Bayesian Word Sense Induction

Samuel Brody[1] Mirella Lapata[2]

[1]Department of Biomedical Informatics
Columbia University

[2]School of Informatics
University of Edinburgh

EACL 2009

Outline

Outline

Motivation

“ ... we find that word sense disambiguation does not yield significantly better translation quality than the statistical machine translation system alone."

– Carpuat and Wu (2005)

“ ... missing correct matches because of incorrect sense resolution has a much more deleterious effect on retrieval performance than does making spurious matches."

– Voorhees (1993)

Why?

*" Major barriers to building a high-performing word sense disambiguation system include the difficulty of labeling data for this task and of predicting **fine-grained sense distinctions**. These issues stem partly from the fact that **the task is being treated in isolation from possible uses** of automatically disambiguated data."*

– Vickrey et al. (2005)

*" ... one of the main problems in word sense disambiguation lies upstream, in the very sense lists used by systems. **Conventional dictionaries are not suited to this task**; they usually contain definitions that are too general ... and there is no guarantee that they reflect the exact content of the particular textbase being queried ... "*

– Véronis (2004)

Solution

Sense Induction / Discrimination

- Detects natural distinctions in the data.
- Independent of any dictionary.
- Distinctions suit the relevant domain and task.

Outline

Clustering Approach

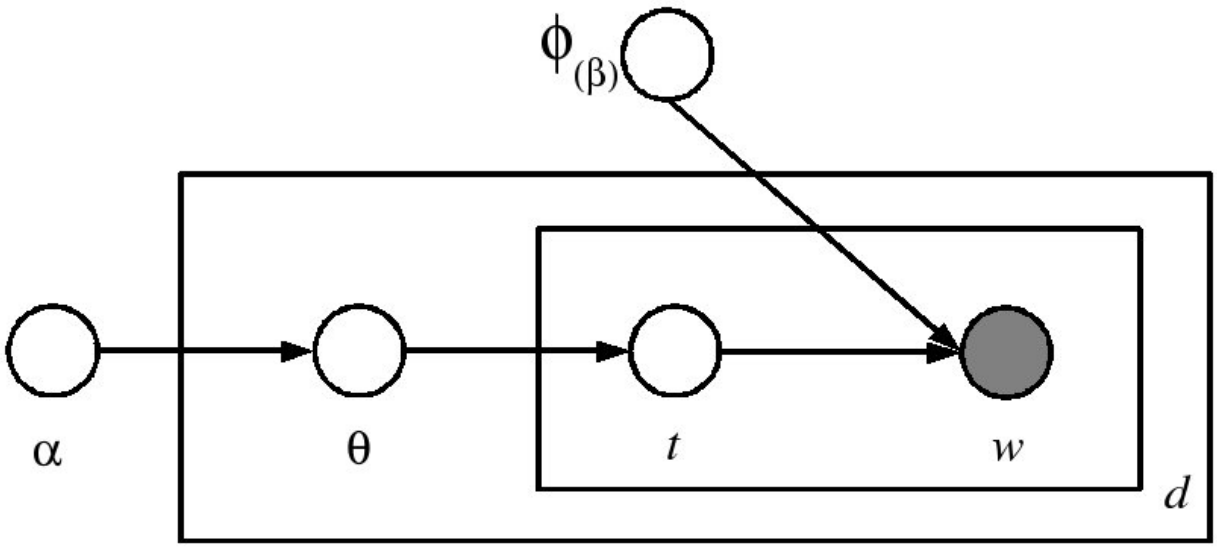
Common Approach : standard clustering task

- Does not take into account the linguistic nature of the data.
- Does not lend itself to easy integration.

Our Approach : probabilistic generative model

+ Generative aspect suits linguistic data
+ Probabilistic nature makes for easy integration (via mixture or product models)

LDA for Document Classification (Blei et al., 2003)



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.

LDA for Document Classification (Blei et al., 2003)

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Previous LDA Approaches to WSD

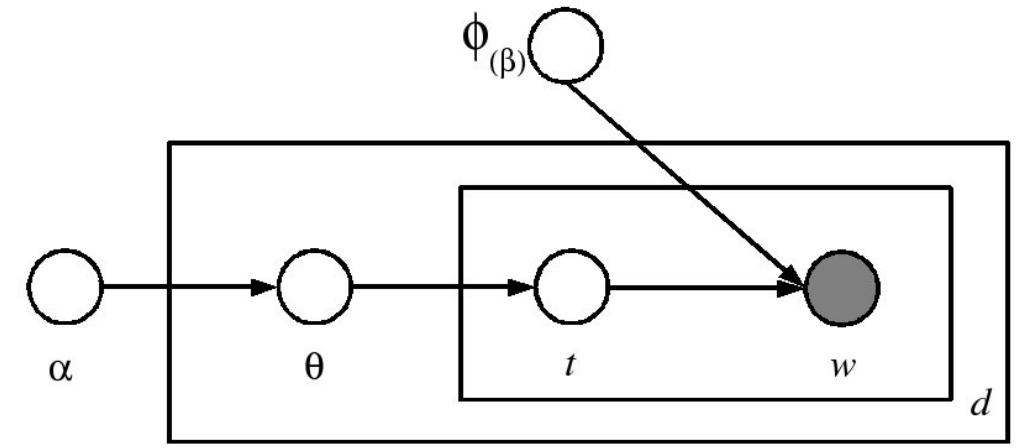
- Supervised - use LDA-derived topics instead of Bag-of-Words. (Cai et al., 2007)
- Unsupervised - integrate distributional similarity approach with LDA. (Boyd-Graber and Blei, 2007)

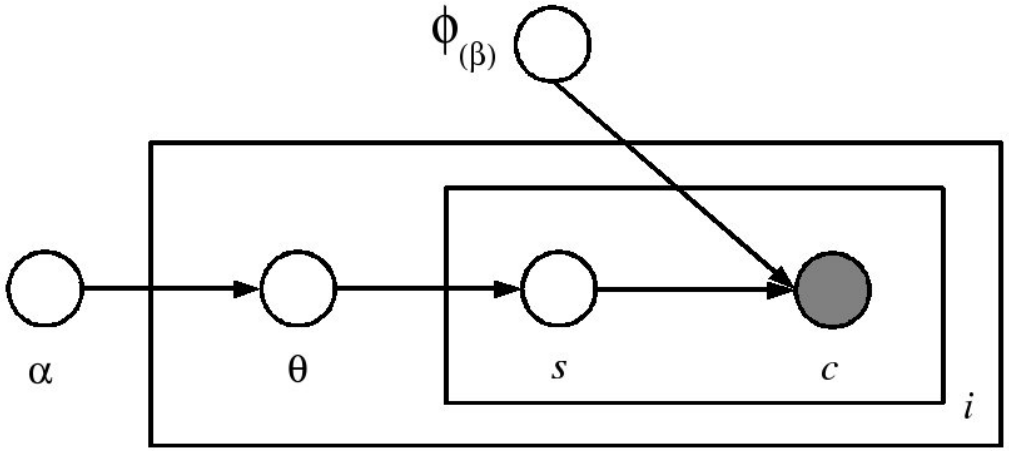
Problems in Previous Approaches

- Treat topics as domain labels.
- Use as an aid in disambiguation.

Outline

Adapting Classic LDA



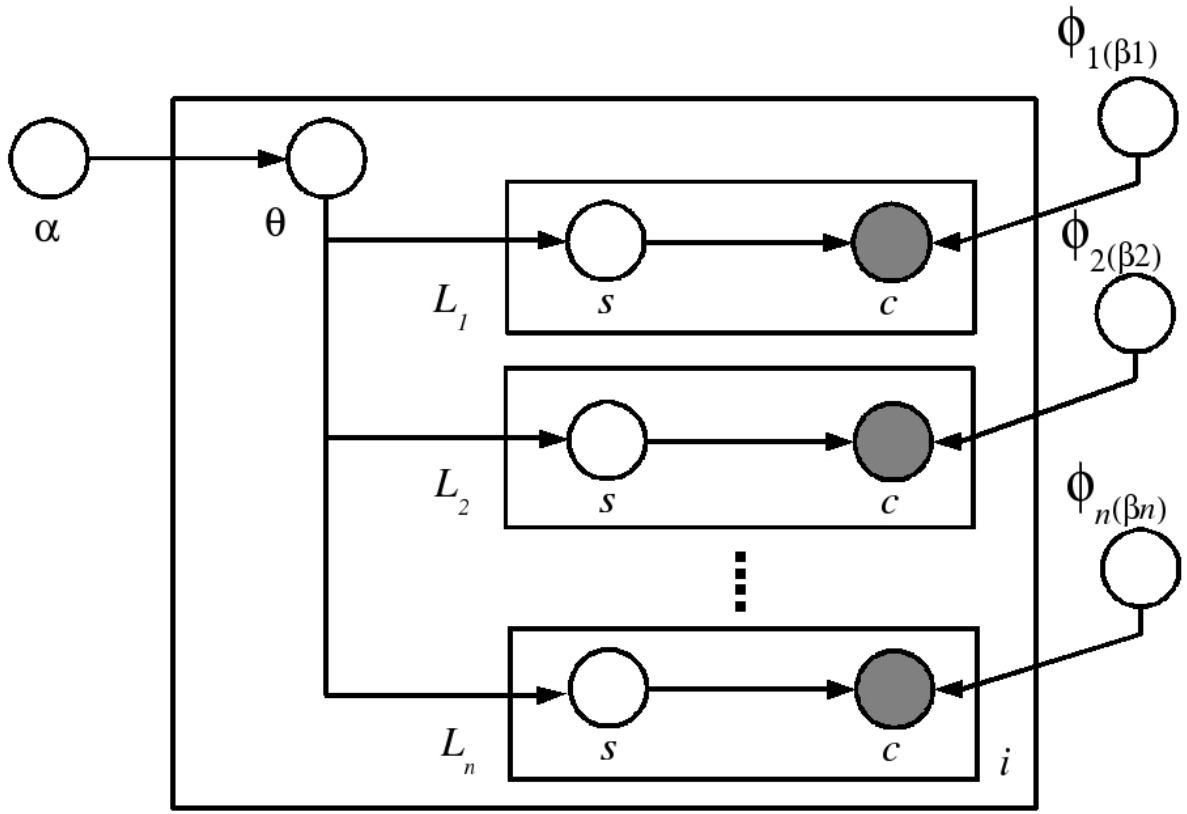


- one model per word
- immediate context instead of whole document
- context elements replace words
- small number of senses (<10)

Multiple Information Sources

- Original LDA model deals with one input layer - words.
- Many classification problems use several sources of information.
- This is common practice in WSD (domain features, local context, syntactic features).
- We extended our model to deal with multiple layers: $\pm$10 word window (10w), $\pm$5 word window (5w), collocations (1w), word bigrams (ng), part-of-speech bigrams (pg), dependency relations (dep)

Layered LDA



Inference

Gibbs Sampling

- An iterative process.
- Start with random assignments (sense-topics) 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 convergeance.

Outline

Experimental Setup

Evaluation Dataset - Semeval (Agirre et al., 2007)

- 35 nouns from the lexical sample.
- Text from the Penn Treebank II. The Treebank data is a collection of articles from first half of the 1989 Wall Street Journal.

In-Domain

Wall Street Journal (WSJ) corpus.

- news with a business and financial perspective
- all articles 1987-89 and 1994 - 740k instances

Out-of-Domain

The British National Corpus (BNC)

- cross-section of 20th century, written & spoken, British English.
- 100 million words - 730k instances

Evaluation

Semeval Sense Discrimination Task (Agirre and Soroa, 2007)

Provided a standardized framework for evaluation of unsupervised sense discrimination systems.

- evaluation dataset
- automated system for mapping to gold-standard
- standardized evaluation metrics

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.

“Enforcement”	“Treatment”	“Industry”	“Research”
U.S.	patient	company	administration
administration	people	million	food
federal	problem	sale	company
against	doctor	maker	approval
war	company	stock	fda
dealer	abuse	inc.	patient
government	aid	market	test
official	user	product	market
enforcement	test	co.	U.S.
testing	prescription	U.S.	approve
charge	cost	sterling	treat
trafficker	year	prescription	aid
money	alcohol	drug	study
president	effect	generic	product
abuse	addict	analyst	treatment
program	treatment	industry	develop
law	Dr.	pharmaceutical	receive

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"
plant	party	plant	bank
company	government	co.	president
computer	political	nuclear	congress
nuclear	military	million	state
electric	president	unit	government
system	economic	utility	security
year	U.S.	electric	federal
U.S.	people	company	executive
utility	world	light	company
price	soviet	corp.	court
line	country	power	law
market	struggle	share	veto
industry	election	inc.	authority

Layer Experiments

Single Layer

Layer	F-Score
10w	**86.9%**
5w	86.8%
1w	84.6%
ng	83.6%
pg	82.5%
dep	82.2%
MFS	80.9%

Remove One

Layer	Diff.	F-Score
-10w	-0.2%	83.1%
-5w	-0.3%	83.0%
-1w	-0.3%	83.0%
-ng	-0.3%	83.0%
-pg	-0.6%	82.7%
-dep	**+1.4%**	84.7%
All	–	83.3%

Combinations

Layer	F-Score
10w+5w	**87.3%**
5w+pg	83.9%
1w+ng	83.2%
10w+pg	83.3%
1w+pg	84.5%
10w+pg+dep	82.2%
MFS	80.9%

Layer Experiments

Single Layer

Layer	F-Score
10w	**86.9%**
5w	86.8%
1w	84.6%
ng	83.6%
pg	82.5%
dep	82.2%
MFS	80.9%

Remove One

Layer	Diff.	F-Score
-10w	-0.2%	83.1%
-5w	-0.3%	83.0%
-1w	-0.3%	83.0%
-ng	-0.3%	83.0%
-pg	-0.6%	82.7%
-dep	**+1.4%**	84.7%
All	–	83.3%

Combinations

Layer	F-Score
10w+5w	**87.3%**
5w+pg	83.9%
1w+ng	83.2%
10w+pg	83.3%
1w+pg	84.5%
10w+pg+dep	82.2%
MFS	80.9%

Layer Experiments

Single Layer

Layer	F-Score
10w	**86.9%**
5w	86.8%
1w	84.6%
ng	83.6%
pg	82.5%
dep	82.2%
MFS	80.9%

Remove One

Layer	Diff.	F-Score
-10w	-0.2%	83.1%
-5w	-0.3%	83.0%
-1w	-0.3%	83.0%
-ng	-0.3%	83.0%
-pg	-0.6%	82.7%
-dep	**+1.4%**	84.7%
All	–	83.3%

Combinations

Layer	F-Score
10w+5w	**87.3%**
5w+pg	83.9%
1w+ng	83.2%
10w+pg	83.3%
1w+pg	84.5%
10w+pg+dep	82.2%
MFS	80.9%

Layer Experiments

Single Layer

Layer	F-Score
10w	**86.9%**
5w	86.8%
1w	84.6%
ng	83.6%
pg	82.5%
dep	82.2%
MFS	80.9%

Remove One

Layer	Diff.	F-Score
-10w	-0.2%	83.1%
-5w	-0.3%	83.0%
-1w	-0.3%	83.0%
-ng	-0.3%	83.0%
-pg	-0.6%	82.7%
-dep	**+1.4%**	84.7%
All	–	83.3%

Combinations

Layer	F-Score
10w+5w	**87.3%**
5w+pg	83.9%
1w+ng	83.2%
10w+pg	83.3%
1w+pg	84.5%
10w+pg+dep	82.2%
MFS	80.9%

Layer Experiments

Single Layer

Layer	F-Score
10w	**86.9%**
5w	86.8%
1w	84.6%
ng	83.6%
pg	82.5%
dep	82.2%
MFS	80.9%

Remove One

Layer	Diff.	F-Score
-10w	-0.2%	83.1%
-5w	-0.3%	83.0%
-1w	-0.3%	83.0%
-ng	-0.3%	83.0%
-pg	-0.6%	82.7%
-dep	**+1.4%**	84.7%
All	–	83.3%

Combinations

Layer	F-Score
10w+5w	**87.3%**
5w+pg	83.9%
1w+ng	83.2%
10w+pg	83.3%
1w+pg	84.5%
10w+pg+dep	82.2%
MFS	80.9%

Comparison to State-of-the-Art

Scores on Semeval

System	F-Score
LDA-IN	**87.3%**
I2R	86.8%
UMND2	84.5%
MFS	80.9%

- LDA system significantly outperforms the MFS baseline
- better performance than highest-scoring system in Semeval

Outline

Conclusions

The Problem:

- traditional WSD uses unsuitable sense inventories

Our Solution:

- a generative, probabilistic model for sense induction
- achieves state-of-the-art results on the induction task

Future Directions

- integrate in an application
  - e.g., WSD in MT (with source & target layers)
- automatic parameter estimation
  - infinite LDA for model order (Teh et al., 2006)
  - integrate hyperparameter estimation into sampling (Goldwater and Griffiths, 2007)
- other uses for layered model
  - document classification (text, abstract, image layers)
  - music recommendation (lyrics and acoustic feature layers)

Thank You!

Bibliography

Agirre, Eneko, Lluís Màrquez, and Richard Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic.

Agirre, Eneko and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, pages 7–12.

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

Boyd-Graber, Jordan and David Blei. 2007. Putop: Turning predominant senses into a topic model for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, pages 277–281.

Cai, J. F., W. S. Lee, and Y. W. Teh. 2007. Improving word sense disambiguation using topic features. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-coNLL)*. pages 1015–1023.

Carpuat, Marine and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, pages 387–394.

Goldwater, Sharon and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 744–751.

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

Véronis, Jean. 2004. Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language* 18(3):223–252.

Vickrey, David, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of the HLT/EMNLP*. Vancouver, pages 771–778.

Voorhees, Ellen M. 1993. Using wordnet to disambiguate word senses for text retrieval. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, pages 171–180.