

CMPT-413

Computational Linguistics

Anoop Sarkar

<http://www.cs.sfu.ca/~anoop>

Lexical Semantics

- So far, we have listed **words** in our **lexicon** or **vocabulary** assuming a single meaning per word:

Consider n -grams $P(w_i | w_{i-2}, w_{i-1}) = P(\text{Bank} | \text{on}, \text{Commerce})$ or prepositional phrase attachment if $p=\text{on}$ and $n2=\text{bank}$ then change N to V

- Consider ... *withdraw twenty dollars on the bank* (correct = V) vs. ... *withdraw the troops on the bank* (correct = N)
- The same word *bank* means two different things but we cannot distinguish between them using the traditional definition of word.

Lexical Semantics

- To deal with this issue, we combine the *spelling* or *pronunciation* of a word and the *meaning*.
In the *lexicon* we now store **lexemes** instead of words. A lexeme pairs a particular spelling or pronunciation with a particular meaning.
- The meaning part of a lexeme is called a **sense**. For CL, our interest is in relations between lexemes or disambiguating different senses of a word.
word: bank → lexeme: **bank**¹ OR word: bank → lexeme: **bank**²
- Note that meanings are often not definitions, but often are simple listings of compatible lexemes.
cf. dictionary defns: *red*, *n.* the color of blood or ruby; *blood*, *n.* red liquid circulating in animals

Homonyms

- Homonyms: *words that have the same form but different meanings*
 1. *Instead, the chemical plant was found in violation of several environmental laws*
 2. *Stanley formed an expedition to find a rare plant found along the Amazon river*
- Same *orthographic* form: *plant* but two senses: **plant**¹ and **plant**²

Homonyms

- Text vs. speech: fly-casting for *bass* vs. rhythmic *bass* chords
These cases are homonyms in text, but not in speech. Referred to as **homographs**
- Speech vs. text: *would* vs. *wood*
These cases are not homonyms in text, but easily confused in speech. Referred to as **homophones**
- Note that this problem in some cases can be solved using *part of speech tagging*
Can you think of a case which cannot be solved using POS tagging?

Applications

- Spelling correction: homophones: *weather* vs. *whether*
- Speech recognition: homophones: *to*, *two*, *too*. Also homonyms (see *n*-gram e.g.)
- Text to speech: homographs: *bass* vs. *bass*
- Information retrieval: homonyms: *latex*

Polysemy

- Consider the homonym: *bank* → commercial **bank**¹ vs. river **bank**²
- Now consider
 1. *A PCFG can be trained using derivation trees from a tree bank annotated by human experts*
- Is this a new sense of *bank*?

Polysemy

- Senses can be derived from a particular lexeme. This process is known as **polysemy**

In previous case we would say that the use of *bank* is a sense derived from commercial **bank**¹

- In some cases, splitting into different lexemes has other supporting evidence: **bank**¹ has Italian origin vs. **bank**² has Scandinavian origin

1. *A PCFG can be trained using a bank of derivation trees called a tree-bank annotated by human experts*

- How can we tell between homonyms and polysemous uses of a word?

Zeugma

- Consider the case for a verb like *serve*
 1. *Does United serve breakfast?*
 2. *Does United serve Philadelphia?*
 3. *Does United serve breakfast and dinner?*
 4. *#Does United serve breakfast and Philadelphia?*

Word Sense Disambiguation

- Consider a noun like *bank*
 1. *How many senses does it have?*
 2. *How are these senses related?*
 3. *How can they be reliably distinguished?*
- For NLP software, among these three questions, typically at runtime we need to automatically find the answer to the last question: given a word in context, map it to the correct lexeme: **word-sense disambiguation**

Word Sense Disambiguation: training data

training_VBG new_JJ Ukrainian_JJ
who_WP are_VBP leaving_VBG the_DT
CC safety_NN procedures_NNS at_IN
t_IN the_DT Orange_NNP County_NNP
Z closing_VBG three_CD missile_NN
_IN the_DT whole_JJ Chernobyl_NNP
IN a_DT hill_NN ,_, gardeners_NNS
\$_\$ 200_CD million_CD printing_NN
of_IN incompletely_JJ oxidated_JJ
whenever_WRB you_PRP eat_VBP a_DT
n_IN return_NN for_IN a_DT new_JJ
T carmaker_NN could_MD finance_VB
n_IN return_NN for_IN a_DT new_JJ

```
plant(1)
plant(1)
plant(1)
plant(1)
plant(1)
plant(1)
plant(2)
plant(1)
plant(2)
plant(2)
plant(1)
plant(1)
plant(1)
```

_NN operators_NNS to_TO replace_V
s_NNS in_IN Ukraine_NNP and_CC im
s_NNS in_IN both_DT countries_NNS
_NN __.
s_NNS in_IN southern_JJ Californi
_NN in_IN 1991_CD __, five_CD yea
_NN begonias_NNS __, making_VBG f
_NN in_IN Brooklyn_NNP __, Ohio_N
_NN and_CC animal_NN sediment_NN
_NN __. ' '_'
_NN near_IN Tuscaloosa_NNP __.
_NN construction_NN with_IN the_D
_NN near_IN Tuscaloosa_NNP __.

Word Sense Disambiguation: learning

- Many different learning methods: let's consider one, Transformation Based Learning
- Let rule condition
 $r \leftarrow W_{-1} = \text{gardeners}, W_{+1} = \text{begonias}, W_{+\text{window}} = \text{floral}$
- If r then change from **plant**¹ (manufacturing plant) to **plant**² (living plant)

Synonyms

- Synonyms: Different lexemes with the same meaning

1. *How big/large is that plane?*

2. *Would I be flying on a big/large or small plane?*

- Synonyms clash with polysemous meanings

1. *Seema is my big sister*

2. *#Seema is my large sister*

WordNet

- WordNet is an electronic database of word relationships, handcrafted from scratch by researchers at Princeton University (George Miller, Christine Fellbaum, et al.)
- WordNet contains 3 databases: for verbs, nouns and one for adjectives and adverbs

Category	Unique Forms	Number of Senses
Noun	94474	116317
Verb	10319	22066
Adjective	20170	29881
Adverb	4546	5677

WordNet

- Ask the question: how many senses per noun or verb? The distribution of senses follows Zipf's (2nd) Law.
- WordNet provides multiple lexeme entries for each word and for each part of speech,
e.g. *plant* as noun has 3 senses; *plant* as verb has 2 senses
- WordNet also provides *domain-independent* lexical relations such as IS-A, HasMember, MemberOf, ...

WordNet: noun relations

Relation	Definition	Example
Hypernym	this is a kind of	<i>breakfast</i> → <i>meal</i>
Hyponym	this has a specific instance	<i>meal</i> → <i>lunch</i>
Has-Member	this has a member	<i>faculty</i> → <i>professor</i>
Member-Of	this is member of a group	<i>copilot</i> → <i>crew</i>
Has-Part	this has a part	<i>table</i> → <i>leg</i>
Part-Of	this is part of	<i>course</i> → <i>meal</i>
Antonym	this is an opposite of	<i>leader</i> → <i>follower</i>

WordNet: verb relations

Relation	Definition	Example
Hypernym	this event is a kind of	<i>fly</i> → <i>travel</i>
Tropynym	this event has a subtype	<i>walk</i> → <i>stroll</i>
Entails	this event entails	<i>snore</i> → <i>sleep</i>
Antonym	this event is opposite of	<i>increase</i> → <i>decrease</i>

WordNet: example from ver1.7.1

Sense1: Canada

⇒North American country,North American nation

⇒country, state, land

⇒administrative district,administrative division,territorial division

⇒district, territory

⇒region

⇒location

⇒entity, physical thing

WordNet: example from ver1.7.1

Sense 3: Vancouver

- ⇒city, metropolis, urban center
 - ⇒municipality
 - ⇒urban area
 - ⇒geographical area
 - ⇒region
 - ⇒location
 - ⇒entity, physical thing
 - ⇒administrative district, territorial division
 - ⇒district, territory
 - ⇒region
 - ⇒location
 - ⇒entity, physical thing
- ⇒port
 - ⇒geographic point
 - ⇒point
 - ⇒location
 - ⇒entity, physical thing

WordNet

- A **synset** in WordNet is a list of synonyms (interchangeable words)
- { chump, fish, fool, gull, mark, patsy, fall guy, sucker, schlemiel, shlemiel, soft touch, mug }
- How can we use this information like synsets, hypernyms, etc. from WordNet to benefit NLP applications?
- Consider one example: PP attachment, words plus word classes extracted from the hypernym hierarchy increase accuracy from 84% to 88% (Stetina and Nagao, 1998)

WordNet

- Another example of WordNet used in NLP applications: **selectional restrictions**
- We have considered subcategorization:
VP-with-NP-complement → *V(eat) NP* “eat six bowls of rice ”
But not selectional restrictions of the verb itself: “ *eat tomorrow* ”
Consider *what do you want to eat tomorrow*
- We can use the **synset** { food, nutrient } to describe the NP argument of *eat* – then the 60K lexemes under these nodes in the WordNet hierarchy will be acceptable.
(however, what about “ *eat my shorts* ”)
→ several other applications have been explored