CS474 Natural Language Processing

- Last class
  - Word sense disambiguation
    » Supervised machine learning methods
    » Issues for WSD evaluation
- Today
  - Word sense disambiguation
    » Weakly supervised (bootstrapping) methods
    » Unsupervised methods

Supervised ML framework

- Examples of task
  (features + class)
  description of context
  correct word sense
  ML Algorithm

  Novel example
  (features) Classifier (program) class
  learn one such classifier for each lexeme to be disambiguated

Weakly supervised approaches

- Problem: Supervised methods require a large sense-tagged training set
- Bootstrapping approaches: Rely on a small number of labeled seed instances

Generating initial seeds

- Hand label a small set of examples
  - Reasonable certainty that the seeds will be correct
  - Can choose prototypical examples
  - Reasonably easy to do
- One sense per collocation constraint (Yarowsky 1995)
  - Search for sentences containing words or phrases that are strongly associated with the target senses
  » Select fish as a reliable indicator of bass₁
  » Select play as a reliable indicator of bass₂
  - Or derive the collocations automatically from machine readable dictionary entries
  - Or select seeds automatically using collocational statistics (see Ch 6 of J&M)
One sense per collocation

Yarowsky's bootstrapping approach

Relies on a **one sense per discourse** constraint:
The sense of a target word is highly consistent
within any given document

- Evaluation on ~37,000 examples

<table>
<thead>
<tr>
<th>Word</th>
<th>Senses</th>
<th>Accuracy</th>
<th>Applicability</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>living/factory</td>
<td>99.8%</td>
<td>72.8%</td>
</tr>
<tr>
<td>tank</td>
<td>vehicle/container</td>
<td>99.6%</td>
<td>50.5%</td>
</tr>
<tr>
<td>poach</td>
<td>steal/boil</td>
<td>100.0%</td>
<td>44.4%</td>
</tr>
<tr>
<td>palm</td>
<td>tree/hand</td>
<td>99.8%</td>
<td>38.5%</td>
</tr>
<tr>
<td>axes</td>
<td>grid/tods</td>
<td>100.0%</td>
<td>36.5%</td>
</tr>
<tr>
<td>sake</td>
<td>benefit/drink</td>
<td>100.0%</td>
<td>33.7%</td>
</tr>
<tr>
<td>bass</td>
<td>fish/music</td>
<td>100.0%</td>
<td>58.8%</td>
</tr>
<tr>
<td>space</td>
<td>volume/outer</td>
<td>99.2%</td>
<td>67.7%</td>
</tr>
<tr>
<td>motion</td>
<td>legal/physical</td>
<td>99.9%</td>
<td>49.8%</td>
</tr>
<tr>
<td>crane</td>
<td>bird/machine</td>
<td>100.0%</td>
<td>48.1%</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td><strong>99.8%</strong></td>
<td><strong>56.1%</strong></td>
</tr>
</tbody>
</table>

Yarowsky's bootstrapping approach

To learn disambiguation rules for a polysemous word:

1. [Find all instances of the word in the training corpus and save the contexts around each instance.]

2. [For each word sense, identify a small set of training examples representative of that sense. Now we have a few labeled examples for each sense.]

3. Build a classifier (e.g., decision list) by training a supervised learning algorithm with the labeled examples.

4. Apply the classifier to all the unlabeled examples. Find instances that are classified with probability > a threshold and add them to the set of labeled examples.

5. Optional: Use the one-sense-per-discourse constraint to augment the new examples.

6. Go to Step 3. Repeat until the unlabelled data is stable.

CS474 Natural Language Processing

- **Last class**
  - Word sense disambiguation
    - Supervised machine learning methods
    - Issues for WSD evaluation

- **Today**
  - Word sense disambiguation
    - Weakly supervised (bootstrapping) methods
    - Unsupervised methods
### Unsupervised WSD

- Rely on **agglomerative clustering** to cluster feature-vector representations (without class/word-sense labels) according to a similarity metric
  - Represent each cluster as the average of its constituent feature-vectors
  - Label the cluster by hand with known word senses
  - Unseen feature-encoded instances are classified by assigning the word sense of the most similar cluster
- Schuetze (1992, 1998) uses a (complex) clustering method for WSD
  - For coarse binary decisions, unsupervised techniques can achieve results approaching those of supervised and bootstrapping methods
  - ...In most cases approaching the 90% range
  - Tested on a small sample of words

### Issues for evaluating clustering

- The correct senses of the instances used in the training data may not be known.
- The clusters are almost certainly heterogeneous w.r.t. the sense of the training instances contained within them.
- The number of clusters is almost always different from the number of senses of the target word being disambiguated.