ECE365: Introduction to NLP

Spring 2021

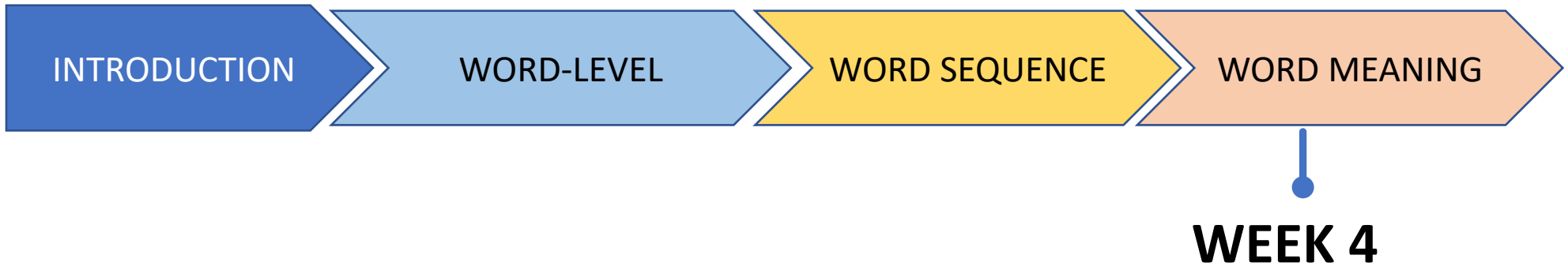
Lecture 6: Lexical Semantics

[Reading J&M 19.1, 19.2, 19.3, 6.1]

Logistics

- Quiz 2 a week from today (05/04)
- Level of difficulty similar to Quiz 1
- Labs 3 and 4 are released; open to early submissions

Course Progress



How do we model meaning of words?

The Story Thus Far

- Words
- Text Classification using words
- Language modeling
- Sequence labeling

What is Lexical Semantics?

What is semantics?

Connects language to real world

What is lexical?

Vocabulary Lexicon

Words

- Types and Tokens
- Morphology
- Sense and meaning

Word sense ambiguity

Word sense ambiguity

- Iraqi head seeks arms
- Drunk gets nine years in violin case

Word sense ambiguity

- Many words have multiple meanings

Terminology

Lemma: Base form (dictionary form) of a word

banks	bank
sung	sing
duermes	dormir

Lemmas have senses

one lemma *bank* can have many meanings:
...a $bank_1$ can hold the investments in a custodial account
...as agriculture burgeons on the east $bank_2$ the river will shrink even more

sense (or **word sense**)
– a discrete representation of an aspect of a word's meaning
Word sense disambiguation (WSD): task of determining sense of word in given context

Why disambiguate word sense?

information retrieval

- query: *bat care* *Animal? Equipment?*

machine translation

- *bat*: *murciélago* (animal) or *bate* (for baseball)

text-to-speech

- *bass* (stringed instrument) vs. *bass* (fish)

Resource for word senses

- WordNet
  - Large online database of word senses and relation between word senses
  - Available in many languages
    - Arabic, Afrikaans, Chinese, English, French, German, Hindi and other languages

Bass on WordNet

Noun

- S: (n) **bass** (the lowest part of the musical range)
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
- **S: (n) bass, basso (an adult male singer with the lowest voice)**
- S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- S: (adj) **bass**, deep (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

“sense” defined in WordNet

- **synset** (**synonym set**): set of near-synonyms
  - instantiates a sense or concept, with **gloss**
  - example: *chump* as a noun with gloss: *“a person who is gullible and easy to take advantage of”*
- this sense of *chump* is shared by 9 words:

chump[1], *fool*[2], *gull*[1], *mark*[9], *patsy*[1], *fall guy*[1], *sucker*[1], *soft touch*[1], *mug*[2]

- each of **these** senses have this same gloss – (not **every** sense; sense 2 of *gull* is the aquatic bird)

Semantic Relations

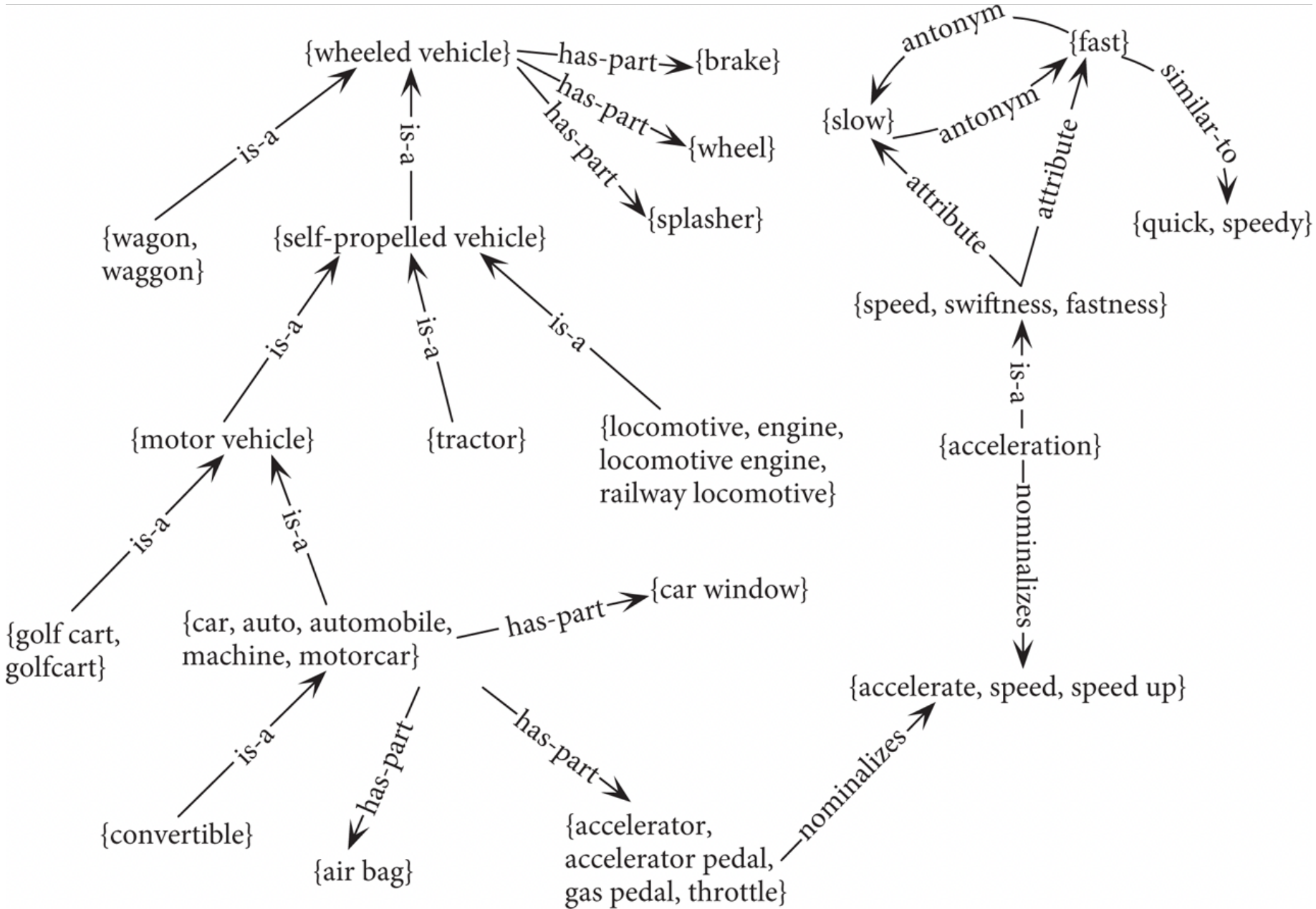
- Synonymy (equivalence)
  - *filbert / hazelnut*, *couch / sofa*, *big / large*

- Antonymy (opposition)

  fast/slow

- Hyponymy – Hypernymy (subset of – superset of)
  - Mango-fruit, breakfast-meal

- Meronymy – Holonymy (part of – has a)
  - Leg-table, course-meal


{wheeled vehicle}
has-part
{brake}
has-part
{wheel}
has-part
{splasher}
is-a
is-a
{wagon, waggon}
{self-propelled vehicle}
is-a
is-a
is-a
{motor vehicle}
{tractor}
{locomotive, engine, locomotive engine, railway locomotive}
is-a
is-a
{golf cart, golfcart}
{car, auto, automobile, machine, motorcar}
has-part
{car window}
is-a
has-part
has-part
{convertible}
{air bag}
{accelerator, accelerator pedal, gas pedal, throttle}
nominalizes
antonym
{fast}
similar-to
{slow}
antonym
attribute
attribute
{quick, speedy}
{speed, swiftness, fastness}
is-a
{acceleration}
nominalizes
{accelerate, speed, speed up}

WordNet Interfaces

- Various interfaces to WordNet are available

– Many languages listed at
https://wordnet.princeton.edu/related-projects

– NLTK (Python)

```
>>> from nltk.corpus import wordnet as wn
>>> wn.synsets('dog') (returns list of Synset objects)
```

http://www.nltk.org/howto/wordnet.html

Limitations

- Intrinsic limits to this type of resource:
  - Many years of manual effort by skilled lexicographers
  - In the case of WordNet, some of the lexicographers were not that skilled, and this has led to inconsistencies
  - The ontology is only as good as the ontologist(s); not data-driven

- We will now look at an approach to lexical semantics that is data driven and does not rely on lexicographers

Distributional Semantics

- What is *tesgüino*?

(a) A bottle of tesgüino is on the table

(b) People like tesgüino.

(c) Don't have tesgüino before you drive.

(d) Tesgüino is made out of corn

Distributional Hypothesis

(C1) A bottle of tesgüino is on the table

(C2) People like tesgüino.

(C3) Don't have tesgüino before you drive.

(C4) Tesgüino is made out of corn

	C1	C2	C3	C4
tesgüino	**1**	**1**	**1**	**1**
loud				
Motor oil				
tortillas				
choices				
wine				

Distributional Hypothesis

	C1	C2	C3	C4
tesgüino	**1**	**1**	**1**	**1**
loud	0	0	0	0
Motor oil	1	0	0	1
tortillas	0	1	0	1
choices	0	1	0	0
wine	1	1	1	1

Distributional hypothesis, stated by linguist John R. Firth (1957) as:

"You shall know a word by the company it keeps."

≈ "words that occur in similar contexts tend to have similar meanings"

One of the most successful ideas of modern statistical NLP!

TESGÜINO, UNA BEBIDA RITUAL DE MAÍZ DE LOS RARÁMURIS

¿QUÉ COMER?

COCINA MEXICANA

28 JUL 2015

1

Distributional Semantic Models

- Distributional statistics are important in NLP
  - they help data-driven approaches learn about rare words that do not appear in labeled training data
  - no complex annotation needed

- Vector semantics = {distributional idea (defining a word by counting what other words occur in its environment) } + {meaning of a word as a vector (a point in N-dimensions)}

- Popularly called word embeddings
  - Various versions depending on how the vector components are computed
  - Latent Semantic Analysis, Word2vec, GloVe

Why model words as vectors?

- We need to model word meaning
- As a way for computing similarity between words
- Useful for unknown words

Question Answering
Q: How tall is Mt. Everest?
Answer candidate: The height of Mt. Everest is 29029 feet

Distributionally Similar Words

Rum	Write	Ancient	Mathematics
vodka	read	old	physics
cognac	speak	modern	biology
brandy	present	traditional	geology
whisky	receive	medieval	sociology
liquor	call	historic	psychology
detergent	release	famous	anthropology
cola	sign	original	astronomy
gin	offer	entire	arithmetic
lemonade	know	main	geography
cocoa	accept	indian	theology
chocolate	decide	various	hebrew
scotch	issue	single	economics
noodle	prepare	african	chemistry
tequila	consider	japanese	scripture
juice	publish	giant	biotechnology

Figure 6.1 A two-dimensional (t-SNE) projection of embeddings for some words and phrases, showing that words with similar meanings are nearby in space. The original 60-dimensional embeddings were trained for sentiment analysis. Simplified from Li et al. (2015).

Why not use a thesaurus?

- We don’t have a thesaurus for every language
- We can’t have a thesaurus for every year

Summary

- Words can be used in different senses
- There are curated sense databases of words manually created
- Semantic relations between words
- Distributional semantics