Nonparametric Bayesian Word Sense Induction

Xuchen Yao[1] and Benjamin Van Durme[1,2]

[1]Department of Computer Science
[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

- Brody and Lapata (2009): Bayesian Word Sense Induction, in EACL 09.
- Evaluation on SemEval-2007 task 02 (Agirre and Soroa, 2007)

	method	**in-domain**	**out-of-domain**	**#senses**
B&L	LDA	86.9%	84.6%	fixed
Our work	HDP	86.7%	85.7%	flexible

Table: F1 measure when training with in-domain (WSJ) or out-of-domain (BNC) data, using only ± 10 word context as feature.

Using Topic Models for WSI

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).

Using Topic Models for WSI

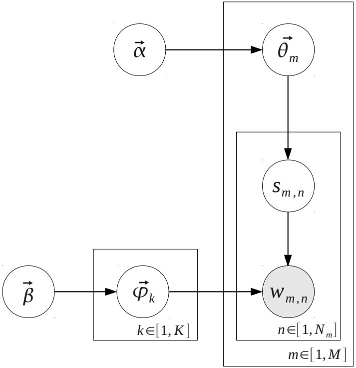
Intuition
the senses of words are hinted at by their contextual information (Yarowsky, 1992).

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)



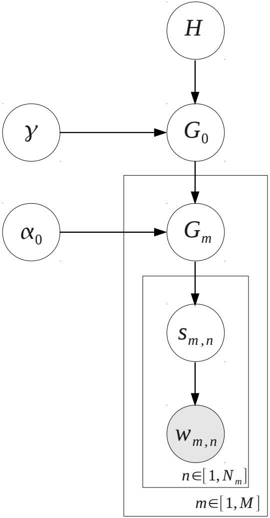
$$p(w_{m,n}) = \sum_{k=1}^{K} p(w_{m,n} \mid s_{m,n} = k) p(s_{m,n} = k)$$

Generative Story:

For $k \in (1, \ldots, K)$ senses:
  Sample mixture component: $\vec{\varphi}_k \sim Dir(\vec{\beta})$.
For $m \in (1, \ldots, M)$ pseudo-docs:
  Sample sense components $\vec{\theta}_m \sim Dir(\vec{\alpha})$.
  For $n \in (1, \ldots, N_m)$ words in pseudo-doc m:
    Sample sense index $s_{m,n} \sim Mult(\vec{\theta}_m)$.
    Sample word $w_{m,n} \sim Mult(\vec{\varphi}_{s_{m,n}})$.

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 $w_{m,n} \sim Mult(s_{m,n})$.

Chinese Restaurant Franchise Interpretation

Hyperparameters γ and α_0

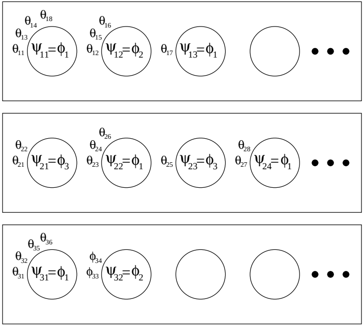


Figure: CRF Interpretation of HDP (Teh et al., 2006)

$$G_0 \sim DP(\gamma, H)$$
$$G_m \sim DP(\alpha_0, G_0)$$

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).

Evaluation

- Feature: $\pm$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

F1

Baseline: 80.9% (the most frequent sense)

WSJ(in-domain)		**BNC**(out-of-domain)	
LDA-4s*	86.9	LDA-8s*	84.6
LDA-4s	86.1	LDA-8s	83.8
HDP	86.7	HDP	85.7$^{\triangle}$

Table: Results with * are taken from B&L. **4** or **8** senses were used per word. $\triangle$: 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

	WSJ		**BNC**	
	Train(WSJ)	Test(WSJ)	Train(BNC)	Test(WSJ)
LDA	4.0	3.9	8.0	7.4
HDP	5.8	3.9	9.4	4.6

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

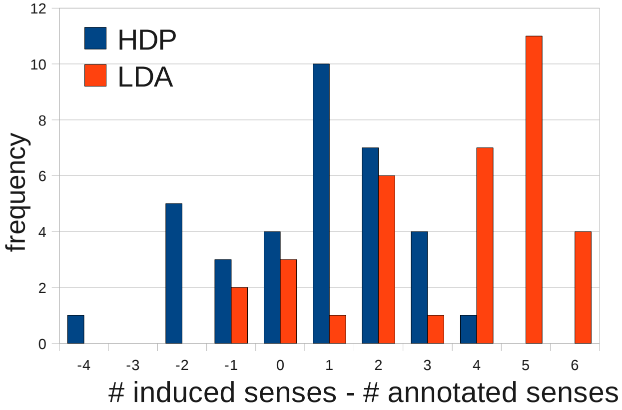
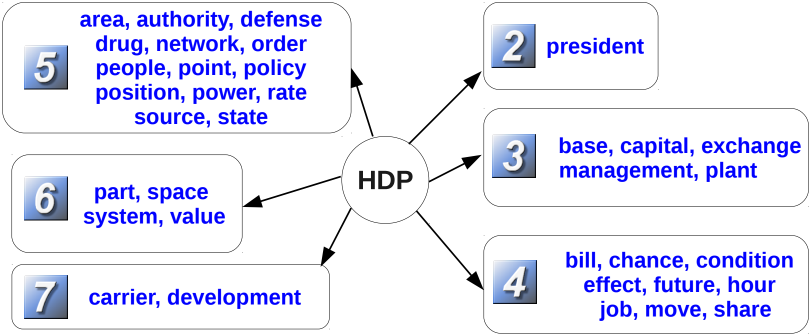


Figure: The difference between induced number of senses and annotated senses with BNC as the training set.

Example on Number of Senses



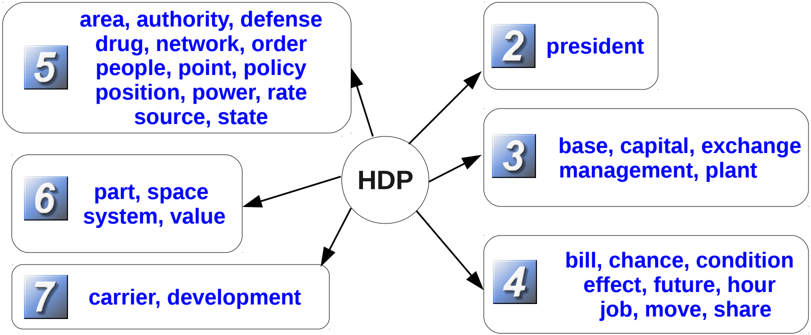
Example: **president**. OntoNotes defines 3 senses:

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

HDP infers 2 senses.

LDA: 8 senses?

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:

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

plant:

	HDP	OntoNotes
1	products, food, power, processing	a building for industrial activity
2	species, water, soil, growth, habitat	living photosynthesizing organism
3?	chapman, regiment, veteran, captain	a contrivance or stratagem

Examples on HDP-selected Senses

with manual mapping to OntoNotes senses

capital:

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

plant:

	HDP	OntoNotes
1	products, food, power, processing	a building for industrial activity
2	species, water, soil, growth, habitat	living photosynthesizing organism
3?	chapman, regiment, veteran, captain	a contrivance or stratagem

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.

Eneko Agirre and Aitor Soroa. Semeval-2007 Task 02: Evaluating Word Sense Induction And Discrimination Systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 7–12, 2007.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022, January 2003.

Samuel Brody and Mirella Lapata. Bayesian Word Sense Induction. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111, 2009.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60, 2006.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

David Yarowsky. Word-Sense Disambiguation Using Statistical Models Of Roget's Categories Trained On Large Corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 454–460, 1992.