Natural Language Processing
DSECL ZG565

Dr. Chetana Gavankar, Ph.D,

IIT Bombay-Monash University Australia

Associate Professor, BITS Pilani

Chetana.gavankar@pilani.bits-pilani.ac.in

BITS Pilani

Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

Session 13 – Word sense disambiguation and WordNet
Date – 16th March 2024

These slides are prepared by the instructor, with grateful acknowledgement of Prof. Jurafsky and Prof. Martin and many others who made their course materials freely available online.

Session Content
(Ref: Chapter 19 Jurafsky and Martin)

- **Word Senses**
- **Relations between Senses**
- **WordNet: A Database of Lexical Relations**
- **Word Sense Disambiguation**
- **Alternate WSD algorithms and Tasks**
- **Using Thesauruses to Improve Embeddings**
- **Word Sense Induction**

What's a word sense?

- Lexeme: An entry in a lexicon consisting of a pairing of a form with a single meaning representation
- A **lemma** or **citation form** is the grammatical form that is used to represent a **lexeme**.
    - ***Carpet*** is the lemma for ***carpets***
- The lemma *bank* has two **senses:**
    - Instead, a **bank** can hold the investments in a custodial account in the client's name
    - But as agriculture burgeons on the east **bank**, the river will shrink even more.
- A **sense** is a discrete representation of one aspect of the meaning of a word

Word senses and relationships between word senses

- Homonymy
- Polysemy
- Synonymy
- Antonymy
- Hypernomy
- Hyponomy

Homonymy

- Lexemes that share a form
  - Phonological, orthographic or both
- But have unrelated, distinct meanings
  - Examples
    - bat (wooden stick-like thing) vs bat (flying scary mammal thing)
    - bank (financial institution) versus bank (riverside)
  - Can be homophones, homographs, or both:
    - Homophones:
      - Write and right
      - Piece and peace

Homonymy causes problems for NLP applications

- Text-to-Speech
  - Same orthographic form but different phonological form
    - bass vs bass
- Information retrieval
  - Different meanings same orthographic form
    - QUERY: bat care
- Machine Translation
- Speech recognition
  - Why?

Polysemy

- The **bank** is constructed from red brick
- I withdrew the money from the **bank**
  - Which sense of bank is this?
    - Is it distinct from the river bank sense?
    - How about the savings bank sense?

Another example:

- His cottage is near a small **wood**.
- The statue was made out of a block of **wood**.

Are those the same sense?

Polysemy

- A single lexeme with multiple related meanings (bank the building, bank the financial institution)
- Most non-rare words have multiple meanings
  - The number of meanings is related to its frequency
  - Verbs tend more to polysemy
  - Distinguishing polysemy from homonymy isn’t always easy (or necessary)

Synonyms

- Word that have the same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - Water / H_20
- Two lexemes are synonyms if they can be successfully substituted for each other in all situations

But

- There are no examples of perfect synonymy
  - Why should that be?
  - Even if many aspects of meaning are identical
  - Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.
- Example:
  - Water and H_20

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, for instance, became a kind of **big** sister to Benjamin.
  - ?Miss Nelson, for instance, became a kind of **large** sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up
  - *large* lacks this sense

Antonyms

- Senses that are opposites with respect to one feature of their meaning
- Otherwise, they are very similar!
  - dark / light
  - short / long
  - hot / cold
  - up / down
  - in / out
- More formally: antonyms can
  - define a binary opposition or at opposite ends of a scale (*long/short, fast/slow*)
  - Be **reverses**: *rise/fall, up/down*

Hyponymy

- 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*
  - *dog* is a hyponym of *animal*
  - *mango* is a hyponym of *fruit*
- Conversely
  - *vehicle* is a hypernym/superordinate of *car*
  - *animal* is a hypernym of *dog*
  - *fruit* is a hypernym of *mango*

superordinate	vehicle	fruit	furniture	mammal
hyponym	car	mango	chair	dog

WordNet

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
    - Versions for other languages are under development
- Avr. noun has 1.23 sense
- Avr. verb has 2.16 senses

Category	Entries
Noun	117,097
Verb	11,488
Adjective	22,141
Adverb	4,601

Format of Wordnet Entries

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)

The adjective "bass" has 1 sense in WordNet.

1. bass^{1}, deep^{6} - (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

Wordnet

- The set of near-synonyms for a WordNet sense is called a **synset (synonym set)**; it's their version of a sense or a concept

Example: chump as a noun to mean

'a person who is gullible and easy to take advantage of'

{chump^1, fool^2, gull^1, mark^9, patsy^1, fall guy^1, soft touch^1, mug^2}

- Each of these senses share this same gloss
- Thus for WordNet, the meaning of this sense of chump *is* this list.

WordNet Noun Relations

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$\textit{breakfast}^1 \rightarrow \textit{meal}^1$
Hyponym	Subordinate	From concepts to subtypes	$\textit{meal}^1 \rightarrow \textit{lunch}^1$
Member Meronym	Has-Member	From groups to their members	$\textit{faculty}^2 \rightarrow \textit{professor}^1$
Has-Instance		From concepts to instances of the concept	$\textit{composer}^1 \rightarrow \textit{Bach}^1$
Instance		From instances to their concepts	$\textit{Austen}^1 \rightarrow \textit{author}^1$
Member Holonym	Member-Of	From members to their groups	$\textit{copilot}^1 \rightarrow \textit{crew}^1$
Part Meronym	Has-Part	From wholes to parts	$\textit{table}^2 \rightarrow \textit{leg}^3$
Part Holonym	Part-Of	From parts to wholes	$\textit{course}^7 \rightarrow \textit{meal}^1$
Antonym		Opposites	$\textit{leader}^1 \rightarrow \textit{follower}^1$

WordNet Verb Relations

Relation	Definition	Example
Hypernym	From events to superordinate events	*fly*^{9} $\rightarrow$ *travel*^{5}
Troponym	From a verb (event) to a specific manner elaboration of that verb	*walk*^{1} $\rightarrow$ *stroll*^{1}
Entails	From verbs (events) to the verbs (events) they entail	*snore*^{1} $\rightarrow$ *sleep*^{1}
Antonym	Opposites	*increase*^{1} $\Longleftrightarrow$ *decrease*^{1}

WordNet Hierarchies

```
Sense 3
bass, 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
               => causal agent, cause, causal agency
                  => physical entity
                     => entity

Sense 7
bass --
(the member with the lowest range of a family of
musical instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
                     => entity
```

WordNet as graph

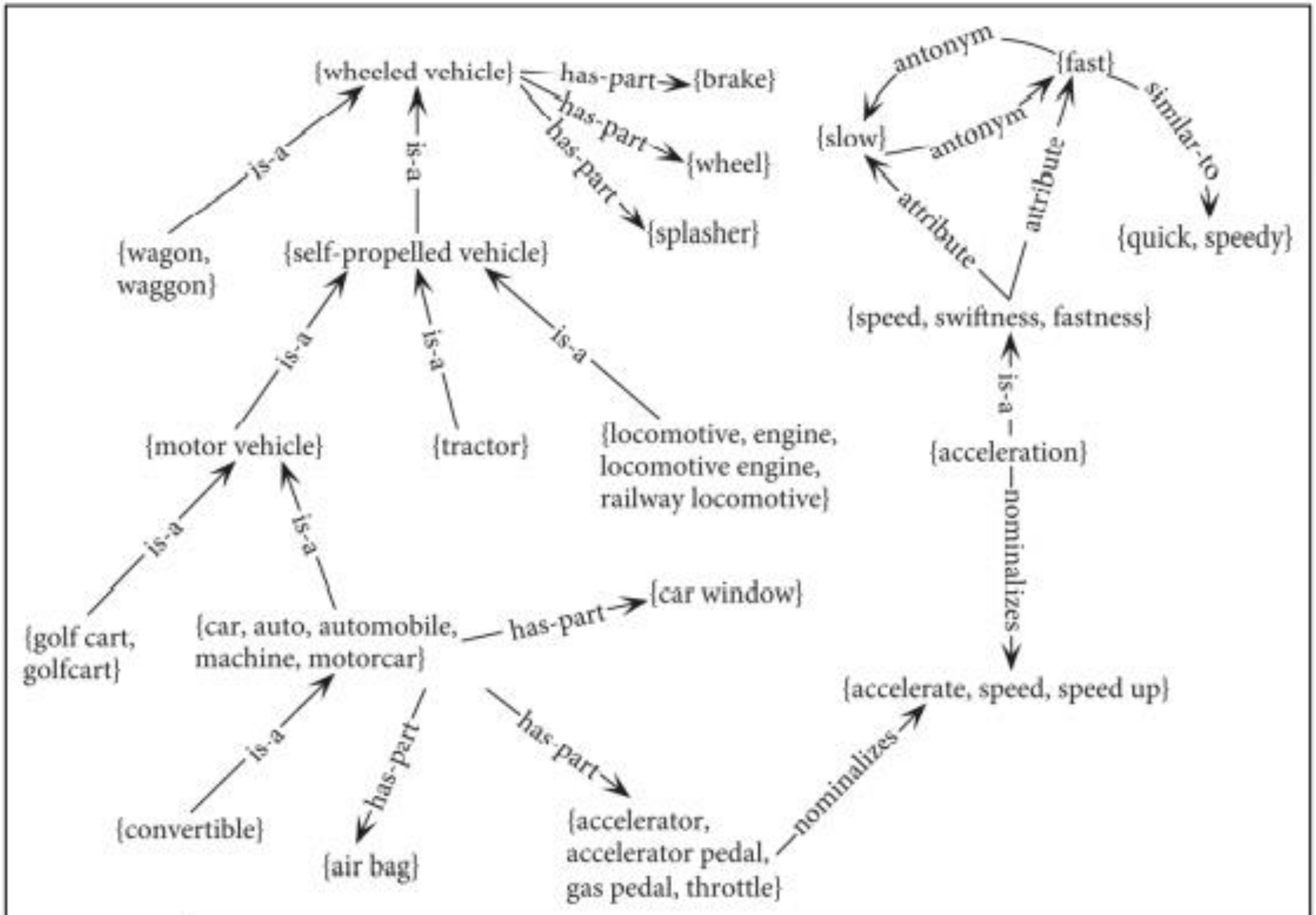


Figure 19.6 WordNet viewed as a graph. Figure from Navigli (2016).

Word Sense Disambiguation (WSD)

- The task of selecting the correct sense for a word is called word sense disambiguation, or WSD
- Given
  - a word in context,
  - a fixed inventory of potential word sense
- Decide which sense of the word this is

- Examples
  - English-to-Spanish MT
    - Inventory is set of Spanish translations
  - Speech Synthesis
    - Inventory is homographs with different pronunciations like *bass* and *bow*

WSD: The Task and Datasets

- The inventory of sense tags depends on the task.
- For sense tagging in the context of translation from English to Spanish, the sense tag inventory for an English word might be the set of different Spanish translations.
- For automatic indexing of medical articles, the sense-tag inventory might be the set of MeSH (Medical Subject Headings) thesaurus entries.
- We can use the set of senses from a resource like WordNet, or supersenses if we want a coarser-grain set.

Inventory of sense tags for *bass*

WordNet Sense	Spanish Translation	Roget Category	Target Word in Context
$bass^4$	lubina	FISH/INSECT	. . . fish as Pacific salmon and striped **bass** and. . .
$bass^4$	lubina	FISH/INSECT	. . . produce filets of smoked **bass** or sturgeon. . .
$bass^7$	bajo	MUSIC	. . . exciting jazz **bass** player since Ray Brown. . .
$bass^7$	bajo	MUSIC	. . . play **bass** because he doesn't have to solo. . .

Two variants of WSD task

- Lexical sample task
  - Small pre-selected set of target words
  - And inventory of senses for each word

- All-words task
  - In this all-words task, the system is given an all-words entire texts and
  - lexicon with an inventory of senses for each entry
  - we have to disambiguate every word in the text (or sometimes just every content word).

WSD Tags

- What's a tag?
  - A dictionary sense?
- For example, for WordNet an instance of "bass" in a text has 8 possible tags or labels (bass1 through bass8).

WordNet Bass

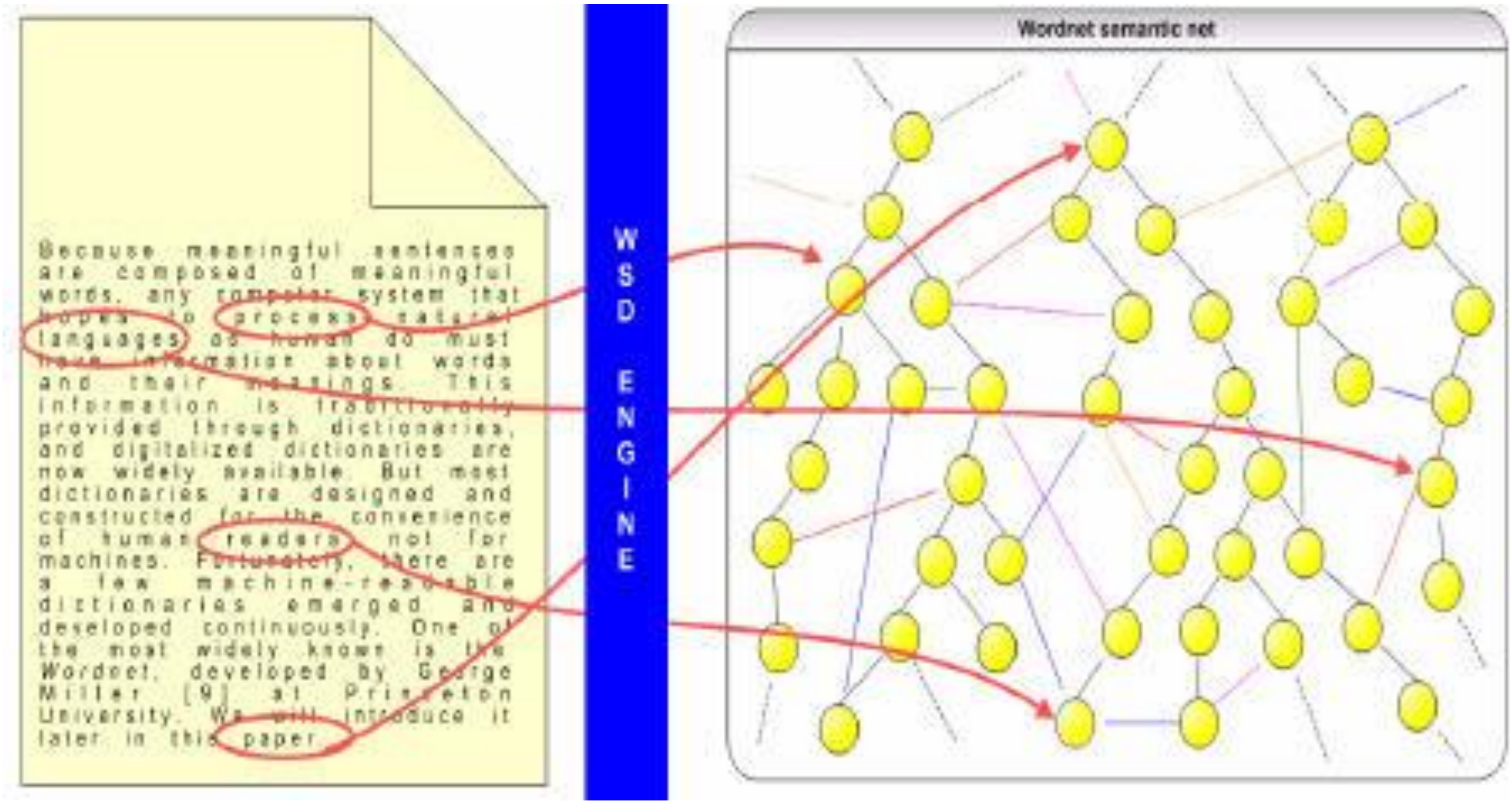
The noun ``bass'' has 8 senses in WordNet

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

Training Corpus

- Lexical sample task:
  - *Line-hard-serve* corpus - 4000 examples of each
  - *Interest* corpus - 2369 sense-tagged examples
  - Since the set of words and the set of senses are small, simple supervised classification approaches work very well.
- All words:
  - **Semantic concordance**: a corpus in which each open-class word is labeled with a sense from a specific dictionary/thesaurus.
    - SemCor: 234,000 words from Brown Corpus, manually tagged with WordNet senses
    - SENSEVAL-3 competition corpora - 2081 tagged word tokens

WSD: Semantic relatedness and word sense disambiguation



Source: Proceedings of the 20th International Conference on Advanced Information Networking and Applications

Example of SemCor with wordnet sense numbers

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

- Given each noun, verb, adjective, or adverb word in the hand-labeled test set. Ex: fruit^1_n (the ripened reproductive body of a seed plant), and the other two senses fruit^2_n (yield;an amount of a product) and fruit^3_n (the consequence of some effort or action).

All word WSD task

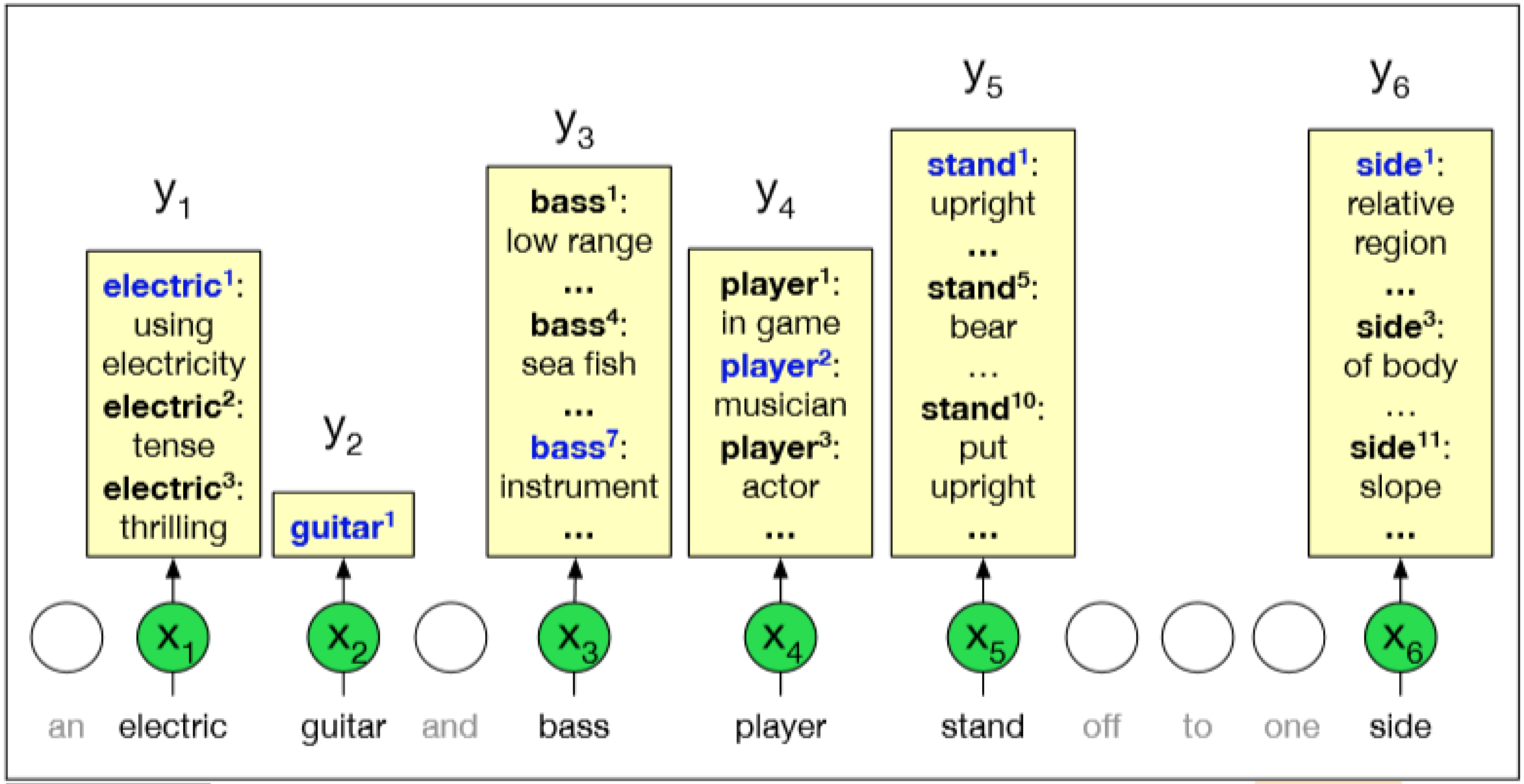


Figure 19.8 The all-words WSD task, mapping from input words (x) to WordNet senses (y). Only nouns, verbs, adjectives, and adverbs are mapped, and note that some words (like *guitar* in the example) only have one sense in WordNet. Figure inspired by Chaplot and Salakhutdinov (2018).

WSD:The Task and Datasets

- Surprisingly strong baseline is simply to choose the **most frequent sense** for most frequent sense each word from the senses in a labeled corpus(Galeetal.,1992a).
- For WordNet, this corresponds to the first sense, since senses in WordNet are generally ordered from most frequent to least frequent
- Most frequent sense baseline can be quite accurate, and is therefore often used as a default, to supply a word sense when a supervised algorithm **has insufficient training data**
- Another heuristic **one sense per discourse**: Word appearing multiple times in a text or discourse often appears with the same sense

WSD algorithms

- Supervised Learning –
  - Feature-Based WSD
- LESK Algorithm
- Word-in-Context Evaluation
- Wikipedia as a source of training data

Supervised Machine Learning Approaches

- Supervised machine learning approach:
  - a training corpus of words tagged in context with their sense
  - used to train a classifier that can tag words in new text
- Summary of what we need:
  - the **tag set** (“sense inventory”)
  - the **training corpus**
  - A set of **features** extracted from the training corpus
  - A **classifier**

Supervised Learning

Uses an SVM classifier to choose the sense for each input word with the following simple features of the surrounding words:

- part-of-speech tags (for a window of 3 words on each side, stopping at sentence boundaries)
- collocation features of words or n-grams of lengths 1, 2, 3) at a particular collocation location in a window of 3 word on each side (i.e., exactly one word to the right, or the two words starting 3 words to the left, and so on).
- weighted average of embeddings (of all words in a window of 10 words on each side, weighted exponentially by distance)

Extract feature vectors

- A simple representation for each observation (each instance of a target word)
  - Vectors of sets of feature/value pairs
    - I.e. files of comma-separated values
  - These vectors should represent the window of words around the target

Two kinds of features in the vectors

- **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

Examples

- Example text (WSJ)

  An electric guitar and **bass** player stand off to one side not really part of the scene, just as a sort of nod to gringo expectations perhaps

  – Assume a window of +/- 2 from the target

Examples

- Example text

  An electric guitar and **bass** player stand off to one side not really part of the scene, just as a sort of nod to gringo expectations perhaps

  – Assume a window of +/- 2 from the target

Collocational

- Position-specific information about the words in the window
- guitar and bass player stand
  - [guitar, NN, and, CC, player, NN, stand, VB]
  - $Word_{n-2,}$ $POS_{n-2,}$ $word_{n-1,}$ $POS_{n-1,}$ $Word_{n+1}$ POS_{n+1}…
  - In other words, a vector consisting of
  - [position n word, position n part-of-speech…]

Collocational

- An electric guitar and bass player stand off to one side, If we used as mall 2-word window, a standard feature vector might include parts-of speech, unigram and bigram collocation features, and a weighted sum of embeddings, that is:
- **[w_{i-2},POS_{i-2},w_{i-1},POS_{i-1},w_{i+1},POS_{i+1},wi_{i+2},POS_{i+2}, $w_{i-2}^{i-1}, w_{i+1}^{i+2}$,,g(E($w_{i-2}$),E($w_{i-1}$),E($w_{i+1}$),E($w_{i+2}$)]**

would yield the following vector:

[guitar, NN, and, CC, player, NN, stand, VB, and guitar, player stand, g(E(guitar) ,E(and), E(player), E(stand))]

Bag-of-words

- Words that occur within the window, regardless of specific position
- First derive a set of terms to place in the vector
- Then note how often each of those terms occurs in a given window

Bag of Words representation

- A very popular and basic representation of documents is the bag of words model.
- Each document is represented by a bag (= multiset) of terms from a predefined vocabulary.

Bag of Words representation

The Jackal was eyeing at the grapes

He was as cunning as a Jackal

These Grapes are too sweet but the poor Jackal could not have it.

1	**2**	**0**		**1**			**1**		**1**				**1**								
a	as	at		Cunning			he		was				jackal								

Co-Occurrence Example

- Assume we've settled on a possible vocabulary of 12 words that includes guitar and player but not and and stand

[*fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band*]

- The vector for:
  - guitar and bass player stand
  - [0,0,0,1,0,0,0,0,0,0,1,0]

Supervised Learning Algorithm

Input:

- a word *w in a text window d (which we'll call a "document")*
- a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
- A training set of *m* hand-labeled text windows again called "documents" $(d_1,c_1),\ldots,(d_m,c_m)$

Output:

- a learned classifier $\gamma: d \rightarrow c$

Applying Naïve Bayes Classifier

P(c) is the prior probability of that sense

- Counting in a labeled training set.

P(w|c) conditional probability of a word given a particular sense

- P(w|c) = count(w,c)/count(c)

We get both of these from a tagged corpus like SemCor

Can also generalize to look at other features besides words.

- Then it would be P(f|c)
  - Conditional probability of a feature given a sense

Applying Naïve Bayes Classifier

Dan Jurafsky

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	fish smoked fish	f
	2	fish line	f
	3	fish haul smoked	f
	4	guitar jazz line	g
Test	5	line guitar jazz jazz	?

V = {fish, smoked, line, haul, guitar, jazz}

Priors:

$P(f) = \frac{3}{4}$

$P(g) = \frac{1}{4}$

Choosing a class:

$P(f|d5) \propto 3/4 * 2/14 * (1/14)^2 * 1/14$
≈ 0.00003

$P(g|d5) \propto 1/4 * 2/9 * (2/9)^2 * 2/9$
≈ 0.0006

Conditional Probabilities:

P(line|*f*) = (1+1) / (8+6) = 2/14
P(guitar|*f*) = (0+1) / (8+6) = 1/14
P(jazz|*f*) = (0+1) / (8+6) = 1/14
P(line|*g*) = (1+1) / (3+6) = 2/9
P(guitar|*g*) = (1+1) / (3+6) = 2/9
P(jazz|*g*) = (1+1) / (3+6) = 2/9

The WSD Algorithm: Simple 1-nearest-neighbor algorithm

- Best-performing WSD algorithm is a simple 1-nearest-neighbor algorithm using contextual word embedding's
- For each token c_i of each sense c of each word, we average the contextual representations to produce a contextual **sense embedding $\mathbf{v_s}$** for c

$$\mathbf{v}_s = \frac{1}{n}\sum_i \mathbf{c}_i$$

- At test time we similarly compute a contextual embedding t for the target word, and choose its nearest neighbor sense (the sense with the highest cosine with t) from the training set.

An important idea in linguistics is that words (or expressions) that can be used in similar ways are likely to have related meanings.

Contextual Embedding

- **Word embedding** is the collective name for a set of language modeling and feature learning techniques in NLP where words or phrases from the vocabulary are mapped to vectors of real numbers
- Intuition of embedding models like word2vec or GloVe is that the meaning of a word can be defined by its co-occurrences, the counts of words that often occur nearby.
- But doesn't tell us how to define the meaning of a word
- Contextual embedding's like ELMo or BERT go further by offering an embedding that represents the meaning of a word in its textual context, and we'll see that contextual embedding's lie at the heart of modern algorithms for word sense disambiguation

Contextual Embedding

context words	v(astronomers)	v(bodies)	v(objects)
’t			1
,		2	1
.	1		1
1			1
And			1
Belt			1
But	1		
Given			1
Kuiper			1
So	1		
and		1	
are		2	1
between			1
beyond		1	
can			1
contains		1	
from	1		
hypothetical			1
ice		1	
including		1	
is	1		
larger		1	
now	1		
of	1		
only			1
out		1	
potential		1	
the	1		1
these		2	1
they	1		
think	2		
those			1
thought		2	
what	1		

$$\text{cosine\_similarity}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|}$$

	astronomers	bodies	objects
astronomers	$\frac{14}{\sqrt{14} \cdot \sqrt{14}} = 1$	$\frac{0}{\sqrt{24} \cdot \sqrt{14}} = 0$	$\frac{1+1}{\sqrt{14} \cdot \sqrt{16}} \approx 0.134$
bodies		$\frac{24}{\sqrt{24} \cdot \sqrt{24}} = 1$	$\frac{2+2+2}{\sqrt{24} \cdot \sqrt{16}} \approx 0.306$
objects			$\frac{16}{\sqrt{16} \cdot \sqrt{16}} = 1$

Nearest-neighbor algorithm for WSD

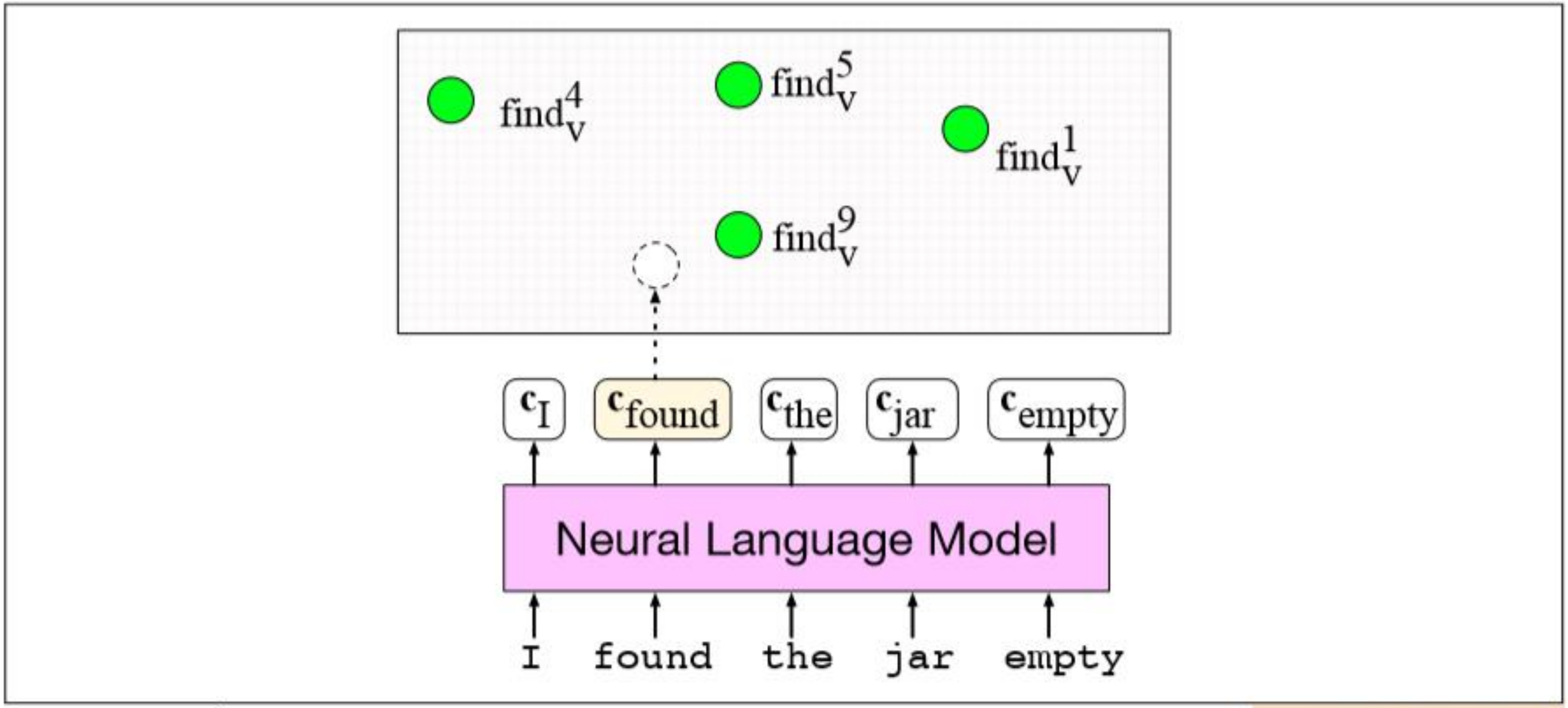


Figure 19.9 The nearest-neighbor algorithm for WSD. In green are the contextual embeddings precomputed for each sense of each word; here we just show a few of the senses for *find*. A contextual embedding is computed for the target word *found*, and the and then the nearest neighbor sense (in this case $\mathbf{find}_n^9$) would be chosen. Figure inspired by Loureiro and Jorge (2019).

The Lesk Algorithm as WSD Baseline

- Knowledge-based algorithms, rely solely on knowledge based WordNet or other such resources and don't require labeled data.
- While supervised algorithms generally work better, knowledge-based methods can be used in languages or domains where thesauruses or dictionaries but not sense labeled corpora are available.
- Lesk algorithm is the most powerful knowledge-based WSD algorithm
- Lesk is really a family of algorithms that choose the sense whose dictionary gloss or definition shares the most words with the target word's neighborhood

Lesk Algorithm Example

- Consider disambiguating the word bank in the following context:

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

- *Wordnet senses of "bank":*

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 bank1 has two non-stopwords overlapping with the context in deposits and mortgage, while sense bank2 has zero words, so bank1 is chosen.

Lesk Algorithm

```
function SIMPLIFIED LESK(word, sentence) returns best sense of word

  best-sense ← most frequent sense for word
  max-overlap ← 0
  context ← set of words in sentence
  for each sense in senses of word do
    signature ← set of words in the gloss and examples of sense
    overlap ← COMPUTEOVERLAP(signature, context)
    if overlap > max-overlap then
          max-overlap ← overlap
          best-sense ← sense
  end
  return(best-sense)
```

Figure 19.10 The Simplified Lesk algorithm. The COMPUTEOVERLAP function returns the number of words in common between two sets, ignoring function words or other words on a stop list. The original Lesk algorithm defines the *context* in a more complex way.

Corpus Lesk Algorithm

- Assumes we have some sense-labeled data (like SemCor)
- Take all the sentences with the relevant word sense:

  These short, "streamlined" meetings usually are sponsored by local ***banks***^{1}*, Chambers of Commerce, trade associations, or other civic organizations.*

- Now add these to the gloss + examples for each sense, call it the "signature" of a sense.
- Choose sense with most word overlap between context and signature.

Wikipedia as a source of training data

- Concept is mentioned in a Wikipedia: article text may contain an explicit link to the concept's Wikipedia page, which is named by a unique identifier (can be used as a sense annotation)
- For example, BAR (LAW), the page BAR (MUSIC), and so on, as in the following Wikipedia
  - *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]].*
- These sentences can then be added to the training data for a supervised system.

Wikipedia as a source of training data

- It is necessary to map from Wikipedia concepts to whatever inventory of senses is relevant for the WSD application.
- Automatic algorithms that map from Wikipedia to WordNet
- Ex: involve finding the WordNet sense that has the greatest lexical overlap with the Wikipedia sense, by comparing the vector of words in the WordNet synset, gloss, and related senses with the vector of words in the Wikipedia page title, outgoing links, and page category
- The resulting mapping has been used to create BabelNet, a large sense-annotated resource.

Using Thesauruses to Improve Embeddings

- Thesauruses have also been used to improve both static and contextual word embeddings.
- For example, static word embeddings have a problem with antonyms.
- A word like expensive is often very similar in embedding cosine to its antonym like cheap.

Unsupervised Learning: Word Sense Induction

- Expensive and difficult to build large corpora in which each word is labeled for its word sense
- Word sense induction or WSI, is an important direction. In word sense induction unsupervised approaches, we don’t use human-defined word senses.
- Instead, the set of “senses” of each word is created automatically from the instances of each word in the training set.

Word Sense Induction

In training, we use three steps:

- For each token w_i of word *w* in a corpus, compute a context vector **c**
- Use a **clustering algorithm** to **cluster** these word-token context vectors c in to a predefined number of groups or clusters. Each cluster defines a sense of *w*.
- Compute the **vector centroid** of each cluster. Each vector centroid **sj** is a **sense vector** representing that sense of w.
- We don’t have names for each of these “senses” of w; we just refer to the jth sense of w.

Word Sense Induction

To disambiguate a particular token t of w we again have three steps:

1. Compute a context vector c for t.
2. Retrieve all sense vectors sj for w.
3. Assign t to the sense represented by the sense vector sj that is closest to t.

All we need is a clustering algorithm and a distance metric between vectors. Ex: Agglomerative clustering

What we covered in todays session

- Word sense is the locus of word meaning; definitions and meaning relations are defined at the level of the word sense rather than word forms.
- Many words are polysemous, having many senses.
- Relations between senses include synonymy, antonymy, meronymy, and taxonomic relations hyponymy and hypernymy.
- WordNet is a large database of lexical relations for English, and exist for a variety of languages.
- WSD is the task of determining the correct sense of a word in context.

What we covered in todays session

- Supervised approaches make use of a corpus of sentences in which individual words (lexical sample task) or all words (all-words task) are hand-labeled with senses from a resource like WordNet.
- SemCor is the largest corpus with WordNet-labeled senses.
- The standard supervised algorithm for WSD is nearest neighbors with contextual embeddings.
- Feature-based algorithms using parts of speech and embeddings of words in the context of the target word also work well

What we covered in todays session

- An important baseline for WSD is the most frequent sense, equivalent, in WordNet, to take the first sense.
- Another baseline is a knowledge-based WSD algorithm called the Lesk algorithm which chooses the sense whose dictionary definition shares the most words with the target word's neighborhood.
- Word sense induction is the task of learning word senses using unsupervised learning

References

- https://wordnet.princeton.edu/
- https://babelnet.org/
- https://aclanthology.org/2022.coling-1.368.pdf
- https://aclanthology.org/2022.wildre-1.4.pdf
- https://www.sciencedirect.com/science/article/abs/pii/S0885230821001303

Thank You