Carnegie Mellon University
Language
Technologies
Institute

11-411/11-611 Natural Language Processing

Lexical Semantics

David R. Mortensen and Kemal Oflazer

March 30, 2023

Language Technologies Institute

Lexical Semantics

The study of meanings of words:

- Decompositional
  - Words have component meanings
  - Total meanings are composed of these component meanings
- Ontological
  - The meanings of words can be defined in relation to other words.
  - Paradigmatic—Thesaurus-based
- Distributional
  - The meanings of words can be defined in relation to their contexts among other words.
  - Syntagmatic—meaning defined by syntactic context

Decompositional Lexical Semantics

Decompositional Lexical Semantics

- Assume that *woman* has (semantic) components [female], [human], and [adult].
- *Man* might have the components [male], [human], and [adult].

Lexical Semantics by Decomposition

A simple (and problematic) example:

boy

$$\begin{bmatrix} -\text{female} \\ -\text{adult} \\ +\text{human} \end{bmatrix}$$

girl

$$\begin{bmatrix} +\text{female} \\ -\text{adult} \\ +\text{human} \end{bmatrix}$$

man

$$\begin{bmatrix} -\text{female} \\ +\text{adult} \\ +\text{human} \end{bmatrix}$$

woman

$$\begin{bmatrix} +\text{female} \\ +\text{adult} \\ +\text{human} \end{bmatrix}$$

What features would you use to define *gawk*?

https://www.newyorker.com/magazine/2012/12/24/utopian-for-beginners

"Hunched over the dining-room table, Quijada showed me how he would translate "gawk" into Ithkuil. First, though, since words in Ithkuil are assembled from individual atoms of meaning, he had to engage in some introspection about what exactly he meant to say. ...As he assembled the evolving word from its constituent meanings, he scribbled its pieces on a notepad. He added the "second degree of the affix for expectation of outcome" to suggest an element of surprise that is more than mere unpreparedness but less than outright shock, and the "third degree of the affix for contextual appropriateness" to suggest an element of impropriety that is less than scandalous but more than simply eyebrow-raising. As he rapped his pen against the notepad, he paged through his manuscript in search of the third pattern of the first stem of the root for "shock" to suggest a "non-volitional physiological response," and then, after several moments of contemplation, he decided that gawking required the use of the "resultative format" to suggest "an event which occurs in conjunction with the conflated sense but is also caused by it." He eventually emerged with a tiny word that hardly rolled off the tongue:apq'uxasiu. He spoke the first clacking syllable aloud a couple of times before deciding that he had the pronunciation right, and then wrote it down in the script he had invented for printed Ithkuil:"

Ontological Lexical Semantics and WordNet

What is an Ontology? Second Definition Below

Dictionary

Definitions from Oxford Languages · Learn more

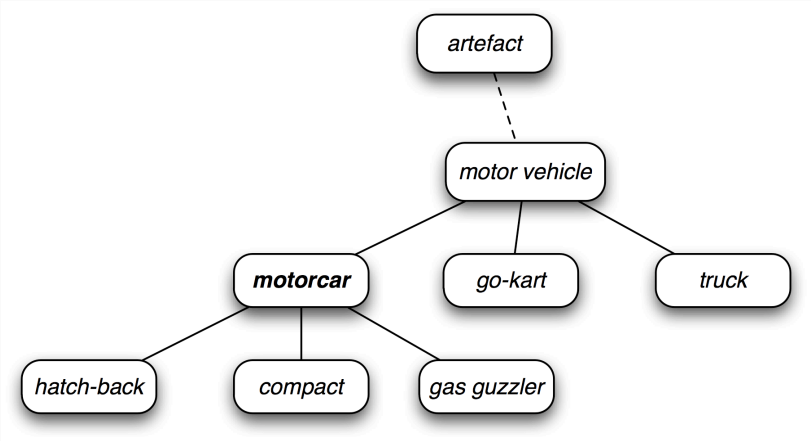
on·tol·o·gy

noun

noun: **ontology**; plural noun: **ontologies**

1. the branch of metaphysics dealing with the nature of being.
2. a set of concepts and categories in a subject area or domain that shows their properties and the relations between them.
   "what's new about our ontology is that it is created automatically from large datasets"

WordNet Ontology Fragment



WordNet: a Widely Used Ontological Lexical Resource

WordNet

- A hierarchically organizated database of (English) word senses.
- George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.
- Available at *wordnet.princeton.edu*
- Provides a set of three lexical databases:
  - Nouns
  - Verbs
  - Adjectives and adverbs.
- Relations are between senses, not lexical items (words).
- Applications Program Interfaces (APIs) are available for many languages and toolkits including a Python interface via NLTK.
- WordNet 3.0

Category	Unique Strings
Noun	117,197
Verb	11,529
Adjective	22,429
Adverb	4,481

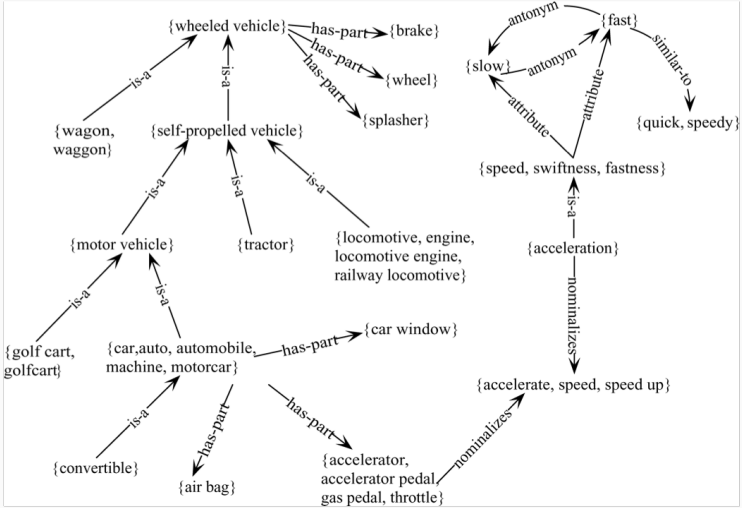
WordNet Noun Relations

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	$destruction^1 \Longleftrightarrow destroy^1$

WordNet Verb Relations

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From events to subordinate event (often via specific manner)	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Semantic opposition between lemmas	$increase^1 \Longleftrightarrow decrease^1$
Derivationally Related Form	Lemmas with same morphological root	$destroy^1 \Longleftrightarrow destruction^1$

WordNet as as Graph



Meronymy/Holonymy: the “part-of” relation

- Lexical item *a* is a meronym of lexical item *b* if a sense of *a* is a part of/a member of a sense of *b*.
  - *hand* is a meronym of *body*.
  - *congressperson* is a meronym of *congress*.
- *Holonymy* (think: whole) is the converse of *meronymy*.
  - *body* is a holonym of *hand*.

Hyponymy/Hypernymy

- A hypernym and a hyponym can be in an “is-a” relation.
  - A screwdriver (hyponym) is a tool (hypernym).
- How to remember which is which:
  - “Hyper” means “over” or “excess” as in *hyperthermia*.
  - “Hypo” means “under” or “not enough” as in *hypothermia* or *hypodermic*.

Hyponymy/Hypernymy

- Lexical item *a* is a **hyponym** of lexical item *b* if *a* is a kind of *b* (if a sense of *b* refers to a superset of the referent of a sense of *a*).
  - *screwdriver* is a hyponym of *tool*.
  - *screwdriver* is also a hyponym of *drink*.
  - *car* is a hyponym of *vehicle*
  - *mango* is a hyponym of *fruit*
- **Hypernymy** is the converse of hyponymy.
  - *tool* and *drink* are hypernyms of screwdriver.
  - *vehicle* is a hypernym of *car*

Hyponymy more formally

- **Extensional**
  - The class denoted by the superordinate (e.g., vehicle) extensionally includes the class denoted by the hyponym (e.g. car).
- **Entailment**
  - A sense *A* is a hyponym of sense *B* if being an *A* entails being a *B* (e.g. if it is car, it is a vehicle)
- Hyponymy is usually transitive
  - If *A* is a hyponym of *B* and *B* is a hyponym of *C* $\Rightarrow$ *A* is a hyponym of *C*.
- Another name is the **IS-A hierarchy**
  - *A* IS-A *B*
  - *B* **subsumes** *A*

Hyponyms and Instances

- An **instance** is an individual, a proper noun that is a unique entity
  - Doha/San Francisco/London are instances of *city*.
- But *city* is a class
  - *city* is a hyponym of *municipality* ...*location* ...

The IS-A Hierarchy for *fish (n)*

- fish (any of various mostly cold-blooded aquatic vertebrates usually having scales and breathing through gills)
- aquatic vertebrate (animal living wholly or chiefly in or on water)
- vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
- chordate (any animal of the phylum Chordata having a notochord or spinal column)
- animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
- organism, being (a living thing that has (or can develop) the ability to act or function independently)
- living thing, animate thing (a living (or once living) entity)
- whole, unit (an assemblage of parts that is regarded as a single entity)
- object, physical object (a tangible and visible entity; an entity that can cast a shadow)
- entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Hierarchy for $bass_3$ in WordNet

- 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))

Supersenses in WordNet

Super senses are top-level hypernyms in the hierarchy.

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 23.2 Supersenses: 26 lexicographic categories for nouns in WordNet.

Terminology Break: Lemma, Wordform, Word Sense, Homonymy, Polysemy, Synonymy

Terminology: Lemma and Wordform

- A **lemma** or **citation form**
  - The form of the word that you find in a dictionary
- A **wordform**
  - The (inflected) word as it appears in text

Wordform	Lemma
banks	bank
sung	sing
sang	sing
went	go
goes	go

Lemmas have Senses (meanings) or Word Senses

- One lemma “bank” can have many meanings:
  - Sense 1: “...a bank_1 can hold the investments in a custodial account ...”
  - Sense 2: “...as agriculture burgeons on the east bank_2 the river will shrink even more.”
- The lemma **bank** here has two senses.

Homonymy

- **Homonyms**: lemmas that share a 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**: Same spelling (wind, bass)
- **Homophones**: Same pronunciation
  - write and right
  - piece and peace

Homonymy causes problems for NLP applications

- Information retrieval
  - “bat care”
- Machine Translation
  - bat: murciélago (animal) or bate (for baseball)
- Text-to-Speech
  - bass (stringed instrument) vs. bass (fish)

How do we know when a lemma has multiple senses?

- The “zeugma” test: a word (e.g., *serve*) applies to two other words (e.g., *breakfast* and *Washington*) with two different meanings (e.g., “give” and ”fly to”).
  - Which flights serve breakfast? (give you breakfast)
  - Does Qatar Airways serve Philadelphia? (fly to Philadelphia)
  - ?Does Qatar Airways serve breakfast and Washington?
- Since this conjunction sounds weird, we say that these are two different senses of the lemma “serve”, $serve_1$ (give) and $serve_2$ (fly to)
- $serve_1$ (give) and $serve_2$ (fly to) are homophones.

More examples of zeugma

- “The farmers in the valley grew potatoes, peanuts, and bored.”
- “He lost his coat and his temper.”
- “I held my tongue and his hand.”
- “He made a good husband and a good omelette”

Polysemy

- The bank_1 was constructed in 1875 out of local red brick.
- I withdrew the money from the bank_2.
- Are those the same sense?
  - bank_1 : "The building belonging to a financial institution"
  - bank_2: "A financial institution"
- A **polysemous** word has multiple related meanings.
- Most non-rare words have multiple meanings

Polysemy vs Homonymy

Polysemy and homonymy are treated differently in a dictionary. The dictionary makes different entries for homonymy. It makes sub-entries for polysemy because the meanings are more predictable or related.

bank-2 1. Any piled up mass of things like snow or clouds. 2. The slope of land adjoining a body of water. etc.

bank-1 1. a business establishment offering financial services. 2. The offices or building in which such a business is located. 3. Any place of safekeeping or storage. etc.

American Heritage Dictionary

A note on lexicography

- Lexicographers make dictionaries
- Lexicography is often corpus-based
  - Human lexicographers collect examples of word usage from corpora (on paper in the old days). Currently they use *concordances* on electronic corpora.
  - They sort all the sentences containing a word into head word entries (the boldface entries in paper dictionaries) and sub-senses (the numbered entries under a head word), until they are pretty sure they have covered enough meanings.
  - They then insert sample sentences based on the corpus (usually in italics in a paper dictionary) into the dictionary to illustrate the senses.
- Lexicographers can be lumpers (make fewer sense distinctions) or splitters (make more sense distinctions).

Synonymy

- Words *a* and *b* share an identical sense or have the same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - water / H_2O
- Synonyms can be substituted for each other in all situations.

True Synonymy is Rare

- True synonymy is relatively rare compared to other lexical relations.
  - may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.
  - water / H_2O
  - big / large
  - bravery / courage
    - Bravery is the ability to confront pain, danger, or attempts of intimidation without any feeling of fear.
    - Courage, on the other hand, is the ability to undertake an overwhelming difficulty or pain despite the eminent and unavoidable presence of fear.

Synonymy

Synonymy is a relation between senses rather than words.

- Consider the words *big* and *large*. Are they synonyms?
  - How **big** is that plane?
  - Would I be flying on a **large** or small plane?
- How about here:
  - Miss Nelson became a kind of **big** sister to Benjamin.
  - ?Miss Nelson became a kind of **large** sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up
  - *large* lacks this sense.

Antonymy

- Lexical items *a* and *b* have senses which are “opposite”, with respect to one feature of meaning
- Otherwise they are similar
  - dark/light
  - short/long
  - fast/slow
  - rise/fall
  - hot/cold
  - up/down
  - in/out
- More formally: antonyms can
  - define a binary opposition or be at opposite ends of a scale (long/short, fast/slow)
  - or be **reversives** (rise/fall, up/down)
- Antonymy is much more common than true synonymy.
- Antonymy is not always well defined, especially for nouns (but for other words as well).

Back to WordNet

Synsets: Synonym Sets

- Word senses that can be given the same gloss or definition.
- Does not require absolute synonymy
- Example:
  - Synset: {*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

Lexical Relations Connect Synsets

Lexical relations, like *antonymy*, *hyponymy*, and *meronymy* are between synsets, not between senses directly.

Synsets for *dog (n)*

- S: (n) **dog, domestic dog, Canis familiaris** (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) "the dog barked all night"
- S: (n) **frump, dog** (a dull unattractive unpleasant girl or woman) "she got a reputation as a frump", "she's a real dog"
- S: (n) **dog** (informal term for a man) "you lucky dog"
- S: (n) **cad, bounder, blackguard, dog, hound, heel** (someone who is morally reprehensible) "you dirty dog"
- S: (n) **frank, frankfurter, hotdog, hot dog, dog, wiener, wienerwurst, weenie** (a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll)
- S: (n) **pawl, detent, click, dog** (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)
- S: (n) **andiron, firedog, dog, dog-iron** (metal supports for logs in a fireplace) "the andirons were too hot to touch"

Synsets for *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"*

WordNets for Other Languages

- *globalwordnet.org/wordnets-in-the-world/* lists WordNets for tens of languages.
- Many of these WordNets are linked through **ILI – Interlingual Index** numbers.

Best and Worst of WordNet

- Best:
  - Like any dictionary it was a massive effort
  - Lexicographers, psychologists, and linguists were involved
  - It is a large resource that represents a lexicon as a graph and is based on an ontology
  - Many other languages were encouraged to make WordNets and align their Synsets with English Synsets, resulting in a large, multi-lingual resource.
- Worst:
  - The lexicographers were splitters. There are more senses than people can distinguish. Corpora that are annoted with WordNet senses may not have high *intercoder agreement*.
  - Many of the multilingual WordNets do not align well with English. The lexicographers did not always do a good job. Perhaps they didn't fully grasp the English senses, or perhaps they aligned Synsets automatically.

Task: Word Sense Disambiguation

Word Sense Disambiguation

Input: You will find that avocado is unlike any other fruit you have ever tasted.

Output: Each word is labeled with its correct sense from WordNet.

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

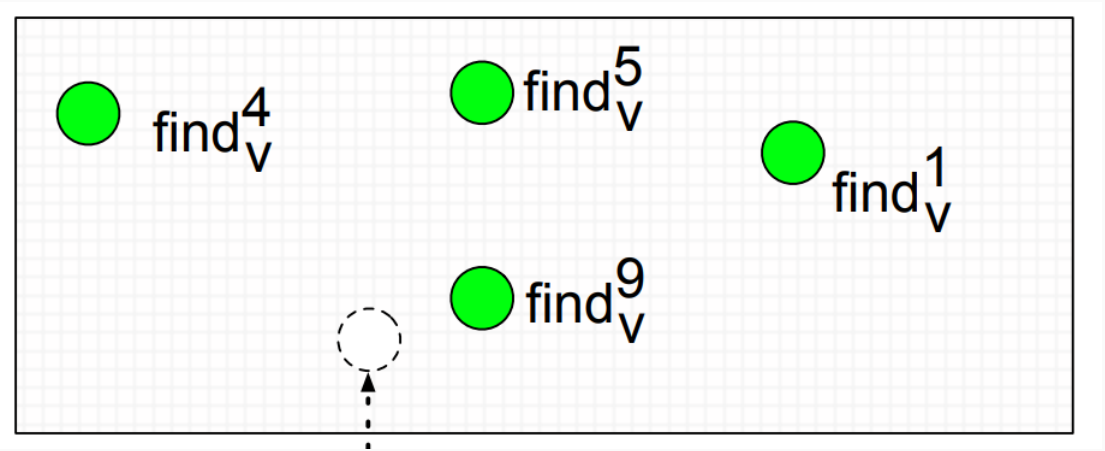
Word Sense Disambiguation

- SemCore Corpus: Human annotators assigned a WordNet sense to each word in a corpus (a subset of the Brown Corpus, 1967).
  - This is very difficult because the WordNet lexicographers were “splitters” (when in doubt, make a new sense). Many senses are too close to call.
- The baselines:
  - always choose the most frequent sense
  - a word tends to have one sense per document

A simple WSD model

- Train on SemCore (sense-labeled corpus)
- Use BERT to get a contextual embedding for each word in SemCore
- Then for each sense (which you get from the SemCore corpus), make a *sense embedding* by averaging all the embeddings for all of the word tokens that have that sense label.
- At run time: get a contextualized embedding for a word from BERT. Find the closest sense embedding.

Step 1: contextual embedding for each word in SemCore



For each sense in SemCor, make a *sense embedding* by averaging all the embeddings for all of the word tokens that have that sense label.

Step 2: Find a nearest neighbor at run time

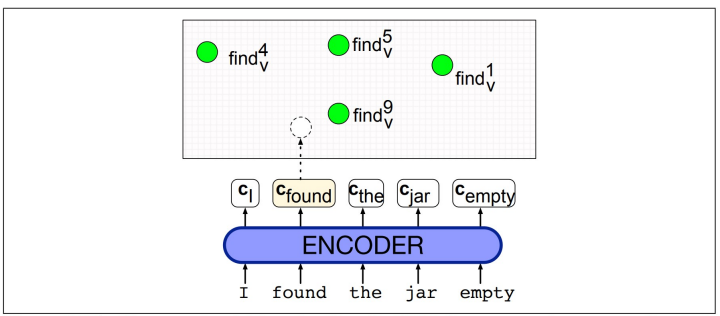


Figure 23.9 The nearest-neighbor algorithm for WSD. In green are the contextual embeddings precomputed for each sense of each word; here we just show a few of the senses for *find*. A contextual embedding is computed for the target word *found*, and then the nearest neighbor sense (in this case $\mathbf{find}^9_v$) is chosen. Figure inspired by Loureiro and Jorge (2019).

At run time: get a contextualized embedding for a word from BERT. Find the closest sense embedding.

For words not in WordNet, but not in SemCor, use SynSet as a backoff

- Suppose *chanced* was not in SemCor and your sentence at run-time is *I chanced upon the jar*.
- *Chanced* is in WordNet.
- It belongs to a Synset that is glossed *Come upon as if by accident*, which includes senses of *find*, *happen*, *chance*, *bump*, and *encounter*.
- It also belongs to two other SynSets: be the case by chance, and take a risk in the hope of a favorable outcome.
- Make embeddings for WordNet SynSets by averaging the embeddings of all of the instances of all of the members of the SynSet in SemCor.
- For example, since *chance* is (for the sake of this example) not in SemCor, the embedding of the SynSet *come upon as if by accident* will be the average all of the embeddings of *find*, *happen*, *bump*, and *encounter* in SemCor that belong to this SynSet.
- For *I chanced upon the jar*, compare the contextual embedding of *chanced* to the embeddings of all three of SynSets it belongs to.
- For more backoffs in case the SynSet doesn't work: there can also be embeddings of hypernyms of the SynSets that *chanced* belongs too and there can also be an embedding of the supersense.