Bayesian Semantics
verb sense induction

Daniel W. Peterson
University of Colorado

Presented at Heinrich Heine Universität Düsseldorf
Roadmap

Sense Induction as Clustering

Nonparametric Clustering

Verb Sense Mixture Model

VerbNet via Hierarchy
Polysemy at Work

The Rhein enters Düsseldorf from the south.

Do not enter the military lightly.

Many students enter university after secondary school.

John entered his essay in the competition.

Mary entered the classroom.

One does not simply enter Mordor.
Identified Senses

The Rhein \textbf{enters} Düsseldorf from the south.

Do not \textbf{enter} the military lightly.

Many students \textbf{enter} university after secondary school.

John \textbf{entered} his essay in the competition.

Mary \textbf{entered} the classroom.

One does not simply \textbf{enter} Mordor.
Sense Induction

Identify and label senses of a word in a corpus

Limited prior knowledge:

- No labeled corpus examples
- Unknown number of senses
This is Clustering!

Sense induction is joining corpus instances into sense clusters
Roadmap

Sense Induction as Clustering

Nonparametric Clustering

Verb Sense Mixture Model

VerbNet via Hierarchy
Polya's Urn

Add colored balls to an urn, one at a time

Select a color by drawing a ball, noting the color, and replacing it

Urn starts off with $\alpha$ black balls; if one is drawn, select a never-before-seen color

Add a new ball of the chosen color to the urn
Polya's Urn

Chance of selecting color $k$ given urn $U$ is

$$P(k|U) \propto \begin{cases} C(k, U), & \text{if } k \in U \\ \alpha, & \text{otherwise,} \end{cases}$$

(1)

where $C(k, U) = |\{x \in U : \text{color}(x) = k\}|$
Polya's Urn

Also called the “Chinese Restaurant Process”
Mathematically, no restriction on $\alpha$ to be integer-valued

Infinite: no upper bound on number of clusters

Conservative: strong “rich get richer” effect tends to use few clusters
Bayesian Clustering

To select a cluster (or color) $k$ for an item with evidence $X$, we compute

$$P(k|X, U) \propto P(k|U)P(X|k).$$  \hspace{1cm} (2)

The first factor, given by Equation 1, says how likely we are to use a particular clustering

The second factor encodes our intuitions about the data: does this $X$ look like it came from cluster $k$?
Gibbs Sampling

**Goal:** Find a complete clustering assignment that is reasonably good

**Problem:** Joint inference is difficult; combinatorially many possible clusterings, most of which are terrible

**Solution:**
- Update one variable at a time, with all others fixed
- After “burn-in” period, we are drawing from overall posterior
- Probabilistic steps mostly avoid local optima: start from anywhere
Roadmap

Sense Induction as Clustering

Nonparametric Clustering

Verb Sense Mixture Model

VerbNet via Hierarchy
Dependency parse gives syntactic context of verbs (subj, dobj, prep_with, etc.)

Shared context words suggest shared sense

Each cluster has a higher likelihood to generate its most frequent established arguments
Treat Senses like Topics

Each cluster is a multinomial distribution $\phi$ (over slot:token pairs)

Distributions drawn from Dirichlet prior with small parameter $\beta$

Probability of drawing an instance from the cluster, given observed counts, is

$$P(X|k) = \prod_{w \in X} \frac{C(w, k) + \beta}{C(\ast, k) + |V| \beta}$$  \hfill (3)
Figure: The proposed graphical model for sense induction. $G$ is the Dirichlet process, $k$ the selected cluster, $M$ the number of verb instances, and $N$ the number of slot:token items in the context.
A Few Useful Speedups

Combine terms with like arguments before clustering ("initial frames")

Discard initial frames with insufficient counts
Enter (sense 1)

nsubj <name>:60080, you:21941, we:13569, he:10760, they:9657, ...
dobj agreement:2768, it:2164, contract:1710, negotiation:1222, her:861, ...
prep_into agreement:24259, contract:13452, <name>:6780, relationship:5243, ...
prep_with <name>:8334, company:597, them:550, him:394, government:310, ...
prep_in <name>:3341, case:432, field:274, box:245, state:213, way:200, ...

...
Enter (sense 2)

nsubj <name>:32854, he:21010, you:13276, they:10276, i:9176, we:8683, ...

dobj <name>:119581, school:18595, college:6889, land:5714, ...

prep_in <name>:4983, fall:857, year:360, field:221, box:212, 1997:132, ...

prep_at <name>:4015, age:1709, time:522, end:226, point:168, level:137, ...

prep_on <name>:2507, day:305, visa:125, side:124, scholarship:91, ...

...
Enter (sense 3)

<table>
<thead>
<tr>
<th>Role</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsubj &lt;name&gt;</td>
<td>&lt;name&gt;:15048, you:12675, you:12675, they:9717, he:8676, we:7917, ...</td>
</tr>
<tr>
<td>dobj house</td>
<td>house:19541, building:16020, building:16020, home:13692, door:8491, ...</td>
</tr>
<tr>
<td>prep_through door</td>
<td>door:2726, gate:915, entrance:586, entrance:586, window:562, &lt;name&gt;:417, ...</td>
</tr>
<tr>
<td>prep_in &lt;name&gt;</td>
<td>&lt;name&gt;:1327, search:143, case:102, case:102, morning:89, middle:86, ...</td>
</tr>
<tr>
<td>prep_on &lt;name&gt;</td>
<td>&lt;name&gt;:1352, side:463, right:347, right:347, left:332, day:194, level:93, ...</td>
</tr>
</tbody>
</table>

...
Enter (sense 4)

<table>
<thead>
<tr>
<th>Role</th>
<th>Nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsubj</td>
<td>&lt;name&gt;:28116, he:11951, you:9188, i:8702, they:7945, she:7498, ...</td>
</tr>
<tr>
<td>dobj</td>
<td>room:62763, office:10113, store:6076, apartment:4150, kitchen:3353, ...</td>
</tr>
<tr>
<td>prep_with</td>
<td>&lt;name&gt;:605, tray:109, smile:91, gun:81, look:73, bag:66, air:65, ...</td>
</tr>
<tr>
<td>prep_in</td>
<td>&lt;name&gt;:794, time:175, room:146, hand:101, case:58, search:54, ...</td>
</tr>
<tr>
<td>prep_at</td>
<td>moment:522, &lt;name&gt;:464, time:357, point:179, end:177, age:85, night:71, ...</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Senses are fairly fine-grained (often duplicate one another)

Some noise: Gibbs sampling is always exploring
Roadmap

Sense Induction as Clustering

Nonparametric Clustering

Verb Sense Mixture Model

VerbNet via Hierarchy
Combining into Semantic Frames

Fine-grained senses for each verb, but no links to other verbs

Semantic frames can often be invoked from multiple verbs (e.g. “enter” and “join”)

Apply the clustering to our induced senses
Verb Clusters DPMM

Figure: The proposed graphical model for sense induction. $G$ is the Dirichlet process, $k$ the selected cluster, $M$ the number of verb senses, and $N$ the number of slot items in a sense.
VerbNet, in particular, is motivated by syntactic groupings of verbs

slot:token features used for sense induction are perhaps too specific

We may choose to use slot features only for either step, independently
## How well does this work?

<table>
<thead>
<tr>
<th>Settings</th>
<th>K</th>
<th>nmPU</th>
<th>niPU</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gigaword/S-S</td>
<td>272.8</td>
<td>63.46</td>
<td>67.66</td>
<td>65.49</td>
</tr>
<tr>
<td>Gigaword/S-SW</td>
<td>36.4</td>
<td>31.49</td>
<td>95.70</td>
<td>47.38</td>
</tr>
<tr>
<td>Gigaword/SW-S</td>
<td>186.2</td>
<td>63.52</td>
<td>64.18</td>
<td>63.84</td>
</tr>
<tr>
<td>Gigaword/SW-SW</td>
<td>30.0</td>
<td>36.27</td>
<td>94.66</td>
<td>52.40</td>
</tr>
<tr>
<td>Web/S-S</td>
<td>363.6</td>
<td>61.32</td>
<td>78.64</td>
<td>68.90</td>
</tr>
<tr>
<td>Web/S-SW</td>
<td>52.2</td>
<td>35.80</td>
<td><strong>99.30</strong></td>
<td>52.62</td>
</tr>
<tr>
<td>Web/SW-S</td>
<td>212.2</td>
<td><strong>66.26</strong></td>
<td>77.38</td>
<td><strong>71.39</strong></td>
</tr>
<tr>
<td>Web/SW-SW</td>
<td>55.0</td>
<td>36.70</td>
<td>96.25</td>
<td>53.13</td>
</tr>
</tbody>
</table>

**Table:** Evaluation on a small, polysemous verb clustering (Korhonen, 2003). K is the average number of induced classes. mPU is modified purity, iPU is inverse purity; both are normalized (n) to account for multiple assignments.
<table>
<thead>
<tr>
<th>Settings</th>
<th>K</th>
<th>mPU</th>
<th>iPU</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gigaword/S-NIL</td>
<td>-</td>
<td>93.43</td>
<td>20.06</td>
<td>33.03</td>
</tr>
<tr>
<td>Gigaword/SW-NIL</td>
<td>-</td>
<td>94.45</td>
<td>41.07</td>
<td>57.05</td>
</tr>
<tr>
<td>Gigaword/S-S</td>
<td>512.2</td>
<td>75.06</td>
<td>45.26</td>
<td>56.47</td>
</tr>
<tr>
<td>Gigaword/S-S</td>
<td>260.6</td>
<td>73.98</td>
<td>56.45</td>
<td>64.04</td>
</tr>
<tr>
<td>Web/S-NIL</td>
<td>-</td>
<td>93.70</td>
<td>32.96</td>
<td>48.78</td>
</tr>
<tr>
<td>Web/SW-NIL</td>
<td>-</td>
<td><strong>94.51</strong></td>
<td>44.95</td>
<td>60.92</td>
</tr>
<tr>
<td>Web/S-S</td>
<td>500.0</td>
<td>72.25</td>
<td>52.48</td>
<td>60.79</td>
</tr>
<tr>
<td>Web/SW-S</td>
<td>255.2</td>
<td>72.65</td>
<td><strong>61.00</strong></td>
<td><strong>66.31</strong></td>
</tr>
</tbody>
</table>

**Table:** Evaluation of direct, instance-level VerbNet alignment for SemLink corpus (WSJ 02-21). K, mPU, and iPU defined as before. NIL means we skipped the second stage of clustering.
Some places for extension

Add supervision (or partial supervision)

Incorporate semantic vectors

Label induced senses, so we can use accuracy instead of just clustering alignment
Adding Verb-Level Partial Supervision

We know about VerbNet class tendencies for some verbs

We can easily model an explicit VerbNet class assignment during sampling

Clusters should have only few VerbNet classes (we’ll draw them categorically with a Dirichlet prior)

Verbs should have only a few VerbNet classes (same, but with a weighted prior for known verbs)
Figure: The Supervised DPMM used for clustering verb senses. $\theta$ is initialized to reflect the VerbNet class preferences for each verb, when they are known.
## Improvements from Supervision

<table>
<thead>
<tr>
<th>Model</th>
<th>nmPU</th>
<th>niPU</th>
<th>F1</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPMM</td>
<td>55.72</td>
<td>60.33</td>
<td>57.93</td>
<td>522</td>
</tr>
<tr>
<td>SDPMM</td>
<td>51.00</td>
<td>75.71</td>
<td>60.95</td>
<td>122</td>
</tr>
</tbody>
</table>

**Table:** Clustering accuracy: all verbs included in clustering, evaluation only on verbs in the (Korhonen, 2003) dataset. *N* is the number of clusters spanned by the evaluation set.
### Comparison of Produced Clusters

<table>
<thead>
<tr>
<th>Model</th>
<th>Example Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>push (0.20), pull (0.17)</td>
</tr>
<tr>
<td>DPMM</td>
<td>push (0.40), drag (0.27), pull (0.08)</td>
</tr>
<tr>
<td>SDPMM</td>
<td>drag (0.87), push (0.43), pull (0.42), pour (0.39), drop (0.31), force (0.09)</td>
</tr>
</tbody>
</table>

**Table:** Example clusters from the evaluation dataset (Gold), and along with the most-aligned clusters from the unsupervised baseline (DPMM) and our semi-supervised clustering scheme (SDPMM). Weights given in parentheses describe the total proportion of verb instances assigned to each cluster.
Comparison of Produced Clusters

<table>
<thead>
<tr>
<th>Model</th>
<th>Example Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>give (1.0), lend (1.0), generate (0.33), allow (0.25), pull (0.17), pour (0.17)</td>
</tr>
<tr>
<td>DPMM</td>
<td>lend (0.30), give (0.13),</td>
</tr>
<tr>
<td>SDPMM</td>
<td>give (0.82), pour (0.02), ship (0.002)</td>
</tr>
</tbody>
</table>

**Table:** More example clusters.
Questions?


Anna Korhonen, Yuval Krymolowski, and Zvika Marx. *Clustering polysemic subcategorization frame distributions semantically.* ACL 2003