Words & their Meaning: Word Sense Disambiguation

CMSC 470
Marine Carpuat
Today: Word Meaning

2 core issues from an NLP perspective

• **Semantic similarity**: given two words, how similar are they in meaning?

• **Word sense disambiguation**: given a word that has more than one meaning, which one is used in a specific context?
“Big rig carrying fruit crashes on 210 Freeway, creates jam”

http://articles.latimes.com/2013/may/20/local/la-me-ln-big-rig-crash-20130520
How do we know that a word (lemma) has distinct senses?

• Linguists often design tests for this purpose

• e.g., zeugma combines distinct senses in an uncomfortable way

Which flight serves breakfast?

Which flights serve BWI?

*Which flights serve breakfast and BWI?
Word Senses

• “Word sense” = distinct meaning of a word

• Same word, different senses
  • **Homonyms** (homonymy): unrelated senses; identical orthographic form is coincidental
    • E.g., financial bank vs. river bank
  • **Polysemes** (polysemy): related, but distinct senses
    • E.g., Financial bank vs. blood bank vs. tree bank
  • **Metonyms** (metonymy): “stand in”, technically, a sub-case of polysemy
    • E.g., use “Washington” in place of “the US government”

• Different word, same sense
  • **Synonyms** (synonymy)
WordNet: a lexical database for English

https://wordnet.princeton.edu/

- Includes most English nouns, verbs, adjectives, adverbs
- Electronic format makes it amenable to automatic manipulation: used in many NLP applications
- “WordNets” generically refers to similar resources in other languages
Synonymy in WordNet

• WordNet is organized in terms of “synsets”
  • Unordered set of (roughly) synonymous “words” (or multi-word phrases)

• Each synset expresses a distinct meaning/concept
WordNet: Example

Noun
{pipe, tobacco pipe} (a tube with a small bowl at one end; used for smoking tobacco)
{pipe, pipage, piping} (a long tube made of metal or plastic that is used to carry water or oil or gas etc.)
{pipe, tube} (a hollow cylindrical shape)
{pipe} (a tubular wind instrument)
{organ pipe, pipe, pipework} (the flues and stops on a pipe organ)

Verb
{shriek, shrill, pipe up, pipe} (utter a shrill cry)
{pipe} (transport by pipeline) “pipe oil, water, and gas into the desert”
{pipe} (play on a pipe) “pipe a tune”
{pipe} (trim with piping) “pipe the skirt”
WordNet 3.0: Size

<table>
<thead>
<tr>
<th>Part of speech</th>
<th>Word form</th>
<th>Synsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>117,798</td>
<td>82,115</td>
</tr>
<tr>
<td>Verb</td>
<td>11,529</td>
<td>13,767</td>
</tr>
<tr>
<td>Adjective</td>
<td>21,479</td>
<td>18,156</td>
</tr>
<tr>
<td>Adverb</td>
<td>4,481</td>
<td>3,621</td>
</tr>
<tr>
<td>Total</td>
<td>155,287</td>
<td>117,659</td>
</tr>
</tbody>
</table>
Different inventories can be used to define senses

<table>
<thead>
<tr>
<th>WordNet Sense</th>
<th>Spanish Translation</th>
<th>Roget Category</th>
<th>Target Word in Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>bass⁴</td>
<td>lubina</td>
<td>FISH/INSECT</td>
<td>...fish as Pacific salmon and striped <strong>bass</strong> and...</td>
</tr>
<tr>
<td>bass⁴</td>
<td>lubina</td>
<td>FISH/INSECT</td>
<td>...produce filets of smoked <strong>bass</strong> or sturgeon...</td>
</tr>
<tr>
<td>bass⁷</td>
<td>bajo</td>
<td>MUSIC</td>
<td>...exciting jazz <strong>bass</strong> player since Ray Brown...</td>
</tr>
<tr>
<td>bass⁷</td>
<td>bajo</td>
<td>MUSIC</td>
<td>...play <strong>bass</strong> because he doesn’t have to solo...</td>
</tr>
</tbody>
</table>

Different inventories do not always agree on sense distinctions
e.g., translation makes some distinctions but not others
Exercise: how many senses of “drive”? 

1. "Can you drive this four-wheel truck?"
2. "We drive to the university every morning"
3. "We drive the car to the garage"
4. "He drives me mad"
5. "She is driven by her passion"
6. "Drive a nail into the wall"
7. "She is driving away at her doctoral thesis"
8. "What are you driving at?"
9. "My new truck drives well"
10. "She drives for the taxi company in Newark"
11. "drive the cows into the barn"
12. "We drive the turnpike to work"
13. "drive a golf ball"
Exercise: how many senses of “drive”?

1. "Can you drive this four-wheel truck?"
2. "We drive to the university every morning"
3. "We drive the car to the garage"
4. "He drives me mad"
5. "She is driven by her passion"
6. "Drive a nail into the wall"
7. "She is driving away at her doctoral thesis"
8. "What are you driving at?"
9. "My new truck drives well"
10. "She drives for the taxi company in Newark"
11. "drive the cows into the barn"
12. "We drive the turnpike to work"
13. "drive a golf ball"

13 distinct senses according to WordNet!
Exercise: how many senses of “drive”?  

1. "We drive to the university every morning" (operate or control a vehicle)  
2. "We drive the car to the garage" (cause someone or something to move by driving)  
3. "He drives me mad" (force into or from an action or state, either physically or metaphorically)  
4. "She is driven by her passion" (to compel or force or urge relentlessly or exert coercive pressure on, or motivate strongly)  
5. "Drive a nail into the wall" (push, propel, or press with force)  
6. "She is driving away at her doctoral thesis" (strive and make an effort to reach a goal)  
7. "What are you driving at?" (move into a desired direction of discourse)  
8. "My new truck drives well" (have certain properties when driven)  
9. "She drives for the taxi company in Newark" (work as a driver)  
10. "drive the cows into the barn" (urge forward)  
11. "We drive the turnpike to work" (proceed along in a vehicle)  
12. "drive a golf ball" (strike with a driver, as in teeing off)  

13 distinct senses according to WordNet!
What can we do when humans who annotate senses disagree?

- Disagreement is inevitable when annotating based on human judgments
  - Even with trained annotators
  - There is no “ground truth”

- We cannot measure “correctness” of annotations directly

- Instead, we can measure reliability of annotation
  - Do human annotators make same decisions consistently?
  - Assumption: high reliability implies validity
Quantifying (dis)agreement between human annotators: Cohen’s Kappa

• Measures agreement between two annotators while taking into account the possibility of chance agreement

\[ K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \]

• Scales for interpreting Kappa

Landis & Koch, 1977

Green, 1997
Quantifying (dis)agreement between human annotators: Cohen’s Kappa

Consider this confusion matrix for sense annotations by A and B of the same 250 examples:

<table>
<thead>
<tr>
<th></th>
<th>Sense 1</th>
<th>Sense 2</th>
<th>Sense 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sense 1</td>
<td>54</td>
<td>28</td>
<td>3</td>
<td>85</td>
</tr>
<tr>
<td>Sense 2</td>
<td>31</td>
<td>18</td>
<td>23</td>
<td>72</td>
</tr>
<tr>
<td>Sense 3</td>
<td>0</td>
<td>21</td>
<td>72</td>
<td>93</td>
</tr>
<tr>
<td>Total</td>
<td>85</td>
<td>67</td>
<td>98</td>
<td>250</td>
</tr>
</tbody>
</table>

Here $Pr(a) = 0.576$, $Pr(e) = 0.403$, $K=0.29$ (agreement is low)
Word Sense Disambiguation
what you should know (so far)

• Word senses distinguish different meanings of same word
• Sense inventories provide definitions of word senses
• Sense distinctions and annotations are based on human judgment
  • no “ground truth”
  • Measure annotation reliability using inter-annotator agreement
Word Sense Disambiguation

• Computational task
  • Given a predefined sense inventory (e.g., WordNet)
  • Goal: automatically select the correct sense of a word
  • Input: a word in context
  • Output: sense of the word

• Motivated by many applications:
  • Information retrieval
  • Machine translation
  • ...
How hard is the problem?

- **Most words in English have only one sense**
  - 62% in Longman’s Dictionary of Contemporary English
  - 79% in WordNet

- **But the others tend to have several senses**
  - Average of 3.83 in LDOCE
  - Average of 2.96 in WordNet

- **Ambiguous words are more frequently used**
  - In the British National Corpus, 84% of instances have more than one sense

- **Some senses are more frequent than others**
Baseline Performance

- **Baseline**: most frequent sense
  - Equivalent to “take first sense” in WordNet
  - Does surprisingly well!

<table>
<thead>
<tr>
<th>Freq</th>
<th>Synset</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>338</td>
<td>plant⁴, works, industrial plant</td>
<td>buildings for carrying on industrial labor</td>
</tr>
<tr>
<td>207</td>
<td>plant², flora, plant life</td>
<td>a living organism lacking the power of locomotion</td>
</tr>
<tr>
<td>2</td>
<td>plant³</td>
<td>something planted secretly for discovery by another</td>
</tr>
<tr>
<td>0</td>
<td>plant⁴</td>
<td>an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience</td>
</tr>
</tbody>
</table>

62% accuracy in this case!
Upper Bound Performance

• Upper bound
  • Fine-grained WordNet sense: 75-80% human agreement
  • Coarser-grained inventories: 90% human agreement possible
Simplest WSD algorithm: Lesk’s Algorithm

• Intuition: note word overlap between context and dictionary entries
  • Unsupervised, but knowledge rich

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

| bank¹ | 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” |
| bank² | 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” |
Lesk’s Algorithm

• Simplest implementation:
  • Count overlapping content words between glosses and context

• Lots of variants:
  • Include the examples in dictionary definitions
  • Include hypernyms and hyponyms
  • Give more weight to larger overlaps
  • Give extra weight to infrequent words (e.g., using idf)
  • ...
Alternative: WSD as **Supervised** Classification

![Diagram showing supervised learning process]

- Training data
- Feature Functions
- Classifier
- Testing
- Supervised machine learning algorithm
- Unlabeled document

**Label** 1, 2, 3, 4
Existing Corpora

• Lexical sample
  • *line-hard-serve* corpus (4k sense-tagged examples)
  • *interest corpus* (2,369 sense-tagged examples)
  • ...

• All-words
  • SemCor (234k words, subset of Brown Corpus)
  • Senseval/SemEval (2081 tagged content words from 5k total words)
  • ...
How are annotated examples used in supervised learning?

• Supervised learning = requires examples annotated with correct prediction

• Used in 2 ways:
  • To find good values for the model (hyper)parameters (training data)
  • To evaluate how good the resulting classifier is (test data)

• How do we know how good a classifier is?
  • Compare classifier predictions with human annotation
  • On held out test examples
  • Evaluation metrics: accuracy, precision, recall
The 2-by-2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>
Precision and recall

- **Precision**: % of selected items that are correct
- **Recall**: % of correct items that are selected

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>
A combined measure: F

• A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

\[ F = \frac{1}{\frac{1}{P} + (1 + \frac{1}{R})} = \frac{(2 + 1)PR}{2P + R} \]

• People usually use balanced F1 measure
  • i.e., with \( \beta = 1 \) (that is, \( \alpha = \frac{1}{2} \)):

\[ F = \frac{2PR}{P + R} \]
Multiclass Classification
Is this spam?

From: "Fabian Starr"
Patrick_Freeman@pamietaniepeereelu.pl
Subject: Hey! Sofware for the funny prices!

Get the great discounts on popular software today for PC and Macintosh
http://iiled.org/Cj4Lmx

70-90% Discounts from retail price!!!
All sofware is instantly available to download - No Need Wait!
What is the subject of this article?

**MEDLINE Article**

**MeSH Subject Category Hierarchy**

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

**SYNTACTIC FRAME AND VERB HIES IN APHASIA: PLACABILITY JUDGMENTS OF UNDERGEN-SUBJECT SENTENCES**

Rusenna Cold*; Lisa Marks; Carl S. Berntsen; Daniel S. Lunde; Elizabeth Elder; Molly Sarraga; and L. Holland Auberg

**Abstract**

The study investigated how referential information influences the acceptability of sentences in aphasia. Previous research has suggested that referential information can influence sentence processing in aphasia. The present study aimed to examine the role of referential information in the acceptability of sentences in aphasia. The study employed a within-subjects design, with participants divided into two groups: those with and without aphasia. The groups were matched on age, gender, and years of education. The participants were asked to judge the acceptability of sentences in a sentence completion task. The results indicated that referential information significantly influenced the acceptability of sentences in aphasia, with participants with aphasia showing a lower level of acceptability than those without aphasia. The findings suggest that referential information plays a crucial role in sentence processing in aphasia, highlighting the importance of referential information in sentence comprehension and production.
Text Classification

• Assigning subject categories, topics, or genres
• Spam detection
• Authorship identification
• Age/gender identification
• Language Identification
• Sentiment analysis
• ...
Word Sense Disambiguation
what you should know

• Word senses distinguish different meanings of same word
• Sense inventories
• Annotation issues and annotator agreement (Kappa)
• Definition of Word Sense Disambiguation Task
• An unsupervised approach: Lesk algorithm
• Supervised classification:
  • Train vs. test data
  • The most frequent class baseline
• Evaluation metrics: accuracy, precision, recall