Motivation

Supervised WSD
- Most accurate WSD systems to date are supervised.
- Rely on sense-labeled data to train standard classifiers.
- Acquiring sufficient labeled data is very expensive.
- Limits the use in new domains and languages.
- Makes supervised WSD unfeasible for many applications.

Unsupervised WSD
- Independent of labeled data.
- Most promising solution for large-scale use.
- Much less accurate than supervised methods.

Motivation

Bayesian Word Sense Induction
- LDA-based model
- Incorporating Features
- Inference
- Evaluation

Word Sense Discrimination

It is a clustering problem! Collect all instances of a word in a corpus and partition them into clusters \(\approx\) senses.
Word Sense Discrimination

It is a clustering problem! Collect all instances of a word in a corpus and partition them into clusters ≈ senses.

LDA for Document Classification (Blei et al. 2003)

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure: Original LDA model; shaded nodes represent observed variables, unshaded nodes indicate latent variables. Arrows indicate conditional dependencies between variables, plates refer to repetitions of sampling steps. Variables in the lower right corner refer to number of samples.
Bayesian Sense Induction Model

- one model per word
- immediate context instead of whole document
- context elements replace words
- small number of senses (<10)

Generative Process

For $j = 1 \ldots T$ topics,

Choose $\phi^{(j)} \sim \text{Dirichlet}(\beta)$ \ 
$\phi_1^{(j)} \ldots \phi_v^{(j)}$: prob. of wd. in topic $j$

For $d = 1 \ldots D$ documents,

Choose $\theta^{(d)} \sim \text{Dirichlet}(\alpha)$ \ 
$\theta_1^{(d)} \ldots \phi_T^{(d)}$: prob. of each topic in $d$

For $i = 1 \ldots N_d$ words in doc $d$,

Choose $z_i \sim \text{Multinomial}(\theta^{(d)})$. \ 
$z_i$: topic of word $i$

Choose $w_i \sim \text{Multinomial}(\phi^{(z_i)})$. \ 
$w_i$: identity of word $i$

Multiple Information Sources

- LDA model deals with one type of features: words.
- Many classification problems use several sources of information.
- This is common practice in WSD (collocation features, local context, syntactic relations).
- Ideally we would like to extend the model so as to deal with multiple feature types.

$$P(w_i) = \sum_{j=1}^{S} P(w_i|s_i = j)P(s_i = j)$$
Extended Model

- Each inner rectangle corresponds to a distinct feature type
- Layers are independent (naive assumption)
- All layers share the same sense distribution ($\theta$)
- Each layer has its own sense-feature distribution ($\phi$).

Inference with Gibbs Sampling

- An iterative process.
- Start with random (sense) assignments for each variable.
- In each iteration, for each variable in the data:
  - Assume you know (from the prev. iteration) the assignments of all other variables.
  - Determine the probabilities of each sense-assignment given the rest of the data.
  - Choose the most probable assignment.
- Iterate until convergence.

The probability of a single sense assignment, $s_i$, is proportional to the product of the likelihood of feature $f_i$, given the rest of the data and the prior probability of the assignment.

$$p(s_i|\bar{s}_i, \bar{T}) \propto p(f_i|s_i, \bar{s}_i, \beta) \cdot p(s_i|\bar{s}_i, \alpha)$$

$$p(f_i|s_i, \bar{s}_i, \beta) = \frac{\#(f_i, s_i) + \beta_1}{\#(s_i) + \phi_1}$$

$$\#(f_i, s_i)$$ number of times $f_i$ assigned sense $s_i$

$$\#(s_i)$$ number of times assignment $s_i$ was observed

$$\beta_1$$ Dirichlet prior for distribution $\phi$ in $l$

$$V_l$$ size of the vocabulary in layer $l$
## Experimental Setup

### Training Data
- Wall Street Journal corpus (WSJ, in-domain)
- The British National Corpus (BNC, out-of-domain)

### Test Data
35 nouns from Semeval 2007 (part of WSJ corpus)

### Model Features
- ±10-word window (10w), ±5-word window (5w), collocations (1w), word n-grams (ng)
- part-of-speech n-grams (pg)
- dependency relations (dp).

### Evaluation
Maps clusters to to gold-standard senses automatically, then uses precision, recall, F-score (Agirre and Soroa 2007).

---

## Sense Induction Example

### OntoNotes Sense Definitions for drug:

**Sense 1** Medicines. A substance that affects the body in some legal, usually-beneficial way. Does not apply to narcotics.

**Sense 2** Narcotics. A substance, usually illegal, that causes bodily pleasure or some other reaction. Has a very negative connotation.

---

### Sense Induction Example

**OntoNotes Sense Definitions for power:**

**Sense 1** An ability to control or influence.

**Sense 2** Entity that possesses ability to control or influence.

**Sense 3** Exerted physical force.

**Sense 4** A mathematical measurement.
"Production" "World Politics" "Financial" "National Politics"

<table>
<thead>
<tr>
<th>Plant</th>
<th>Party</th>
<th>Plant</th>
<th>Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company</td>
<td>Government</td>
<td>Co.</td>
<td>President</td>
</tr>
<tr>
<td>Computer</td>
<td>Political</td>
<td>Nuclear</td>
<td>Congress</td>
</tr>
<tr>
<td>Nuclear</td>
<td>Military</td>
<td>Million</td>
<td>State</td>
</tr>
<tr>
<td>Electric</td>
<td>President</td>
<td>Unit</td>
<td>Government</td>
</tr>
<tr>
<td>System</td>
<td>Economic</td>
<td>Utility</td>
<td>Security</td>
</tr>
<tr>
<td>Year</td>
<td>U.S.</td>
<td>Electric</td>
<td>Federal</td>
</tr>
<tr>
<td>U.S.</td>
<td>People</td>
<td>Company</td>
<td>Executive</td>
</tr>
<tr>
<td>Utility</td>
<td>World</td>
<td>Light</td>
<td>Company</td>
</tr>
<tr>
<td>Price</td>
<td>Soviet</td>
<td>Corp.</td>
<td>Court</td>
</tr>
<tr>
<td>Line</td>
<td>Country</td>
<td>Power</td>
<td>Law</td>
</tr>
<tr>
<td>Market</td>
<td>Struggle</td>
<td>Share</td>
<td>Veto</td>
</tr>
<tr>
<td>Industry</td>
<td>Election</td>
<td>Inc.</td>
<td>Authority</td>
</tr>
</tbody>
</table>

---

**Results on WSJ**

<table>
<thead>
<tr>
<th>1-Layer</th>
<th>5-Layers</th>
<th>Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>10w</td>
<td>66.9</td>
<td>-10w</td>
</tr>
<tr>
<td>5w</td>
<td>86.8</td>
<td>-5w</td>
</tr>
<tr>
<td>1w</td>
<td>84.6</td>
<td>-1w</td>
</tr>
<tr>
<td>ng</td>
<td>83.6</td>
<td>-ng</td>
</tr>
<tr>
<td>pg</td>
<td>82.5</td>
<td>-pg</td>
</tr>
<tr>
<td>dp</td>
<td>82.2</td>
<td>-dp</td>
</tr>
<tr>
<td>MFS</td>
<td>80.9</td>
<td>all</td>
</tr>
</tbody>
</table>

Table: Model performance (F-score) on the WSJ with one layer (left), five layers (middle), and selected combinations of layers (right).

**Results on BNC**

<table>
<thead>
<tr>
<th>1-Layer</th>
<th>5-Layers</th>
<th>Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>10w</td>
<td>84.6</td>
<td>-10w</td>
</tr>
<tr>
<td>5w</td>
<td>84.6</td>
<td>-5w</td>
</tr>
<tr>
<td>1w</td>
<td>83.6</td>
<td>-1w</td>
</tr>
<tr>
<td>pg</td>
<td>83.1</td>
<td>-pg</td>
</tr>
<tr>
<td>ng</td>
<td>82.8</td>
<td>-ng</td>
</tr>
<tr>
<td>dp</td>
<td>81.1</td>
<td>-dp</td>
</tr>
<tr>
<td>MFS</td>
<td>80.9</td>
<td>all</td>
</tr>
</tbody>
</table>

Table: Model performance (F-score) on the BNC with one layer (left), five layers (middle), and selected combinations of layers (right).

**Discussion**

**Strengths:**
- The method is conceptually simple, and requires no human involvement
- Can integrate several features beyond words
- Probabilistic formulation allows integration with applications (via mixture or product models)

**Weaknesses:**
- Sense proportions may vary for different features!
- Fixed number of senses for all words (infer number of senses)
- Focuses only on local topics (what about global topics?)

