heig-vd
HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Dictionnaires sémantiques : WordNet et la désambiguïsation lexicale

(based on Ch. 17 of *Speech and Language Processing*,
by D. Jurafsky & J. H. Martin, 3rd ed. draft & slides,
https://web.stanford.edu/~jurafsky/slp3/)

Andrei Popescu-Belis

IICT | TIC | HEIG-VD

Cours 10 – 30 avril 2019

heig-vd

Plan of this course (#10)

1. Word senses and their relations
2. The WordNet lexical database
3. Word Sense Disambiguation (WSD)
   - using distance in WordNet
   - using the Lesk Algorithm & WordNet definitions
   - using supervised machine learning
   - using word2vec → Lab 5

1. WORD SENSES AND THEIR RELATIONS

heig-vd

Words have senses

- One word can have many meanings
- Example: *bank* has here two different senses

  A bank can hold the investments in a custodial account.

  As agriculture burgeons on the east bank the river will shrink even more.

- Sense of a word (lexical meaning)
  - discrete representation of an aspect of a word's meaning

heig-vd

Homonymy

- Homonyms: words that share a (written and spoken) form but have unrelated, distinct meanings:
  - *$bank_1$: financial institution* *$bank_2$: sloping land*
  - *bat_1: club for hitting a ball* *bat_2: nocturnal flying mammal*
- Homographs: share the written form (bank, bat) but not necessarily the spoken one (record)
- Homophones: share the spoken form (but in general not the written one)
  - write and right
  - piece and peace

heig-vd

Problems for NLP applications

- Information retrieval (query)
  - "bat care"
- Machine Translation (EN → ES)
  - bat → murciélago (animal) | bate (for baseball)
- Text-to-Speech (*homographs*)
  - bass (stringed instrument) vs. bass (fish)

heig-vd

Polysemy

- Polysemous word: a word with multiple related meanings
  - if meanings are not related, we have two homonyms
  - **most non-rare words have multiple meanings**
    - check (in French) this dictionary: http://www.cnrtl.fr/definition/

- Example

  1. The bank was constructed in 1875 out of local red brick.
  2. I withdrew the money from the bank

  Are those the same sense? No.

    - sense 2 = “A financial institution”
    - sense 1 = “The building hosting a financial institution”

  Typical cases: organization/building, work/author, etc.

heig-vd

Synonyms

- Word that have the same meaning in some or all contexts
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - water / H_20
- Two words are synonyms
  - if they can be substituted for each other in all situations
  - i.e. they have the same propositional meaning
- But there are few (or no) examples of perfect synonymy
  - even if many aspects of meaning are identical
  - still may not preserve the acceptability based on politeness, register, genre, …
- Examples
  - water/H_20
  - big/large
  - brave/courageous

heig-vd

Synonyms and antonyms

- Synonymy: identity between senses rather than words
- Example: *big* vs. *large*
  - synonyms or not?

    How **big** is that plane?

    Would I be flying on a **large** plane?

    Miss Nelson became a kind of **big** sister to Benjamin.

    ?Miss Nelson became a kind of **large** sister to Benjamin.
  - *big* has one sense that means older, or grown up, but *large* lacks it

- Antonyms: senses that are opposites with respect to one feature of meaning
  - otherwise, very similar
  - dark/light, short/long, fast/slow, rise/fall, hot/cold, up/down
  - can define a binary opposition, or be at opposite ends of a scale (fast/slow) or be reversives (rise/fall)

heig-vd

Hyponyms and hypernyms

- One sense is a hyponym of another if the first sense is more specific, denoting a subclass of the other
  - car is a hyponym of vehicle
  - mango is a hyponym of fruit
- Conversely: the hypernym or superordinate
  - vehicle is a hypernym of car
  - fruit is a hypernym of mango
- Other equivalent perspectives
  - the class (set of entities) denoted by the superordinate includes the class denoted by the hyponym
  - a sense *X* is a hyponym of sense *Y* if *being an X* **entails** *being a Y*
    - X is-a Y or X ISA Y → hierarchy, because hyponymy is usually transitive

heig-vd

Exercice

Quelles sont les relations de sens entre les mots des deux colonnes ?

- clair
- primevère
- mort
- animal
- ouvert
- granit
- grand
- véhicule

- minéral
- vivant
- caniche
- fermé
- vélo
- limpide
- fleur
- petit

2. WORDNET: A LEXICAL DATABASE

heig-vd

WordNet 3.0

- A hierarchically organized **English lexical database**
  - most of it completed in the 1990s, but still updated
  - in electronic form: online queries or download
  - many software libraries offer access to it

- Online thesaurus + aspects of a dictionary
  - many other languages available or under development
  - en.wikipedia.org/wiki/WordNet

Category	Word types
Noun	117,798
Verb	11,529
Adjective	22,479
Adverb	4,481

heig-vd

WordNet 3.0

- Where it is
  - homepage: https://wordnet.princeton.edu/
  - online search interface: http://wordnetweb.princeton.edu/perl/webwn
- Libraries
  - Python
    - WordNet from NLTK: http://www.nltk.org/Home
  - Java
    - JWNL (Java WordNet Library), extJWNL (extended JWNL)

heig-vd

Senses of “bass” in 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"*

heig-vd

How are senses defined in WordNet?

- Synsets = synonym sets
  - sets of near-synonyms, instantiating one sense or concept, with a gloss (= definition) and sometimes examples
- Example of a synset
  - chump (noun) has one sense, with the gloss: "a person who is gullible and easy to take advantage of"
  - this sense of chump is shared by senses of 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, marked with superscripts, form a single synset with a single gloss
  - but, for instance gull[2] is the *aquatic bird*, another synset

heig-vd

WordNet synsets form a hierarchy e.g., hypernym hierarchy for “bass”

- S: (n) **bass**, basso (an adult male singer with the lowest voice)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) singer, vocalist, vocalizer, vocaliser (a person who sings)
      - S: (n) musician, instrumentalist, player (someone who plays a musical instrument (as a profession))
        - S: (n) performer, performing artist (an entertainer who performs a dramatic or musical work for an audience)
          - S: (n) entertainer (a person who tries to please or amuse)
            - S: (n) person, individual, someone, somebody, mortal, soul (a human being) *"there was too much for one person to do"*
              - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
                - S: (n) living thing, animate thing (a living (or once living) entity)
                  - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                    - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                      - S: (n) physical entity (an entity that has physical existence)
                        - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

heig-vd

A variety of relations between synsets are encoded in WordNet

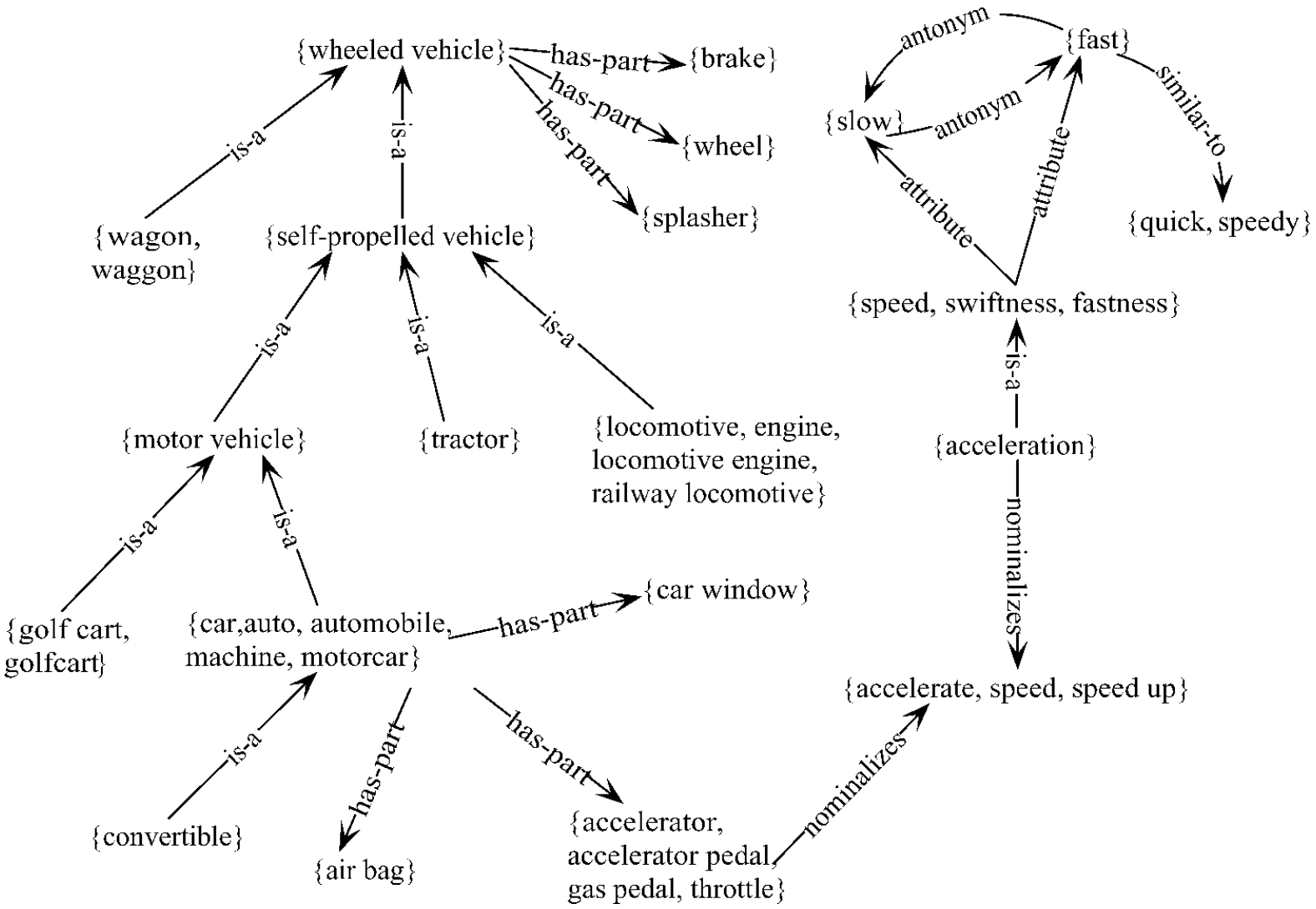
- Example for synsets of nouns

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$breakfast^1 \rightarrow meal^1$
Hyponym	Subordinate	From concepts to subtypes	$meal^1 \rightarrow lunch^1$
Instance Hypernym	Instance	From instances to their concepts	$Austen^1 \rightarrow author^1$
Instance Hyponym	Has-Instance	From concepts to concept instances	$composer^1 \rightarrow Bach^1$
Member Meronym	Has-Member	From groups to their members	$faculty^2 \rightarrow professor^1$
Member Holonym	Member-Of	From members to their groups	$copilot^1 \rightarrow crew^1$
Part Meronym	Has-Part	From wholes to parts	$table^2 \rightarrow leg^3$
Part Holonym	Part-Of	From parts to wholes	$course^7 \rightarrow meal^1$
Substance Meronym		From substances to their subparts	$water^1 \rightarrow oxygen^1$
Substance Holonym		From parts of substances to wholes	$gin^1 \rightarrow martini^1$
Antonym		Semantic opposition between lemmas	$leader^1 \Longleftrightarrow follower^1$
Derivationally Related Form		Lemmas w/same morphological root	$destruction^1 \Longleftrightarrow destroy^1$

- There are fewer ones for verbs

heig-vd

Relations in WordNet can also be viewed as a graph



heig-vd

Using WordNet for word similarity

- Similarity is properly a relation between senses
  - the word "bank" is not obviously similar to the word "slope", but the sense bank^{2} is similar to slope^{5}

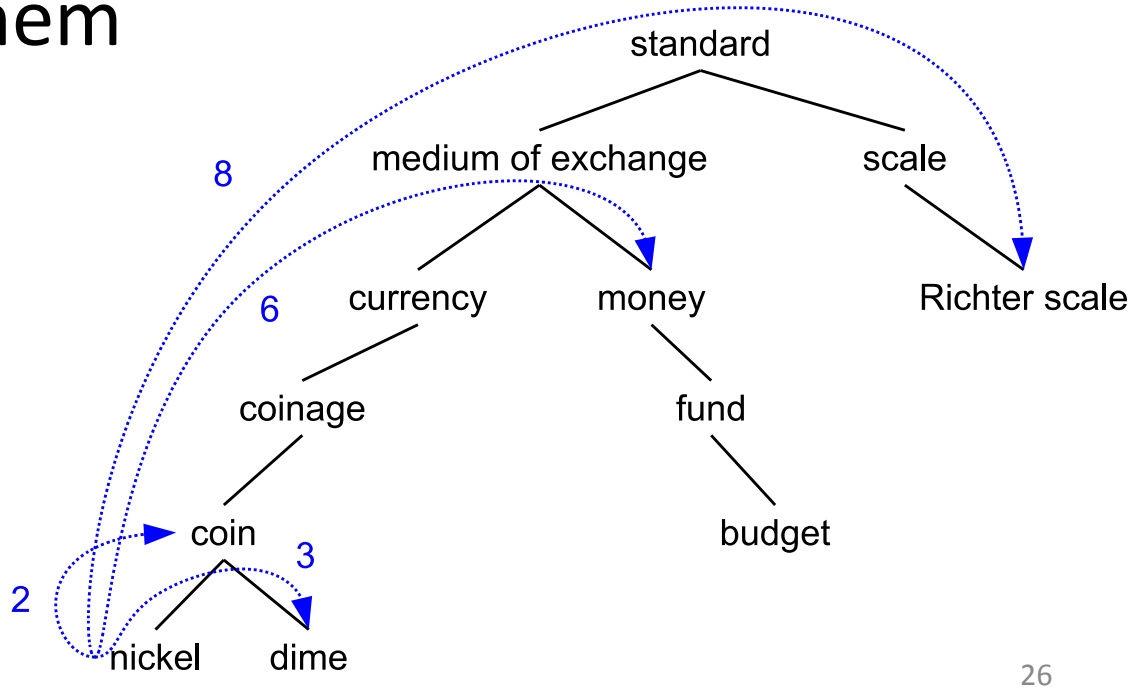
- Similarity algorithms assume that similarity between two words is given by the similarity of (1) either their closest senses, or (2) their most frequent senses

- Two classes of similarity algorithms
  1. **Thesaurus-based algorithms**: are words/senses close in hypernym hierarchy? Do words/senses have similar glosses (definitions)?
  2. **Distributional algorithms**: do words have similar distributional contexts? Are their low-dimensional embeddings close?

heig-vd

Example: path-based similarity

- Two concepts (senses/synsets) are similar if they are near each other in the thesaurus hierarchy, i.e. they have a short path between them



3. WORD SENSE DISAMBIGUATION

heig-vd

Word Sense Disambiguation (WSD)

- Given
  - a word in context (sentence, text, dialogue, …)
  - a fixed inventory of potential word senses
  - ➢ decide which sense of the word this is
- Why? Machine translation, QA, speech synthesis, …
- What set of senses?
  - MT, e.g. English-to-Spanish: set of Spanish translations
  - speech synthesis: homographs like *bass* and *bow*
  - in general: the senses in a thesaurus like **WordNet**

heig-vd

Two variants of the WSD task

1. Lexical sample task
   - disambiguate a small pre-selected set of target words
     - *line*, *plant*, *interest*, etc.
   - given an inventory of senses for each word
   - example of solution: use supervised machine learning to train a sense classifier for each word

2. All-words task
   - disambiguate every word in an entire text
   - given a lexicon with senses for each word
   - data sparseness: hard to train word-specific classifiers

heig-vd

WSD evaluation

- Best evaluation: extrinsic (‘end-to-end’, `task-based’) evaluation
  - embed WSD algorithm in a task and see if it improves the task
- What we often do for convenience: intrinsic evaluation
  - exact match / sense accuracy
    - % of words tagged identically with the human-manual sense tags
  - usually evaluated using held-out data from same labeled corpus
- Baselines
  - most frequent sense
  - the Lesk algorithm

heig-vd

Most frequent sense

- WordNet senses are ordered in frequency order
  - the most frequent sense in WordNet is the first one
  - these frequencies come from the *SemCor* corpus

Freq	Synset	Gloss
338	plant[1], works, industrial plant	buildings for carrying on industrial labor
207	plant[2], flora, plant life	a living organism lacking the power of locomotion
2	plant[3]	something planted secretly for discovery by another
0	plant[4]	an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

heig-vd

The simplified Lesk algorithm

- Let's disambiguate "bank" in this sentence

  The bank can guarantee that deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

- Given the following two WordNet senses

bank[1]	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
bank[2]	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

heig-vd

The simplified Lesk algorithm

Choose sense with most word overlap between gloss and context (not counting function words)

The bank can guarantee that deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

bank[1]	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
bank[2]	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

heig-vd

WSD using machine learning: building feature vectors

- Observation = instance of a target word
- A simple representation for each observation
  - vectors of sets of feature/value pairs
  - represented as a ordered list of values
  - they represent, e.g., the window of words around the target

- Collocational features and bag-of-words features
  - Collocational**:** Features about words at **specific** positions near target word: Often limited to just word identity and POS
  - Bag-of-words: Features about words that occur anywhere in the window (regardless of position): Typically limited to frequency counts

heig-vd

Example: collocational features

- Example text (WSJ):

  "An electric guitar and bass player stand off to one side not really part of the scene"

  – assume a window of +/- 2 from the target, -stopwords

- Position-specific information about the words and collocations in window

$$[w_{i-2}, \mathrm{POS}_{i-2}, w_{i-1}, \mathrm{POS}_{i-1}, w_{i+1}, \mathrm{POS}_{i+1}, w_{i+2}, \mathrm{POS}_{i+2}, w_{i-2}^{i-1}, w_{i}^{i+1}]$$

```
[guitar, NN, and, CC, player, NN, stand, VB, and guitar, player stand]
```

- Word 1,2,3 grams in window of ±3 is common

heig-vd

Bag-of-words features

- Unordered set of words = position is ignored
- Counts of words occurring within the window

- First choose a vocabulary
- Then count how often each of those terms occurs in a given window, or sometimes just use a binary “indicator” 1 or 0

heig-vd

Classification Methods: Supervised Machine Learning

- *Input*
  - a word *w in a text window d ("document")*
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - a training set of *m* hand-labeled text windows $(d_1,c_1),\ldots,(d_m,c_m)$
- *Output:*
  - a learned classifier $\gamma : d \rightarrow c$
- *Method*
  - any kind of classifier: Naive Bayes, Logistic regression, Neural Networks, Support-vector machines, k-Nearest Neighbors

Conclusion

- Capturing the sense of a text is often a matter of capturing the senses of its words
  - enables tasks such as classification, but also question answering
- Distributional semantics is powerful
  - but not ideal for dealing with WSD
  - use dictionary based or supervised ML
- Lab 6 : use word2vec for WSD