Nonparametric Bayesian Word Sense Induction

Xuchen Yao\textsuperscript{1} and Benjamin Van Durme\textsuperscript{1,2}

\textsuperscript{1}Department of Computer Science
\textsuperscript{2}Human Language Technology Center of Excellence
Johns Hopkins University

TextGraphs-6
June 23 2011
Word Sense Induction (WSI) v.s. Word Sense Disambiguation (WSD)

- the task of automatically discovering latent senses for each word type, across a collection of that word’s tokens situated in context.
  - “a bank loan” → Cluster1
  - “the Willamette River bank” → Cluster2

- WSD: has a predefined sense inventory, such as WordNet, OntoNotes.
  - “a bank loan” → bank.n.1 (place for money)
  - “the Willamette River bank” → bank.n.2 (land along the side of a river or lake)

- We perform the task of WSI instead of WSD mainly because:
  - WSI requires no dictionaries (which have various shortcomings)
  - WSI can also be used to disambiguate senses (sufficient to tell different senses apart)
Word Sense Induction (WSI) v.s. Word Sense Disambiguation (WSD)

- the task of automatically discovering latent senses for each word type, across a collection of that word’s tokens situated in context.
  - “a bank loan” → Cluster1
  - “the Willamette River bank” → Cluster2

- WSD: has a predefined sense inventory, such as WordNet, OntoNotes.
  - “a bank loan” → bank.n.1 (place for money)
  - “the Willamette River bank” → bank.n.2 (land along the side of a river or lake)

- We perform the task of WSI instead of WSD mainly because:
  - WSI requires no dictionaries (which have various shortcomings)
  - WSI can also be used to disambiguate senses (sufficient to tell different senses apart)
Word Sense Induction (WSI) v.s. Word Sense Disambiguation (WSD)

• the task of automatically discovering latent senses for each word type, across a collection of that word’s tokens situated in context.
  • “a bank loan” → Cluster1
  • “the Willamette River bank” → Cluster2

• WSD: has a predefined sense inventory, such as WordNet, OntoNotes.
  • “a bank loan” → bank.n.1 (place for money)
  • “the Willamette River bank” → bank.n.2 (land along the side of a river or lake)

• We perform the task of WSI instead of WSD mainly because:
  • WSI requires no dictionaries (which have various shortcomings)
  • WSI can also be used to disambiguate senses (sufficient to tell different senses apart)
Bayesian WSI
Parametric v.s. Nonparametric

- Evaluation on SemEval-2007 task 02 (Agirre and Soroa, 2007)

<table>
<thead>
<tr>
<th>method</th>
<th>in-domain</th>
<th>out-of-domain</th>
<th>#senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>B&amp;L</td>
<td>LDA</td>
<td>86.9%</td>
<td>84.6%</td>
</tr>
<tr>
<td>Our work</td>
<td>HDP</td>
<td>86.7%</td>
<td>85.7%</td>
</tr>
</tbody>
</table>

**Table**: F1 measure when training with in-domain (WSJ) or out-of-domain (BNC) data, using only ±10 word context as feature.
Intuition
the senses of words are hinted at by their contextual information (Yarowsky, 1992).

Example
given the word \textit{bank} with a sense \textit{river bank}, it is more likely that the neighboring words are \textit{river, lake and water} than \textit{finance, money and loan}.

Simplication
We only use the \textit{±10 word context} as feature since B&L saw no improvements using syntactic features (pos, dependency, which also depend on a mature NLP pipeline).
Intuition
the senses of words are hinted at by their contextual information (Yarowsky, 1992).

Example
given the word bank with a sense river bank, it is more likely that the neighboring words are river, lake and water than finance, money and loan.

Simplication
We only use the the ±10 word context as feature since B&L saw no improvements using syntactic features (pos, dependency, which also depend on a mature NLP pipeline).
Parametric Bayesian WSI
Latent Dirichlet Allocation (LDA, Blei et al., 2003)

\[
p(w_{m,n}) = \sum_{k=1}^{K} p(w_{m,n} \mid s_{m,n}=k)p(s_{m,n}=k)
\]

**Generative Story:**

For \( k \in (1, \ldots, K) \) senses:
- Sample mixture component: \( \varphi_k \sim \text{Dir}(\vec{\beta}) \).

For \( m \in (1, \ldots, M) \) pseudo-docs:
- Sample sense components \( \hat{\theta}_m \sim \text{Dir}(\vec{\alpha}) \).

For \( n \in (1, \ldots, N_m) \) words in pseudo-doc \( m \):
- Sample sense index \( s_{m,n} \sim \text{Mult}(\hat{\theta}_m) \).
- Sample word \( w_{m,n} \sim \text{Mult}(\varphi_{s_{m,n}}) \).
Nonparametric Bayesian WSI
Hierarchical Dirichlet Process (HDP, Teh et al., 2006)

Generative Story:

Select base distribution $G_0 \sim DP(\gamma, H)$ which provides an unlimited inventory of senses.
For $m \in (1, ..., M)$ pseudo-docs:
   Draw $G_m \sim DP(\alpha_0, G_0)$.
For $n \in (1, ..., N_m)$ words in pseudo-doc $m$:
   Sample $s_{m,n} \sim G_m$.
   Sample $w_{m,n} \sim Mult(s_{m,n})$. 
Let us now compute marginals under a hierarchical Dirichlet process when \( G_0 \) and \( G_j \) are

\[ G_0 \sim DP(\gamma, H) \]
\[ G_m \sim DP(\alpha_0, G_0) \]

Multiple restaurants (documents) share a set of dishes (senses).
\( \gamma \sim Gamma \): controls the variability of the global sense distribution.
\( \alpha_0 \sim Gamma \): controls the variability of each customer’s (word) choice of dishes (senses).

**Figure**: CRF Interpretation of HDP (Teh et al., 2006)
Evaluation

- Feature: ±10 word context
- Test data
  - SemEval-2007 task 2, with 15,852 instances of 35 nouns
  - “Supervised Evaluation”: 72% mapping, 14% dev, 14% test
  - annotated with OntoNotes (Hovy et al., 2006) senses, on average 3.9 senses/word.
- Training data
  - In-domain: WSJ in years 87/88/90/94, 930K instances
  - out-of-domain: BNC, 930K instances
Baseline: 80.9% (the most frequent sense)

<table>
<thead>
<tr>
<th></th>
<th>WSJ (in-domain)</th>
<th>BNC (out-of-domain)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA-4s*</td>
<td>86.9</td>
<td>LDA-8s*</td>
</tr>
<tr>
<td>LDA-4s</td>
<td>86.1</td>
<td>LDA-8s</td>
</tr>
<tr>
<td>HDP</td>
<td>86.7</td>
<td>HDP</td>
</tr>
</tbody>
</table>

Table: Results with * are taken from B&L. 4 or 8 senses were used per word. \(\Delta\): statistically significant against LDA-8s by paired permutation test with \(p < 0.001\).

- our F1 measures on LDA are 0.8% lower than reported by B&L.
- the HDP model appears to better adapt to data in other domains.
**Number of Senses**

test set average: 3.9 senses/word

<table>
<thead>
<tr>
<th></th>
<th>WSJ</th>
<th>BNC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train(WSJ)</td>
<td>Test(WSJ)</td>
</tr>
<tr>
<td>LDA</td>
<td>4.0</td>
<td>3.9</td>
</tr>
<tr>
<td>HDP</td>
<td>5.8</td>
<td>3.9</td>
</tr>
</tbody>
</table>

**Table:** The average number of senses the LDA and HDP models output when training with WSJ/BNC and testing on SemEval-2007 (genre: WSJ).
Number of Senses
Deviation from the number of annotated senses

Figure: The difference between induced number of senses and annotated senses with BNC as the training set.
Example: **president**. OntoNotes defines 3 senses:

1. chair of an organization.
2. head of a country.
3. head of U.S.

HDP infers 2 senses.

LDA: 8 senses?
Example on Number of Senses

Example: **president**. OntoNotes defines 3 senses:
1. chair of an organization.
2. head of a country.
3. head of U.S.

HDP infers 2 senses.
LDA: 8 senses?
Examples on HDP-selected Senses
with manual mapping to OntoNotes senses

capital:

<table>
<thead>
<tr>
<th>HDP</th>
<th>OntoNotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  property, tax, cost, year, income</td>
<td>Wealth in the form of money or property</td>
</tr>
<tr>
<td>2  national, region, ottawa, cultural</td>
<td>a seat of government or influence</td>
</tr>
<tr>
<td>3? de, mark, xxxx, letter, expression</td>
<td>a letter represented in uppercase</td>
</tr>
<tr>
<td>?</td>
<td>a book by Karl Marx</td>
</tr>
<tr>
<td>?</td>
<td>uppermost part of a column</td>
</tr>
</tbody>
</table>

plant:

<table>
<thead>
<tr>
<th>HDP</th>
<th>OntoNotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  products, food, power, processing</td>
<td>a building for industrial activity</td>
</tr>
<tr>
<td>2  species, water, soil, growth, habitat</td>
<td>living photosynthesizing organism</td>
</tr>
<tr>
<td>3? chapman, regiment, veteran, captain</td>
<td>a contrivance or stratagem</td>
</tr>
</tbody>
</table>
Examples on HDP-selected Senses
with manual mapping to OntoNotes senses

### capital:

<table>
<thead>
<tr>
<th>HDP</th>
<th>OntoNotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 property, tax, cost, year, income</td>
<td>Wealth in the form of money or property</td>
</tr>
<tr>
<td>2 national, region, ottawa, cultural</td>
<td>a seat of government or influence</td>
</tr>
<tr>
<td>3? de, mark, xxxx, letter, expression</td>
<td>a letter represented in uppercase</td>
</tr>
<tr>
<td>?</td>
<td>a book by Karl Marx</td>
</tr>
<tr>
<td>?</td>
<td>uppermost part of a column</td>
</tr>
</tbody>
</table>

### plant:

<table>
<thead>
<tr>
<th>HDP</th>
<th>OntoNotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 products, food, power, processing</td>
<td>a building for industrial activity</td>
</tr>
<tr>
<td>2 species, water, soil, growth, habitat</td>
<td>living photosynthesizing organism</td>
</tr>
<tr>
<td>3? chapman, regiment, veteran, captain</td>
<td>a contrivance or stratagem</td>
</tr>
</tbody>
</table>
Conclusion

- **Performance in F1**
  - HDP and LDA are equivalent
  - HDP adapts better to balanced-domain data

- **Number of Senses**
  - LDA: fixed, hard to use in applications
  - HDP: flexible, only have to tune the hyper-parameters.


