.47+0.48)}} = 0.255$$

...

2 is for 2
words in a given context

P(s1) = \sum_i P(s1, ci) =
P(s1, money) + P(s1, lend) +
P(s1, borrow) + … + P(s1,
river) + P(s1, river)

# Iteration

- Iterate E-step and M-step until convergence

| | |
|---|---|
| P(s1) = 0.75 | |
| P(s2) = 0.25 | |
| P(money\|s1)=0.33 | P(money\|s2)=0.00 |
| P(borrow\|s1)=0.33 | P(borrow\|s2)=0.00 |
| P(lend\|s1)=0.33 | P(lend\|s2)=0.00 |
| P(river\|s1)=0.00 | P(river\|s2)=1.00 |

# Supervised Disambiguation

- Disambiguated corpora for training
  - SemCor
    - 200K words in Brown Corpus were manually annotated with a sense tag (synset) of WordNet
      - e.g., Grabbing his Winchester from its sheath, Cook prepared to fight from behind the arroyo <u>bank</u>.

09213434-n:
a long ridge or pile

  - Wikipedia
    - Internal links in Wikipedia can be regarded as manually disambiguated sense tags

The signing of basketball player <u>Michael Jordan</u> in 1984, with his subsequent promotion of Nike over the course of his career …

http://en.wikipedia.org/wiki/Nike,_Inc.      http://en.wikipedia.org/wiki/Michael_Jordan

# Supervised Disambiguation

- Bayesian classification

$$\arg\max_{s_k} P(s_k \mid c)$$

$$= \arg\max_{s_k} \frac{P(c \mid s_k)}{P(c)} \cdot P(s_k)$$

$$= \arg\max_{s_k} P(c \mid s_k) \cdot P(s_k)$$

$$= \arg\max_{s_k} \prod_{v_j \text{in} c} P(v_j \mid s_k) \cdot P(s_k)$$

$$\frac{C(v_j, s_k)}{\sum_t C(v_t, s_k)} \qquad \frac{C(s_k)}{C(w)}$$

count of sk / count of w

- Other machine learning methods

svm, NN

# Wikification

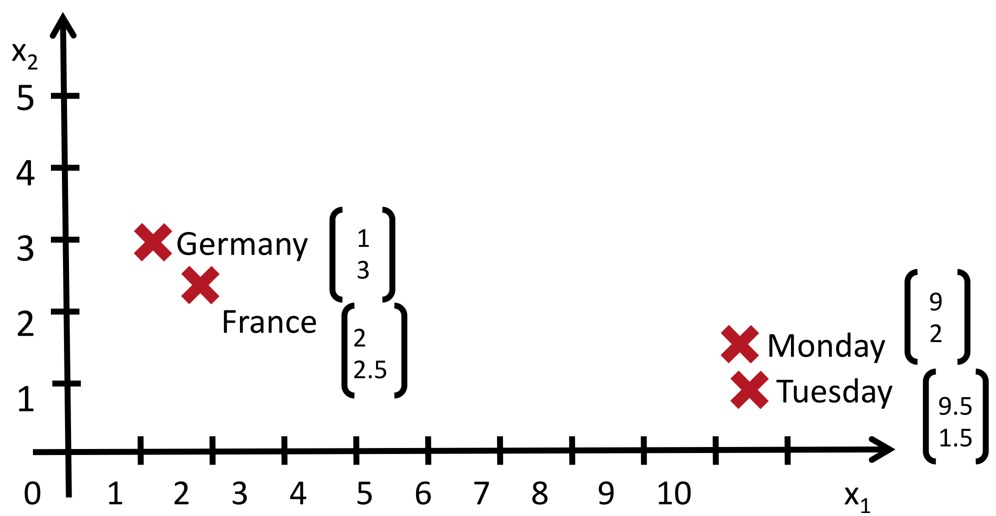
On Saturday, Michael Jordan and Tom Brady played a game of pickup basketball in the Bahamas.

http://en.wikipedia.org/wiki/Michael_Jordan


http://en.wikipedia.org/wiki/Michael_I._Jordan

# Word2vec [Mikolov+ 2013]

- Learning (dense) word vectors using a neural network
- Based on the distributional hypothesis

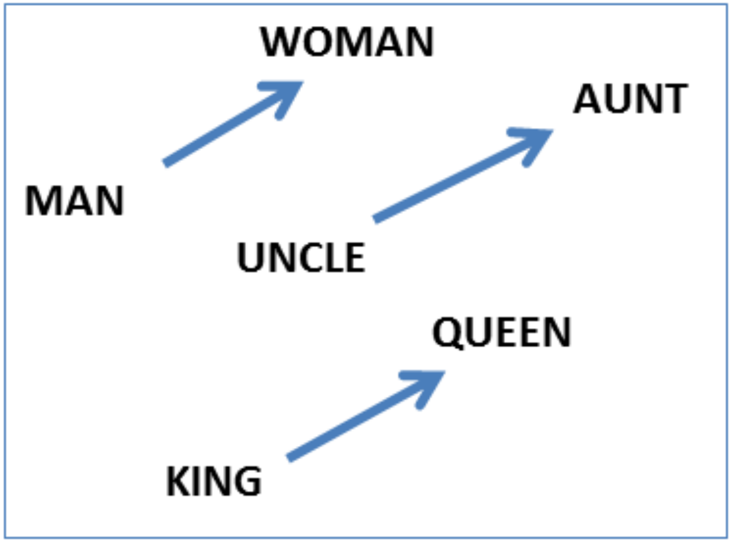


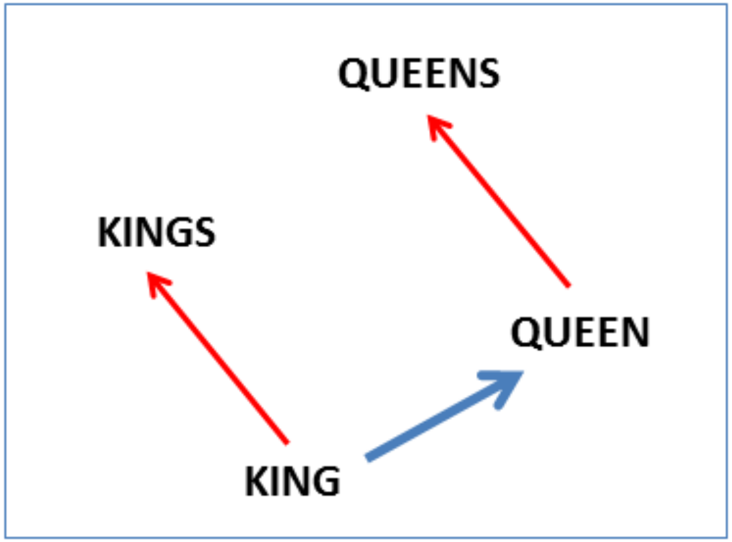
# Linguistic Regularities

[Mikolov+ 2013]

- KINGS – KING + QUEEN = QUEENS

distributional hypothesis



# Linguistic Regularities

| *Expression* | *Nearest token* |
|---|---|
| Paris - France + Italy | Rome |
| bigger - big + cold | colder |
| sushi - Japan + Germany | bratwurst |
| Cu - copper + gold | Au |
| Windows - Microsoft + Google | Android |
| Montreal Canadiens - Montreal + Toronto | Toronto Maple Leafs |

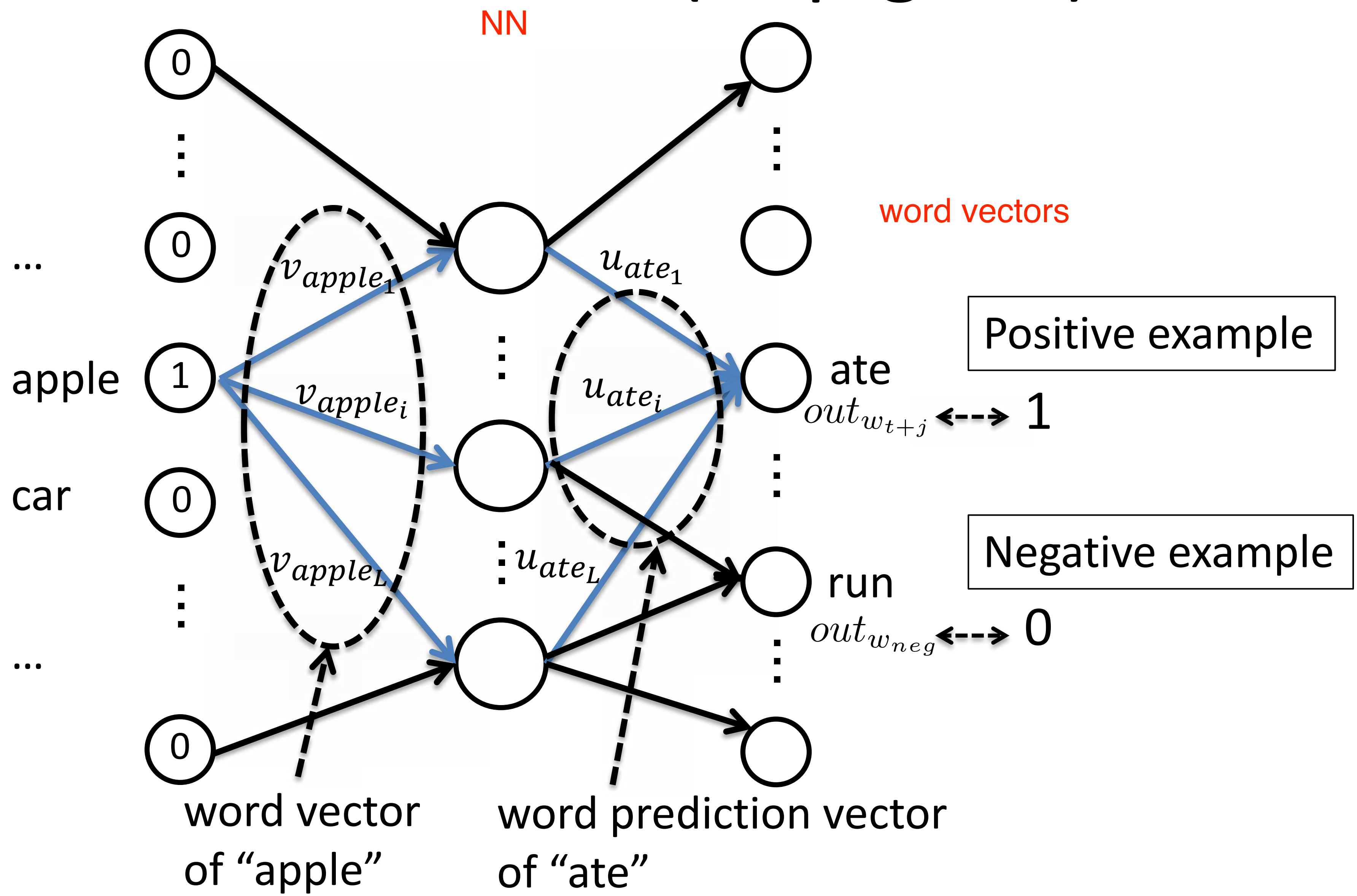
| *Expression* | *Nearest tokens* |
|---|---|
| Czech + currency | koruna, Czech crown, Polish zloty, CTK |
| Vietnam + capital | Hanoi, Ho Chi Minh City, Viet Nam, Vietnamese |
| German + airlines | airline Lufthansa, carrier Lufthansa, flag carrier Lufthansa |
| Russian + river | Moscow, Volga River, upriver, Russia |
| French + actress | Juliette Binoche, Vanessa Paradis, Charlotte Gainsbourg |

# Word2vec (Skip-gram)

$p(w_{t-2}|w_t)$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\cdots$ | he | ate | an | apple | yesterday | and | $\cdots$ |
| | $w_{t-3}$ | $w_{t-2}$ | $w_{t-1}$ | $w_t$ | $w_{t+1}$ | $w_{t+2}$ | |
| Pseudo negative examples | friend | run | company | | ink | mountain | |
| | play | flow | desk | | say | run | |
| | feature | center | human | | light | swallow | |
| | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ | |

# Word2vec (Skip-gram)



# Problems of Word2vec

- Sense ambiguities
  - One vector is defined for a word
  
  → Contextualized embeddings
    - ELMo [Peters+ 2018], BERT [Devlin+ 2019]
- Out-of-vocabulary words

  → Use of subwords
- A vector for a phrase or a document?
- Antonyms tend to have similar vectors

# Summary

- What is word sense?
- Synonym / Homonym / Polysemic words
- Word sense disambiguation
  - Dictionary-based disambiguation
  - Unsupervised disambiguation
  - Supervised disambiguation