Unsupervised Methods for ~~NLP~~ WSD

Samuel Brody

Department of Biomedical Informatics
Columbia University
samuel.brody@dbmi.columbia.edu

July 2010

Outline

1. Introduction - Unsupervised NLP
   - The Competition - Supervised Methods
   - Colleagues - Human Knowledge
   - Unsupervised Learning
2. Word Sense Disambiguation (WSD)
   - Unsupervised Labeling
   - Bayesian Sense Induction
3. Work in Progress - Aspect & Sentiment in Reviews
4. Conclusion

Outline

1. Introduction - Unsupervised NLP
   - The Competition - Supervised Methods
   - Colleagues - Human Knowledge
   - Unsupervised Learning
2. Word Sense Disambiguation (WSD)
   - Unsupervised Labeling
   - Bayesian Sense Induction
3. Work in Progress - Aspect & Sentiment in Reviews
4. Conclusion

Outline

1. Introduction - Unsupervised NLP
   - The Competition - Supervised Methods
   - Colleagues - Human Knowledge
   - Unsupervised Learning
2. Word Sense Disambiguation (WSD)
   - Unsupervised Labeling
   - Bayesian Sense Induction
3. Work in Progress - Aspect & Sentiment in Reviews
4. Conclusion

The Competition - Supervised Machine Learning

Supervised methods are used for many NLP tasks (parsing, relation extraction, WSD)

Why?

+ high accuracy with sufficient annotation
+ full collection of powerful and easy-to-use tools (e.g., SVM, kNN, Maximum Entropy)

The Competition - Supervised Machine Learning

Why not?

- annotation is expensive
- doesn't transfer well between domains and tasks
- is it a good model for human learning?
  - do humans perform singular-value decomposition?
  - discriminative rather than generative
  - concepts come from the annotation rather than the data

Outline

1. Introduction - Unsupervised NLP
   - The Competition - Supervised Methods
   - Colleagues - Human Knowledge
   - Unsupervised Learning
2. Word Sense Disambiguation (WSD)
   - Unsupervised Labeling
   - Bayesian Sense Induction
3. Work in Progress - Aspect & Sentiment in Reviews
4. Conclusion

Colleagues - Knowledge Bases

Many “unsupervised" approaches make use of manually compiled knowledge bases.

- Dictionaries
- Thesauri
- FrameNet
- PropBank

The Problem with Knowledge

WordNet senses for *bank*:

1. river bank
2. financial institution
3. bank of earth

...

9. bank building
10. a flight maneuver

- lack of coverage
- no domain/task specificity
- over representation of marginal cases
- based on a specific theory

Colleagues - Scientific Theory

- Linguistic Theory
- Psychology
- Neurology
- Formal Logic

Ignorance = Bliss?

"Whenever I fire a linguist our system performance improves"

- Fred Jelinek

Why? (see *"Some Of My Best Friends Are Linguists"* - Fred Jelinek)

- strict models do not allow for "grey" areas
- attempts to cover rare cases leads to excessive complexity
- models do not scale to practical cases

Outline

1. Introduction - Unsupervised NLP
   - The Competition - Supervised Methods
   - Colleagues - Human Knowledge
   - Unsupervised Learning
2. Word Sense Disambiguation (WSD)
   - Unsupervised Labeling
   - Bayesian Sense Induction
3. Work in Progress - Aspect & Sentiment in Reviews
4. Conclusion

Unsupervised Learning

Unsupervised techniques offer many tools and insights:

- EM
  - classification / generalization
- Automatic Alignment
  - corpus statistics
  - information theory
- Bayesian Models, LDA
  - probabilistic view
  - minimal assumptions

Competition & Colleagues

We can still benefit from:

- insights and tools from supervised learning
- careful use of knowledge bases
- aspects of scientific theory

Outline

1. Introduction - Unsupervised NLP
   - The Competition - Supervised Methods
   - Colleagues - Human Knowledge
   - Unsupervised Learning
2. Word Sense Disambiguation (WSD)
   - Unsupervised Labeling
   - Bayesian Sense Induction
3. Work in Progress - Aspect & Sentiment in Reviews
4. Conclusion

Outline

1. Introduction - Unsupervised NLP
   - The Competition - Supervised Methods
   - Colleagues - Human Knowledge
   - Unsupervised Learning
2. Word Sense Disambiguation (WSD)
   - Unsupervised Labeling
   - Bayesian Sense Induction
3. Work in Progress - Aspect & Sentiment in Reviews
4. Conclusion

Good Senses Make Good Neighbors:

Exploiting Distributional Similarity for Unsupervised WSD

Brody and Lapata (2008)

Motivation

Supervised WSD

- Most accurate WSD systems to date are supervised.
- Rely on sense-labeled training 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.

Solution

The Idea: Automatic Labeling

- go directly to the data
- replace manual annotation
- retain use of supervised classifiers

Prev. Approach - Linguistic Knowledge

Synonyms from a Lexical Resource (Leacock et al., 1998; Mihalcea, 2002)

- Obtain synonymous / related words for each sense.
- Search a large corpus / web for the synonyms.
- Find good sense indicators from the retrieved contexts.

Example

WordNet senses for the word *"sense"*:

1. A general conscious **awareness**.
   (e.g., *a sense of security*)
2. The **meaning** of a word.
   (e.g., *The dictionary gave several senses for the word*)
3. Sound practical **judgment**.
   (e.g., *I can't see the sense in doing it now*)
4. A natural appreciation or **ability**.
   (e.g., *keen musical sense*).

Using WordNet

Semantic Neighbors from WordNet

- **Neighbors of *awareness*:** sentience , sensation, sensitivity, sensitiveness, sensibility, modality, module, knowingness, ...
- **Neighbors of *meaning*:** signified , acceptation, signification, significance, meaning, import, symbolization, symbolisation,...
- **Neighbors of *judgment*:** gumption , logic, sagacity, judgment, judgement, discernment, prudence, judiciousness, eye, ...
- **Neighbors of *ability*:** hold, grasp, appreciation

- few exact synonyms
- many related words
- neighbors are not “substitutable”
- neighbors are themselves polysemous

Neighbor Polysemy

Monosemous Semantic Neighbors

- **Neighbors of *awareness*:** cognisance, self-awareness
- **Neighbors of *meaning*:** signified, signification, nuance, moral, intention

- greatly reduced number of neighbors
- no monosemous neighbors for last two senses
- neighbors may be rare

Our Approach

Distributional Neighbors

- Extension of McCarthy et al. (2004).
- Based on distributional similarity - words are related if used in similar contexts.
- Uses semantic similarity to associate neighbors with senses.

Method Advantages

+ relates directly to context cues
+ domain specific
+ polysemy restricted by similarity

Using Statistics

Distributional Neighbors

- **Neighbors of *awareness*:** awareness, feeling, instinct, enthusiasm, sensation, vision, tradition, consciousness, anger, panic, loyalty
- **Neighbors of *meaning*:** emotion, belief, meaning, manner, necessity, tension, motivation

- No neighbors for last two senses.
- Not prevalent in the corpus (corroborated by the test data).

Associating Neighbors and Senses

Neighbors from a lexical resource are already associated.

Distributional neighbors are not.

- Use semantic similarity on the knowledge base.
  (WordNet::Similarity – Pedersen et al. 2004)
- Choose target sense most similar to *any* sense of the neighbor.

Methodology

1. Acquire “neighbors” - words related to (a sense of) the target
2. Extract instances of neighbors from a large corpus
3. Label instances with associated sense
4. Use labeled data to train supervised classifier

*“... an attempt to state the **meaning** of a word”*

becomes

*“... an attempt to state the **sense** (s#2) of a word.”*

Data

Corpus

The British National Corpus (BNC)

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

Evaluation

Nouns from Senseval 2 & 3 lexical samples

- instances from BNC
- coarse-grain senses

	# Words	Ambiguity	1st Sense
SE-2	25	3.28	65.96%
SE-3	20	4.35	60.90%

Tools

Distributional Neighbors

- dependency based (Lin, 1998)
- co-occurrence based (InfoMap)

Classifiers

Evaluated on a variety of classifiers, from different paradigms:

- SVM - multi-class bound-constrained SVC (Hsu and Lin, 2001)
- Maximum Entropy (Megam, Daumé III 2004)
- Label Propagation (SemiL, Zhu and Ghahramani 2002)

Baselines

- McCarthy et al - predominant sense detection
- Lesk - overlap between context and dictionary definition

Results

Senseval 2	AllWN	MonoWN	Depend	Manual
SVM	48.12%	53.29%	64.38%	72.52%
MaxEnt	40.93%	52.11%	62.32%	71.91%
LP	42.67%	49.54%	63.32%	69.28%
McCarthy	59.98%			
Lesk	48.12%			

Senseval 3	AllWN	MonoWN	Depend	Manual
SVM	53.16%	46.32%	57.47%	71.22%
MaxEnt	49.67%	44.85%	57.35%	71.75%
LP	47.41%	43.60%	60.60%	67.57%
McCarthy	57.14%			
Lesk	48.66%			

Results

Senseval 2	AllWN	MonoWN	Depend	Manual
SVM	48.12%	53.29%	64.38%	72.52%
MaxEnt	40.93%	52.11%	62.32%	71.91%
LP	42.67%	49.54%	63.32%	69.28%
McCarthy	59.98%			
Lesk	48.12%			

Senseval 3	AllWN	MonoWN	Depend	Manual
SVM	53.16%	46.32%	57.47%	71.22%
MaxEnt	49.67%	44.85%	57.35%	71.75%
LP	47.41%	43.60%	60.60%	67.57%
McCarthy	57.14%			
Lesk	48.66%			

Results

Senseval 2	AllWN	MonoWN	Depend	Manual
SVM	48.12%	53.29%	64.38%	72.52%
MaxEnt	40.93%	52.11%	62.32%	71.91%
LP	42.67%	49.54%	63.32%	69.28%
McCarthy	59.98%			
Lesk	48.12%			

Senseval 3	AllWN	MonoWN	Depend	Manual
SVM	53.16%	46.32%	57.47%	71.22%
MaxEnt	49.67%	44.85%	57.35%	71.75%
LP	47.41%	43.60%	60.60%	67.57%
McCarthy	57.14%			
Lesk	48.66%			

Results

Senseval 2	AllWN	MonoWN	Depend	Manual
SVM	48.12%	53.29%	64.38%	72.52%
MaxEnt	40.93%	52.11%	62.32%	71.91%
LP	42.67%	49.54%	63.32%	69.28%
McCarthy	59.98%			
Lesk	48.12%			

Senseval 3	AllWN	MonoWN	Depend	Manual
SVM	53.16%	46.32%	57.47%	71.22%
MaxEnt	49.67%	44.85%	57.35%	71.75%
LP	47.41%	43.60%	60.60%	67.57%
McCarthy	57.14%			
Lesk	48.66%			

Conclusions

- statistics + knowledge-base is better than just knowledge-base
- surpasses state-of-the-art unsupervised methods
- utility similar to supervised framework
  better classifier $\longrightarrow$ better scores

Outline

1. Introduction - Unsupervised NLP
   - The Competition - Supervised Methods
   - Colleagues - Human Knowledge
   - Unsupervised Learning
2. Word Sense Disambiguation (WSD)
   - Unsupervised Labeling
   - Bayesian Sense Induction
3. Work in Progress - Aspect & Sentiment in Reviews
4. Conclusion

Bayesian Sense Induction

Brody and Lapata (2009)

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.

Clustering Approach

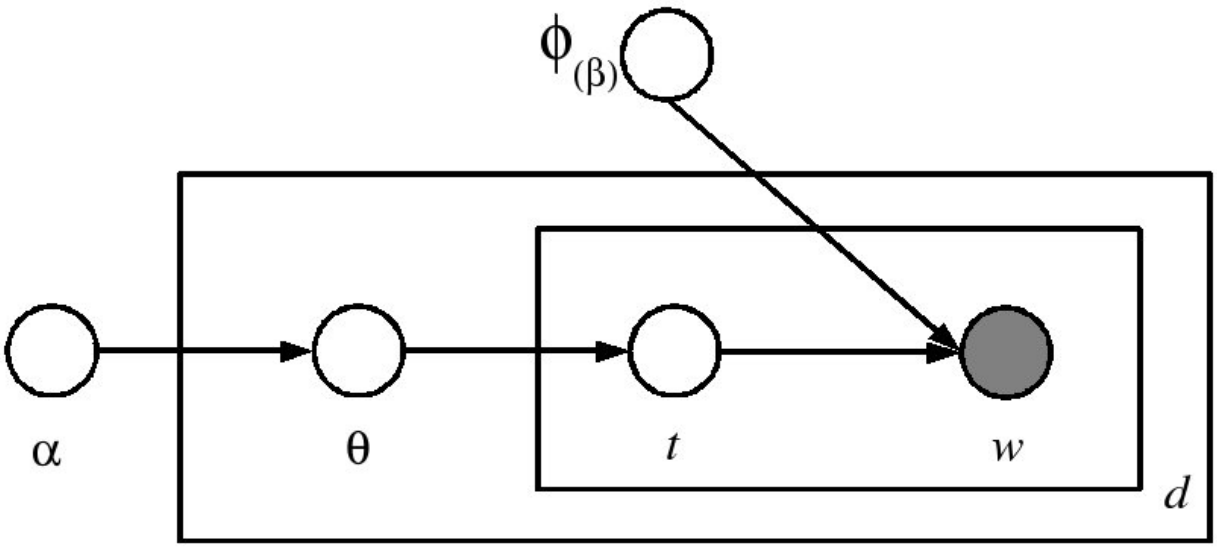
Common Approach : standard clustering task

- 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

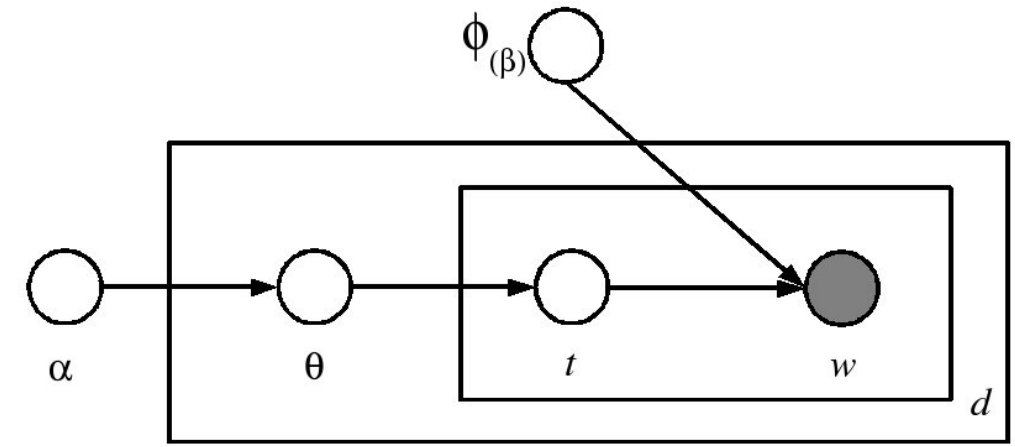
Previous LDA Approaches to WSD

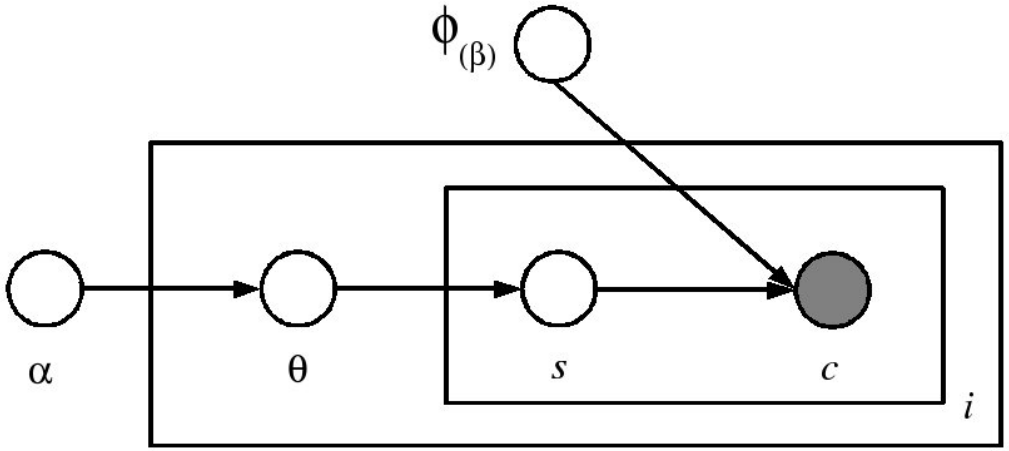
- Supervised - use LDA-derived topics instead of Bag-of-Words. (Cai et al., 2007)
- Unsupervised - integrate distributional similarity approach with LDA. (Boyd-Graber and Blei, 2007)

Problems in Previous Approaches

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

Adapting Classic LDA



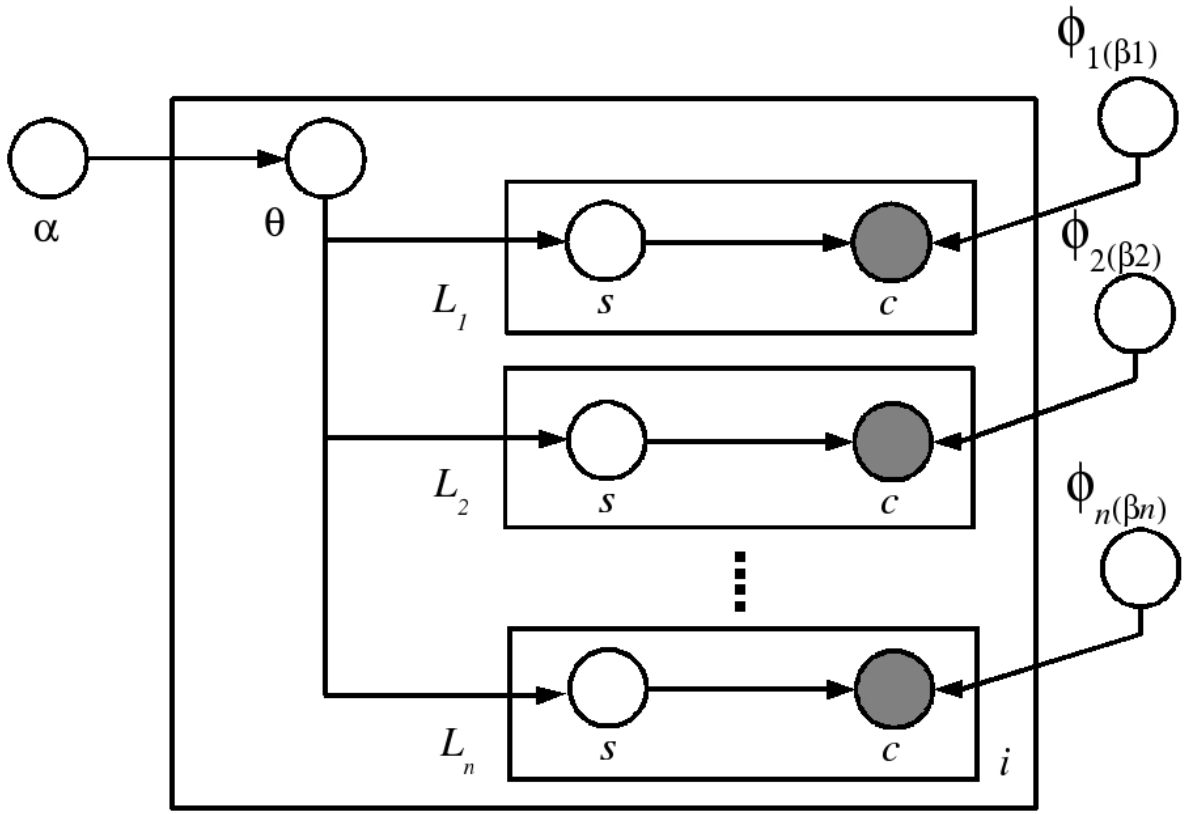


- one model per word
- immediate context instead of whole document
- context elements replace words
- small number of senses (<10)

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



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

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

Induced Senses - 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

Induced Senses - 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%

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
- induced senses match the domain

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
- induced senses match the target domain

Outline

1. Introduction - Unsupervised NLP
   - The Competition - Supervised Methods
   - Colleagues - Human Knowledge
   - Unsupervised Learning
2. Word Sense Disambiguation (WSD)
   - Unsupervised Labeling
   - Bayesian Sense Induction
3. Work in Progress - Aspect & Sentiment in Reviews
4. Conclusion

Online Reviews

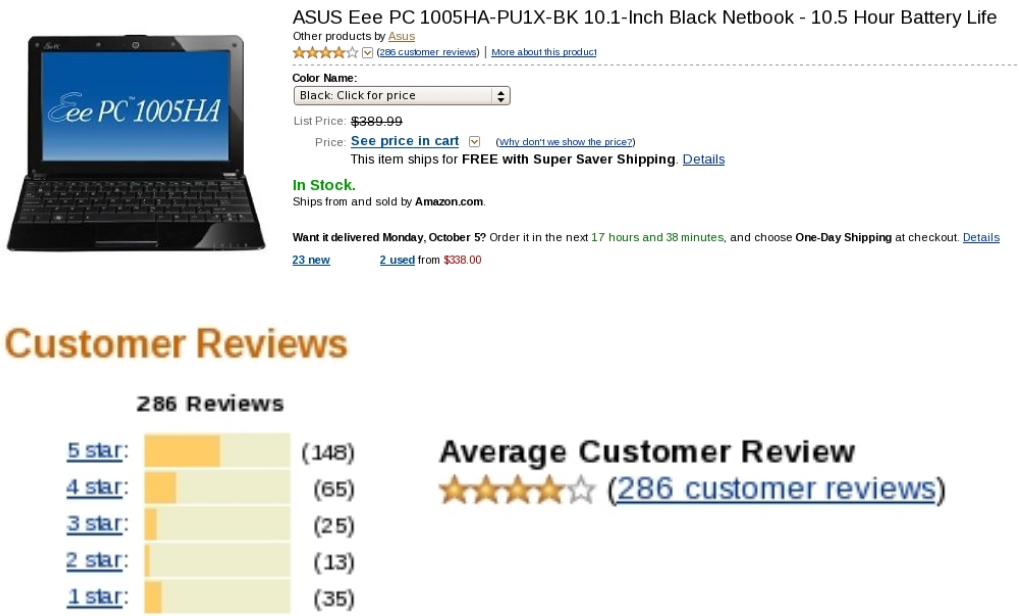
What we have:

- overall score
- details in free-form text

What we want:

- overall score
- summary of details:

Aspect	Score
Weight	80% Pos
Keyboard	50% Pos
Wireless	30% Pos
Battery	70% Pos



Why Unsupervised?

- manual annotation is not feasible
- aspects are unpredictable
- varying ways of expressing sentiment
- spelling errors and typos are an issue for lexicons

Outline

1. Introduction - Unsupervised NLP
   - The Competition - Supervised Methods
   - Colleagues - Human Knowledge
   - Unsupervised Learning
2. Word Sense Disambiguation (WSD)
   - Unsupervised Labeling
   - Bayesian Sense Induction
3. Work in Progress - Aspect & Sentiment in Reviews
4. Conclusion

Summary

- Unsupervised methods
  - \+ don't require annotation
  - \+ fit themselves to the task and data
  - \+ cognitively interesting/plausible
  - – are less accurate
  - – require more thought

Conclusions

- Unsupervised methods *can* be used to achieve good results
- We can harness supervised tools and knowledge resources
- Inducing classes from the data itself is a huge advantage

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.

Brody, Samuel and Mirella Lapata. 2008. Good neighbors make good senses: Exploiting distributional similarity for unsupervised WSD. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Coling 2008 Organizing Committee, Manchester, UK, pages 65–72.

Brody, Samuel and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. The Association for Computer Linguistics, Athens, Greece.

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.

Daumé III, Hal. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`.

Hsu, C. and C. Lin. 2001. A comparison of methods for multi-class support vector machines.

Leacock, Claudia, Martin Chodorow, and George A. Miller. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics* 24(1):147–165.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*. Association for Computational Linguistics, Morristown, NJ, pages 768–774.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42th ACL*. Barcelona, Spain, pages 280–287.

Mihalcea, Rada F. 2002. Word sense disambiguation with pattern learning and automatic feature selection. *Nat. Lang. Eng.* 8(4):343–358.

Pedersen, T., S. Patwardhan, and J. Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In *Human*