# 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 **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.

# 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

$$P(k|U) \propto \begin{cases} C(k, U), & \text{if } k \in U \\ \alpha, & \text{otherwise,} \end{cases} \tag{1}$$

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

$$P(k|X, U) \propto P(k|U)P(X|k). \tag{2}$$

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

# Verb Sense Induction

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

$$P(X|k) = \prod_{w \in X} \frac{C(w,k) + \beta}{C(*,k) + |V|\beta} \tag{3}$$

# Verb Instance DPMM

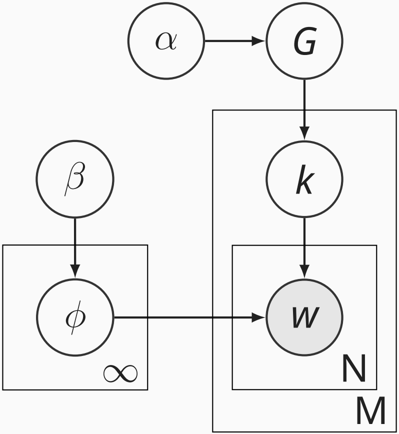


Figure: The proposed graphical model for sense induction. $G$ is the Dirichlet process, $k$ the selected cluster, $M$ the number of verb instances, and $N$ the number of `slot:token` items in the context.

# A Few Useful Speedups

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

Discard initial frames with insufficient counts

# Senses from a Web Corpus

## 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, … |

…

# Senses from a Web Corpus

## 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, … |
| … | |

# Senses from a Web Corpus

## Enter (sense 3)

| | |
|---|---|
| nsubj | <name>:15048, you:12675, they:9717, he:8676, we:7917, … |
| dobj | house:19541, building:16020, home:13692, door:8491, … |
| prep_through | door:2726, gate:915, entrance:586, window:562, <name>:417, … |
| prep_in | <name>:1327, search:143, case:102, morning:89, middle:86, … |
| prep_on | <name>:1352, side:463, right:347, left:332, day:194, level:93, … |

…

# Senses from a Web Corpus

## Enter (sense 4)

| | |
|---|---|
| nsubj | <name>:28116, he:11951, you:9188, i:8702, they:7945, she:7498, … |
| dobj | room:62763, office:10113, store:6076, apartment:4150, kitchen:3353, … |
| prep_with | <name>:605, tray:109, smile:91, gun:81, look:73, bag:66, air:65, … |
| prep_in | <name>:794, time:175, room:146, hand:101, case:58, search:54, … |
| prep_at | moment:522, <name>:464, time:357, point:179, end:177, age:85, night:71, … |
| … | |

# Automatic sense takeaways

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

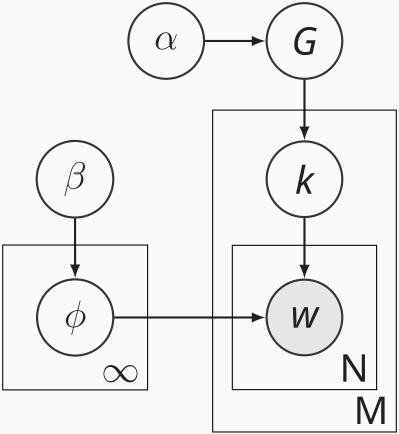


Figure: The proposed graphical model for sense induction. $G$ is the Dirichlet process, $k$ the selected cluster, $M$ the number of verb senses, and $N$ the number of `slot` items in a sense.

# A note about features

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?

| Settings | K | nmPU | niPU | F1 |
| --- | --- | --- | --- | --- |
| Gigaword/S-S | 272.8 | 63.46 | 67.66 | 65.49 |
| Gigaword/S-SW | 36.4 | 31.49 | 95.70 | 47.38 |
| Gigaword/SW-S | 186.2 | 63.52 | 64.18 | 63.84 |
| Gigaword/SW-SW | 30.0 | 36.27 | 94.66 | 52.40 |
| Web/S-S | 363.6 | 61.32 | 78.64 | 68.90 |
| Web/S-SW | 52.2 | 35.80 | **99.30** | 52.62 |
| Web/SW-S | 212.2 | **66.26** | 77.38 | **71.39** |
| Web/SW-SW | 55.0 | 36.70 | 96.25 | 53.13 |

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.

# What about actual labels?

| Settings | K | mPU | iPU | F1 |
|---|---|---|---|---|
| Gigaword/S-NIL | - | 93.43 | 20.06 | 33.03 |
| Gigaword/SW-NIL | - | 94.45 | 41.07 | 57.05 |
| Gigaword/S-S | 512.2 | 75.06 | 45.26 | 56.47 |
| Gigaword/S-S | 260.6 | 73.98 | 56.45 | 64.04 |
| Web/S-NIL | - | 93.70 | 32.96 | 48.78 |
| Web/SW-NIL | - | **94.51** | 44.95 | 60.92 |
| Web/S-S | 500.0 | 72.25 | 52.48 | 60.79 |
| Web/SW-S | 255.2 | 72.65 | **61.00** | **66.31** |

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)

# Partial Supervision Model

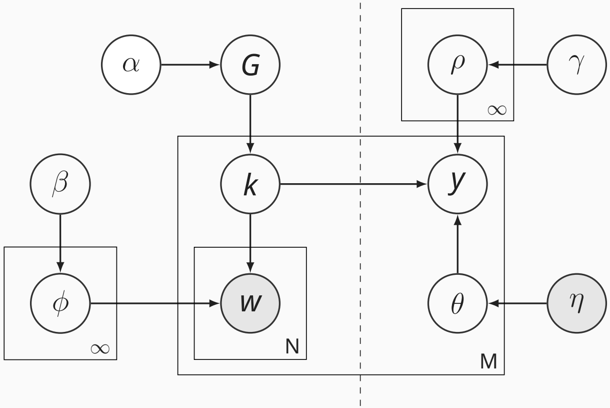


Figure: The Supervised DPMM used for clustering verb senses. $\theta$ is initialized to reflect the VerbNet class preferences for each verb, when they are known.

# Improvements from Supervision

| **Model** | **nmPU** | **niPU** | **F1** | *N* |
|---|---|---|---|---|
| **DPMM** | **55.72** | 60.33 | 57.93 | 522 |
| **SDPMM** | 51.00 | **75.71** | **60.95** | 122 |

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

| **Model** | Example Clusters |
|---|---|
| **Gold** | **push** (0.20), **pull** (0.17) |
| **DPMM** | **push** (0.40), **drag** (0.27), **pull** (0.08) |
| **SDPMM** | **drag** (0.87), **push** (0.43), **pull** (0.42), **pour** (0.39), **drop** (0.31), **force** (0.09) |

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

| **Model** | Example Clusters |
|---|---|
| **Gold** | **give** (1.0), **lend** (1.0), **generate** (0.33), **allow** (0.25), **pull** (0.17), **pour** (0.17) |
| **DPMM** | **lend** (0.30), **give** (0.13), |
| **SDPMM** | **give** (0.82), **pour** (0.02), **ship** (0.002) |

Table: More example clusters.

# Questions?

Daisuke Kawahara, Daniel W. Peterson, and Martha Palmer. **A Step-wise Usage-based Method for Inducing Polysemy-aware Verb Classes.** ACL 2014.

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

Daniel W. Peterson, Martha Palmer, Jordan Boyd-Graber and Daisuke Kawahara. 2016. **Leveraging VerbNet to build Corpus-Specific Verb Clusters.** *SEM 2016