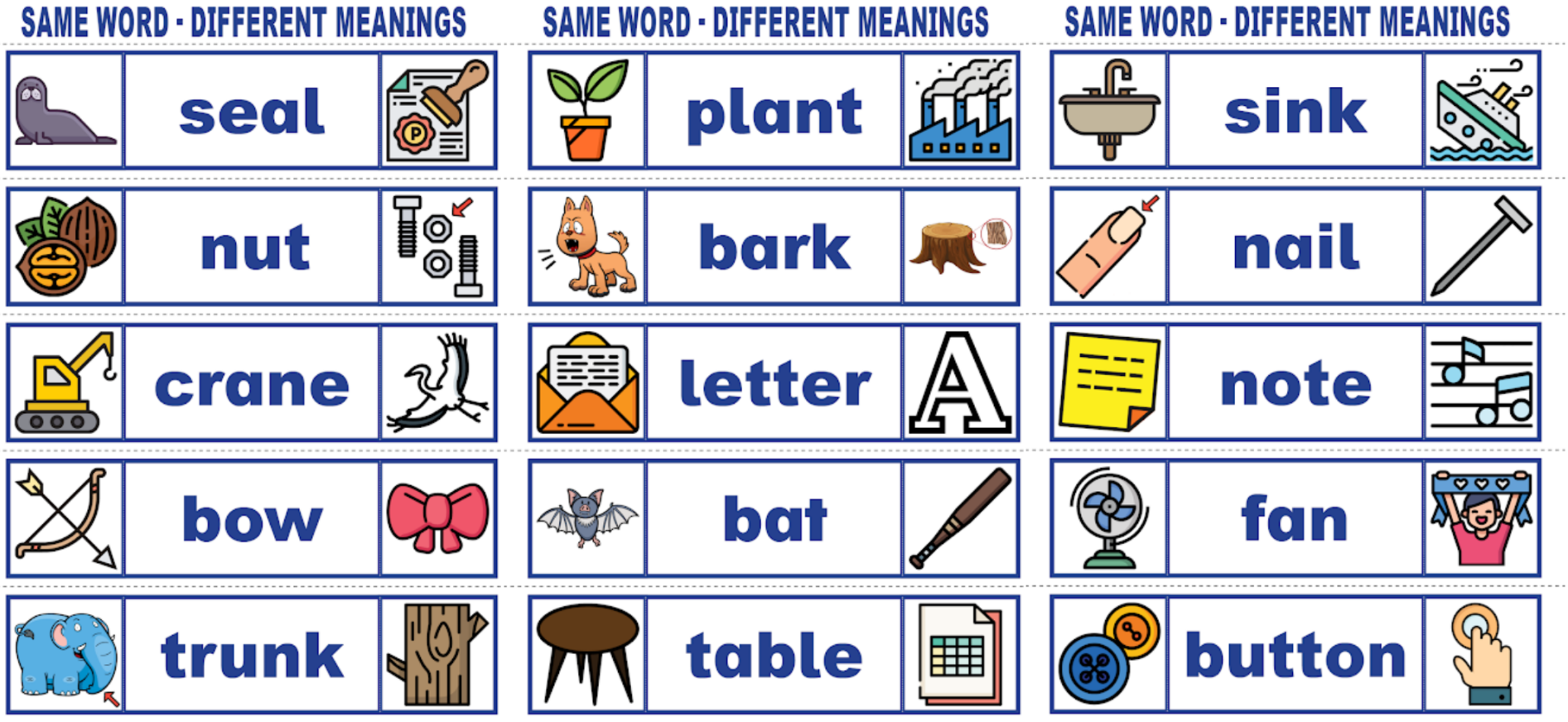
Concepts and Applications in NLP

Word Senses and WordNet

Marion Di Marco

December 17, 2024

Challenges in NLP: Word Senses



Illustrations from https://www.englishforkidz.com/2020/01/vocabulary-cards-word-different-meanings.html

Outline

Word Senses

- **Word sense**:
  discrete representation of one aspect of the meaning of a word
- Context helps to differentiate the meanings:
- $mouse^1$: ... a mouse controlling a computer system in 1968.
  $mouse^2$: ... a quiet animal like a mouse.
- $bank^1$: ... a bank can hold the investments in a custodial account ...
  $bank^2$: ... as agriculture burgeons on the east bank, the river ...

Defining Word Senses

- Embeddings: define the meaning of a word by its co-occurrences
- Contextual embeddings: embedding that represents the meaning of a word in its textual contex
- How to define the meaning of a a word *sense*?
- Textual definitions for each sense (glosses) as given in dictionaries or thesauri
- Glosses for *bank*:
  1. `financial institution that accepts deposits and channels the money into lending activities`
  2. `sloping land (especially the slope beside a body of water)`

Glosses

- Glosses are not a formal meaning representation
- Examples from the definitions for *right*, *left*, *red*, *blood*

right	*adj.* located nearer the right hand esp. being on the right when facing the same direction as the observer.
left	*adj.* located nearer to this side of the body than the right.
red	*n.* the color of blood or a ruby.
blood	*n.* the red liquid that circulates in the heart, arteries and veins of animals.

- Circularity in definitions (e.g. *right*)
  red and *blood* reference each other
- Useful for humans → sufficient grasp of the other terms
- Still useful for computational modeling of senses
  - use sentence to build embeddings
  - make use of *sense relations*

Sense Relations

right	*adj.* located nearer the right hand esp. being on the right when facing the same direction as the observer.
left	*adj.* located nearer to this side of the body than the right.
red	*n.* the color of blood or a ruby.
blood	*n.* the red liquid that circulates in the heart, arteries and veins of animals.

  - *left – right*: similar words that stand in alternation or contrast
  - *red*: color
  - *blood*: liquid

- Sense relations: "*is-a*" relation or antonymy
  listed in databases like WordNet

- Given a large database of relations → allows model to perform semantic tasks

How Many Senses Do Words Have?

- Dictionaries and thesauri: discrete lists of senses

  Embeddings (static/contextual): high dimensional model of meaning, does not divide up into discrete senses

- How to determine when differing uses of a word should be represented with different sense?

- Example: *serve* from WSJ corpus

  (1) They rarely *serve* red meat, preferring to prepare seafood.
  (2) He *served* as U.S. ambassador to Norway in 1976 and 1977.
  (3) He might have *served* his time, come out and led an upstanding life.

- *serve red meat* and *serve time*: different truth conditions and presuppositions
  serve as ambassador: distinct subcategorization structure *serve as NP*

$\Rightarrow$ 3 distinct senses of *serve*

Outline

Relations: Synonymy

- General idea of synonyms: when two senses of two different words/lemmas are (nearly) identical
  - *couch – sofa*
  - *filbert – hazelnut*
  - *car – automobile*
- Synonymy: relationship between senses rather than words
- Example: *big – large*
  - How big is that plane?
    Would I be flying on a large or small plane?
  - Miss Nelson, for instance, became a kind of big sister to Benjamin.
    Miss Nelson, for instance, became a kind of large sister to Benjamin. (?)
- *big* has a sense of *older/grown up*, while *large* lacks this sense

Relations: Antonymy

- Antonyms: words with an opposite meaning
  - *long – short*
  - *big –little*
  - *fast – slow*
  - *dark – light*
  - *rise – fall*
- Binary opposition, or at opposite ends of a scale
- Reversives: change or movement into in opposite directions
- Antonyms differ with respect to one aspect of their meaning, but otherwise are very similar.

Taxonomic Relations

- A word is a **hyponym** of another word if the first sense is more specific
  - *car* is a hyponym of *vehicle*
  - *dog* is a hyponym of *animal*
  - *mango* is a hyponym of *fruit*
- A **hypernym** is a more generic word
  - *vehicle* is a hypernym of *car*
  - *animal* is a hypernym of *dog*
  - *fruit* is a hypernym of *mango*
- **Meronymy**: a part–whole relation
  - a *leg* is a part of a *chair*
  - a *wheel* is a part of a *car*
- Also: *car* is a **holonym** of *wheel*

Outline

WordNet: A Database of Lexical Relations

- Most commonly used resource for sense relations in English Fellbaum (1998)
- English WordNet consists of three databases:

  nouns, verbs, adjectives/adverbs
- Each database contains a set of lemmas annotated with a set of senses
  - 117,798 nouns (on average: 1.23 senses)
  - 11,529 verbs (on average: 2.16 senses)
  - 22,479 adjectives
  - 4,481 adverbs

WordNet: Example

The noun "bass" has 8 senses in WordNet.
1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso[1] - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Figure G.1 A portion of the WordNet 3.0 entry for the noun *bass*.

- Set of near synonyms: "*synset*"
  - {bass[1], deep[6]}
  - {bass[6], bass voice[1], basso[2]}

→ Lists of word senses that can be used to express the concept

- Glosses are properties of a synset
  each sense included in the synset has the same gloss

WordNet: Synset Example

- {chump^1, fool^2, gull^1, mark^9, patsy^1, fall guy^1, sucker^1, soft touch^1, mug^2}
- **Gloss**: a person who is gullible and easy to take advantage of.

- Synsets are the fundamental unit associated with WordNet entries
- Synsets are labeled with a lexicographic category

WordNet: Synsets

- Categories are called **supersenses**: coarse semantic categories or groupings of senses

Category	Example	Category	Example	Category	Example
ACT	*service*	GROUP	*place*	PLANT	*tree*
ANIMAL	*dog*	LOCATION	*area*	POSSESSION	*price*
ARTIFACT	*car*	MOTIVE	*reason*	PROCESS	*process*
ATTRIBUTE	*quality*	NATURAL EVENT	*experience*	QUANTITY	*amount*
BODY	*hair*	NATURAL OBJECT	*flower*	RELATION	*portion*
COGNITION	*way*	OTHER	*stuff*	SHAPE	*square*
COMMUNICATION	*review*	PERSON	*people*	STATE	*pain*
FEELING	*discomfort*	PHENOMENON	*result*	SUBSTANCE	*oil*
FOOD	*food*			TIME	*day*

Figure G.2 Supersenses: 26 lexicographic categories for nouns in WordNet.

- Nouns: 26 supersenses
  Verbs: 15 supersenses

Sense Relations in WordNet

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 their instances	$composer^1 \rightarrow Bach^1$
Part Meronym	Has-Part	From wholes to parts	$table^2 \rightarrow leg^3$
Part Holonym	Part-Of	From parts to wholes	$course^7 \rightarrow meal^1$
Antonym		Semantic opposition between lemmas	$leader^1 \Longleftrightarrow follower^1$
Derivation		Lemmas w/same morphological root	$destruction^1 \Longleftrightarrow destroy^1$

Figure G.3 Some of the noun relations in WordNet.

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From events to subordinate event	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Semantic opposition between lemmas	$increase^1 \Longleftrightarrow decrease^1$

Figure G.4 Some verb relations in WordNet.

- WordNet represents hyponymy by relating each synset to its immediately more general and more specific synsets
- Produce longer chains of more general or more specific synsets

Sense Relations in WordNet

```
bass[3], basso (an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
bass[7] (member with the lowest range of a family of instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
                     => entity
```

Figure G.5 Hyponymy chains for two separate senses of the lemma *bass*. Note that the chains are completely distinct, only converging at the very abstract level *whole, unit*.

Sense Relations in WordNet

- WordNet has two kinds of taxonomic entities: classes and instances
  - instance: individual, a proper noun that is a unique entity (*San Francisco* is an instance of *city*)
  - *city* is a class, a hyponym of *municipality* and eventually of *location*

Sense Relations in WordNet

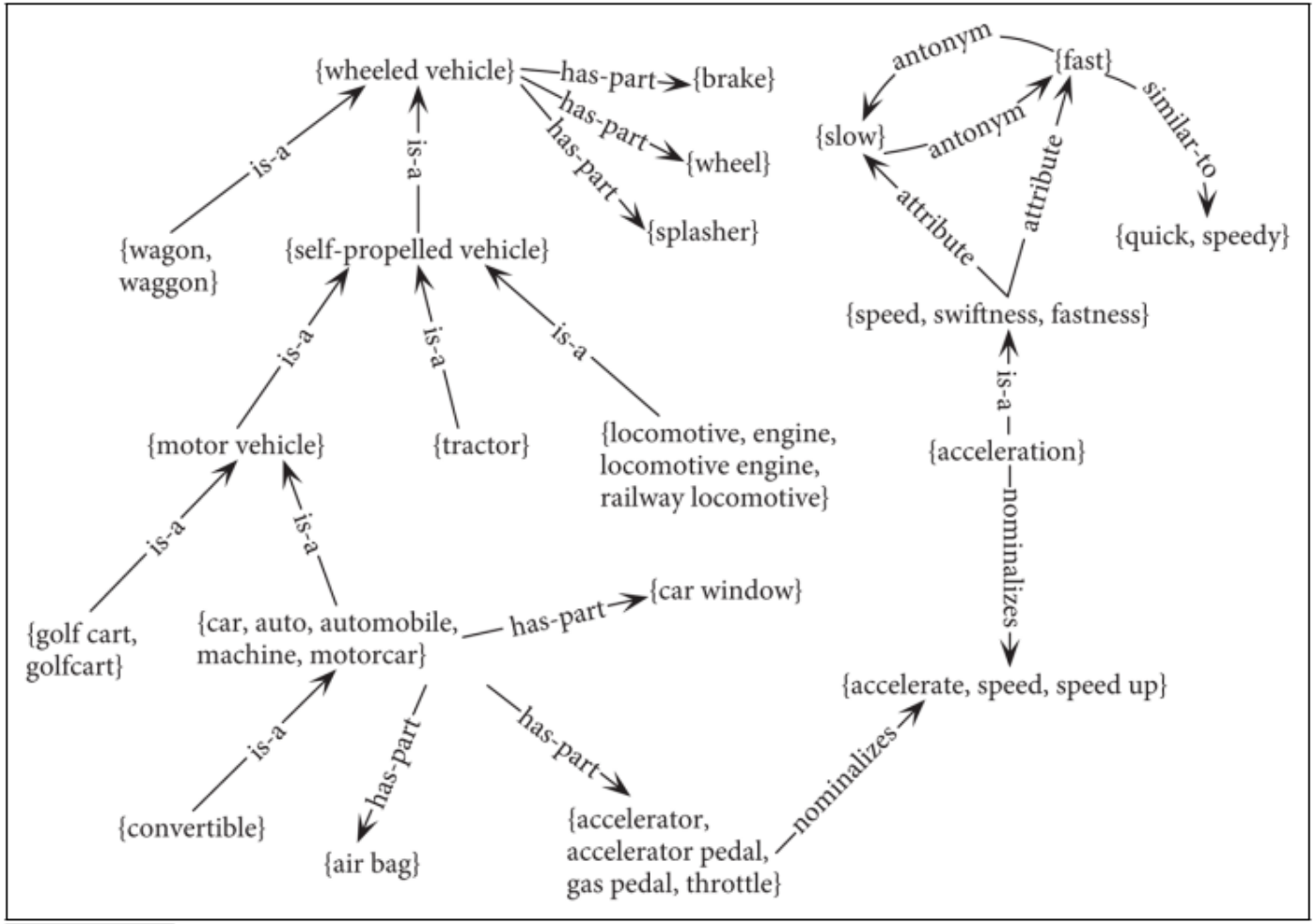


Figure G.6 WordNet viewed as a graph. Figure from Navigli (2016).

Outline

Word Sense Disambiguation

- **Word sense disambiguation (WSD)**: the task of selecting the correct sense for a word
- Inventory of sense tags depends on task

WordNet Sense	Spanish Translation	WordNet Supersense	Target Word in Context
bass^4	lubina	FOOD	...fish as Pacific salmon and striped **bass** and...
bass^7	bajo	ARTIFACT	...play **bass** because he doesn't have to solo...

Figure G.7 Some possible sense tag inventories for *bass*.

Word Sense Disambiguation

- WSD task:
  given an entire texts and a lexicon with an inventory of senses for each entry → disambiguate every word in the text

- Supervised WSD tasks are typically trained on sense-annotated corpora (semantic concordance)

- For example the *SemCor corpus*: subset of the Brown Corpus (226,036 words manually tagged with WordNet senses)

  $\texttt{You}\ \texttt{will}\ \texttt{find}^{9}_{v}\ \texttt{that}\ \texttt{avocado}^{1}_{n}\ \texttt{is}^{1}_{v}\ \texttt{unlike}^{1}_{j}\ \texttt{other}^{1}_{j}\ \texttt{fruit}^{1}_{n}$
  $\texttt{you}\ \texttt{have}\ \texttt{ever}^{1}_{r}\ \texttt{tasted}^{2}_{v}$

- For *fruit*, choose the correct sense:
  - $\texttt{fruit}^{1}_{n}$: the ripened reproductive body of a seed plant
  - $\texttt{fruit}^{2}_{n}$: yield; an amount of a product
  - $\texttt{fruit}^{3}_{n}$: the consequence of some effort or action

Word Sense Disambiguation

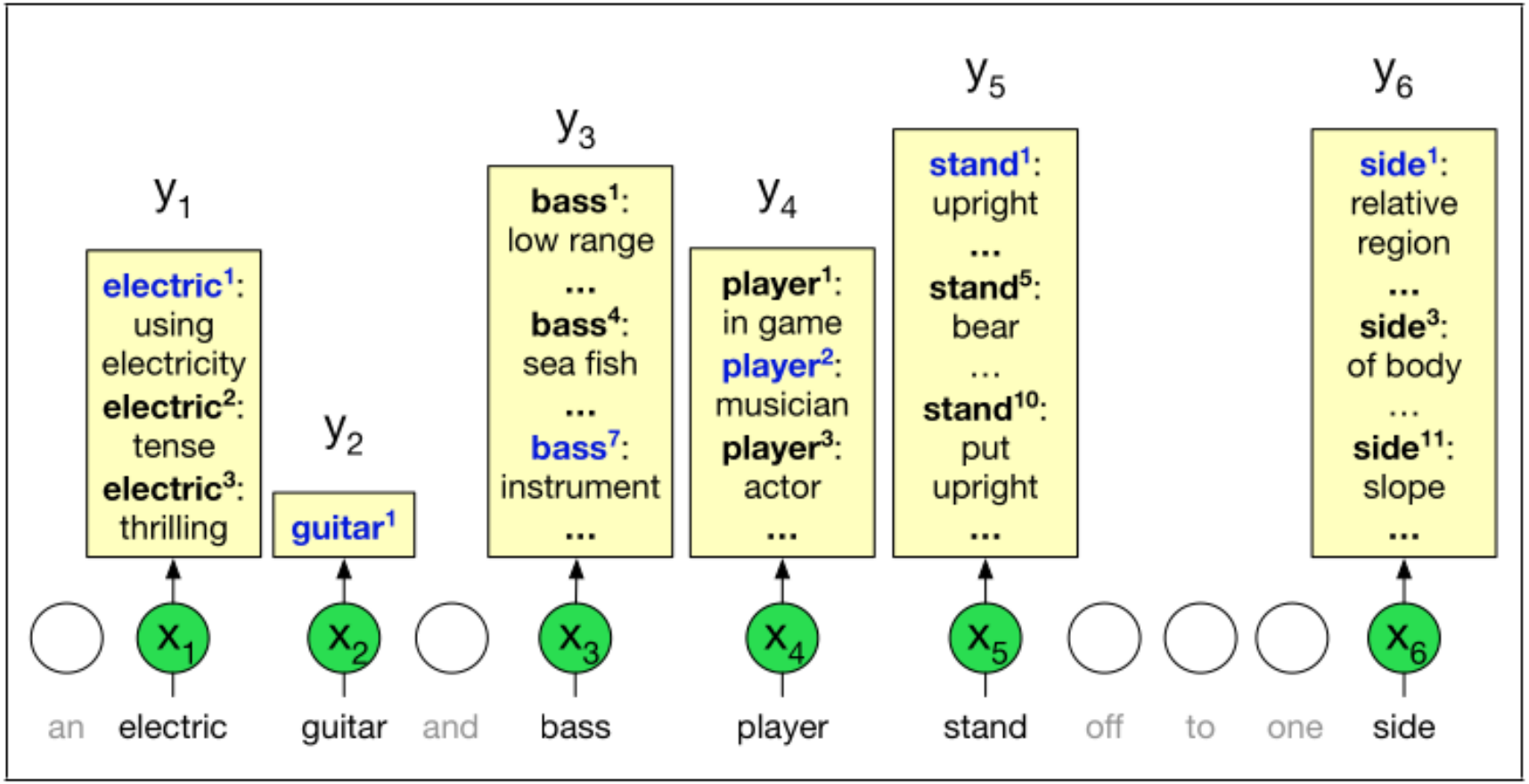


Figure G.8 The all-words WSD task, mapping from input words (x) to WordNet senses (y). Only nouns, verbs, adjectives, and adverbs are mapped, and note that some words (like *guitar* in the example) only have one sense in WordNet. Figure inspired by Chaplot and Salakhutdinov (2018).

Word Sense Disambiguation: Simple Strategies

- **Most frequent sense**: surprisingly strong baseline
- Senses in WordNet are generally ordered from most frequent to least frequent based on their counts
- Quite accurate, and is therefore often used as a default

- **One sense per discourse**: a word appearing multiple times in a text often appears with the same sense
- Better for coarse-grained senses and particularly when a word's senses are unrelated

Word Sense Disambiguation: Contextual Embeddings

- Simple 1-nearest-neighbor algorithm using contextual word embeddings
- Training time:
  - pass each sentence in SemCore labeled dataset through contextual embedding (e.g., BERT)
  - Produce a contextual sense embedding v_s for s:

    $\mathbf{v}_s = \frac{1}{n}\sum_i \mathbf{v}_i \quad \forall \mathbf{v}_i \in token(s)$

    for each of the n tokens of sense s: average n contextual representations
- Test time:
  - given target word t in context: compute its contextual embedding $\mathbf{t}$
  - choose its nearest neighbor sense from the training set

$$\text{sense}(t) = \underset{s \in \text{senses}(t)}{\text{argmax}} \ \text{cosine}(\mathbf{t}, \mathbf{v}_s)$$

Word Sense Disambiguation: Nearest Neighbor

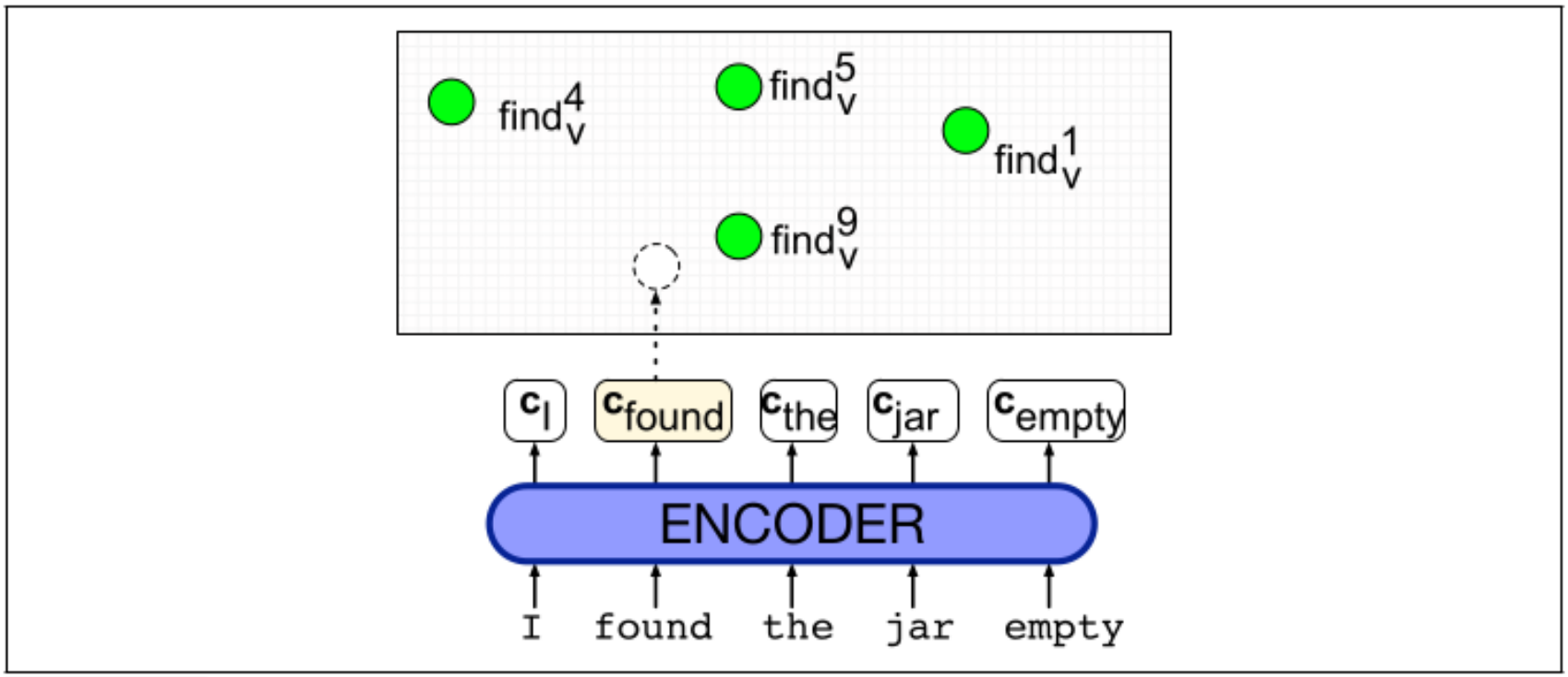


Figure G.9 The nearest-neighbor algorithm for WSD. In green are the contextual embeddings precomputed for each sense of each word; here we just show a few of the senses for *find*. A contextual embedding is computed for the target word *found*, and then the nearest neighbor sense (in this case $\textbf{find}^9_v$) is chosen. Figure inspired by Loureiro and Jorge (2019).

The Simplified Lesk Algorithm as WSD Baseline

- Generating sense labeled corpora like SemCor: difficult and expensive
- Knowledge-based algorithms rely solely on WordNet or and don't require labeled data
  - supervised algorithms typically better
- Lesk algorithm: choose the sense whose dictionary gloss or definition shares the most words with the target word's neighborhood

The Simplified Lesk Algorithm as WSD Baseline

- Disambiguate *bank* in the following context

  The bank can guarantee 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"

  - sense *bank*^{1} has two overlaps with the context: *deposits*, *mortgage*
  - sense *bank*^{2}: no overlap

- Variations:
  - weighting overlapping words by inverse document frequency (IDF)
  - use word embedding cosine instead of word overlap

Word-in-Context Evaluation

- Context-free word similarity task: how similar is *cup* to *mug*?
- WSD as contextualized similarity task: distinguish the meaning of a word in one context from another context
- Word-in-Context task:
  two sentences with the same target word → used in the same sense?

F	There's a lot of trash on the **bed** of the river — I keep a glass of water next to my **bed** when I sleep
F	**Justify** the margins — The end **justifies** the means
T	**Air** pollution — Open a window and let in some **air**
T	The expanded **window** will give us time to catch the thieves — You have a two-hour **window** of clear weather to finish working on the lawn

Figure G.11 Positive (T) and negative (F) pairs from the WiC dataset (Pilehvar and Camacho-Collados, 2019).

Wikipedia as Source of Training Data

- Wikipedia as a source of sense-labeled data
- Articles in Wikipedia contain explicit links to concepts → link as sense annotation

> In 1834, Sumner was admitted to the **[[bar (law)|bar]]** at the age of twenty-three, and entered private practice in Boston.
>
> It is danced in 3/4 time (like most waltzes), with the couple turning approx. 180 degrees every **[[bar (music)|bar]]**.
>
> Jenga is a popular beer in the **[[bar (establishment)|bar]]**s of Thailand.

- Add senses to training data of a supervised system
- Map Wikipedia concepts to relevant sense inventory (e.g. WordNet)
  - find WordNet sense with lexical overlap with the Wikipedia concept
  - vector of words in WordNet synset, gloss and related senses with
  - vector of words in Wikipedia page title, outgoing links and page category

Thesauri to Improve Embeddings

- Thesauri have been used to improve word embeddings
- Static word embeddings have a problem with antonyms
  - *expensive* is often similar in embedding cosine to *cheap*
- Two strategies to include information from thesauri
- Retraining:
  modify the embedding training to incorporate thesaurus relations
- Retro-fitting or counterfitting:
  after embedding training, learn a second mapping using thesaurus information: move synonyms closer and antonyms further apart

	Before counterfitting			After counterfitting		
east	west	north	south	eastward	eastern	easterly
expensive	pricey	cheaper	costly	costly	pricy	overpriced
British	American	Australian	Britain	Brits	London	BBC

Figure G.12 The nearest neighbors in GloVe to *east*, *expensive*, and *British* include antonyms like *west*. The right side showing the improvement in GloVe nearest neighbors after the counterfitting method (Mrkšić et al., 2016).

Outline

Credits

Content based on:

- Dan Jurafsky and James H. Martin (2024)
  Speech and Language Processing: Appendix Chapter G
  `https://web.stanford.edu/ jurafsky/slp3/`

References

- Fellbaum, C., editor. 1998.
WordNet: An Electronic Lexical Database. MIT Press.