Sparsity in Dependency Grammar Induction

Jennifer Gillenwater[1] Kuzman Ganchev[1] João Graça[2]
Ben Taskar[1] Fernando Pereira[3]

[1]Computer & Information Science
University of Pennsylvania

[2]L[2]F INESC-ID, Lisboa, Portugal

[3]Google, Inc.

July 12, 2010

Outline

- A generative dependency parsing model

Outline

- A generative dependency parsing model
- The ambiguity problem this model faces

Outline

- A generative dependency parsing model
- The ambiguity problem this model faces
- Previous attempts to reduce ambiguity

Outline

- A generative dependency parsing model
- The ambiguity problem this model faces
- Previous attempts to reduce ambiguity
- How posteriors provide a good measure of ambiguity

Outline

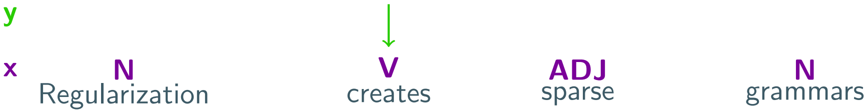
- A generative dependency parsing model
- The ambiguity problem this model faces
- Previous attempts to reduce ambiguity
- How posteriors provide a good measure of ambiguity
- Applying posterior regularization to the likelihood objective

Outline

- A generative dependency parsing model
- The ambiguity problem this model faces
- Previous attempts to reduce ambiguity
- How posteriors provide a good measure of ambiguity
- Applying posterior regularization to the likelihood objective
- Success with respect to EM and parameter prior baselines

Dependency model with valence

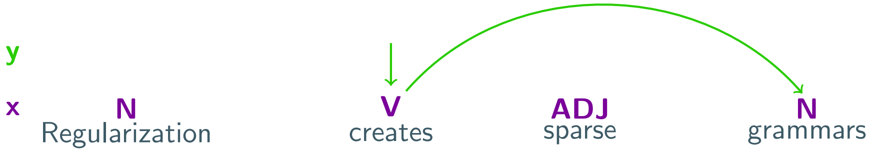
(Klein and Manning, ACL 2004)



$$p_\theta(\mathbf{x}, \mathbf{y}) = \theta_{root(V)}$$

Dependency model with valence

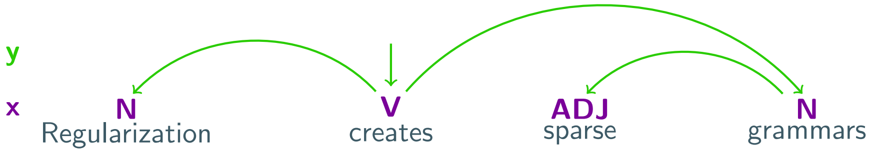
(Klein and Manning, ACL 2004)



$$p_\theta(\mathbf{x}, \mathbf{y}) = \theta_{root(V)}$$
$$\cdot \theta_{stop(nostop|V,right,false)} \cdot \theta_{child(N|V,right)}$$

Dependency model with valence

(Klein and Manning, ACL 2004)



$$p_\theta(\mathbf{x}, \mathbf{y}) = \theta_{root(V)}$$
$$\cdot\theta_{stop(nostop|V,right,false)} \cdot \theta_{child(N|V,right)}$$
$$\cdot\theta_{stop(stop|V,right,true)} \cdot \theta_{stop(nostop|V,left,false)} \cdot \theta_{child(N|V,left)}$$
$$\ldots$$

Traditional objective optimization

- **Traditional objective**: marginal log likelihood

$$\max_{\theta} \mathcal{L}(\theta) = E_X[\log p_\theta(\mathbf{x})] = E_X[\log \sum_{\mathbf{y}} p_\theta(\mathbf{x}, \mathbf{y})]$$

Traditional objective optimization

- **Traditional objective**: marginal log likelihood

$$\max_{\theta} \mathcal{L}(\theta) = E_X[\log p_\theta(\mathbf{x})] = E_X[\log \sum_{\mathbf{y}} p_\theta(\mathbf{x}, \mathbf{y})]$$

- **Optimization method**: expectation maximization (EM)

Traditional objective optimization

- **Traditional objective**: marginal log likelihood

$$\max_{\theta} \mathcal{L}(\theta) = E_X[\log p_\theta(\mathbf{x})] = E_X[\log \sum_{\mathbf{y}} p_\theta(\mathbf{x}, \mathbf{y})]$$

- **Optimization method**: expectation maximization (EM)
- **Problem**: EM may learn a very ambiguous grammar
  - Too many non-zero probabilities
  - Ex: V $\longrightarrow$ N should have non-zero probability,
    but V $\longrightarrow$ DET, V $\longrightarrow$ JJ, V $\longrightarrow$ PRP$, etc. should be 0

Previous approaches to improving performance

- Structural annealing[1]

1. Smith and Eisner, ACL 2006
2. Headden et al., NAACL 2009
3. Liang et al., EMNLP 2007; Johnson et al., NIPS 2007; Cohen et al., NIPS 2008, NAACL 2009

Previous approaches to improving performance

- Structural annealing[1]
- $\mathcal{L}(\theta')$: Model extension[2]

1 Smith and Eisner, ACL 2006

2 Headden et al., NAACL 2009

3 Liang et al., EMNLP 2007; Johnson et al., NIPS 2007; Cohen et al., NIPS 2008, NAACL 2009

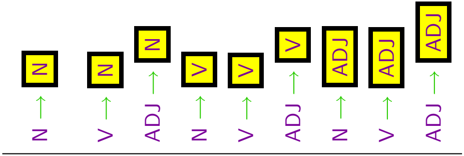
Previous approaches to improving performance

- Structural annealing[1]
- $\mathcal{L}(\theta')$: Model extension[2]
- $\mathcal{L}(\theta) + \log p(\theta)$: Parameter regularization[3]
  - Tend to reduce unique # of children per parent, rather than directly reducing # of unique parent-child pairs
  - $\theta_{child(Y|X,dir)} \neq posterior(X \rightarrow Y)$

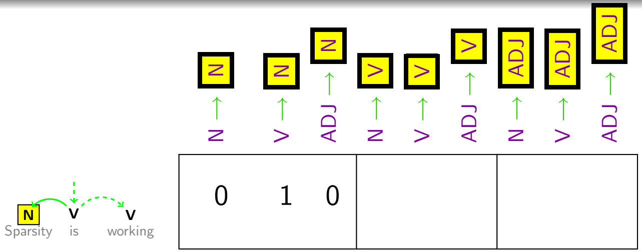
1. Smith and Eisner, ACL 2006
2. Headden et al., NAACL 2009
3. Liang et al., EMNLP 2007; Johnson et al., NIPS 2007; Cohen et al., NIPS 2008, NAACL 2009

Ambiguity measure using posteriors: $L_{1/\infty}$

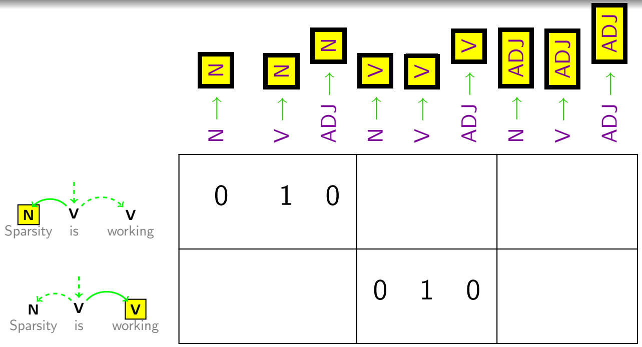
Intuition: True # of unique parent tags for a child tag is small



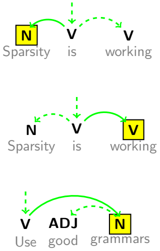
Ambiguity measure using posteriors: $L_{1/\infty}$



Ambiguity measure using posteriors: $L_{1/\infty}$

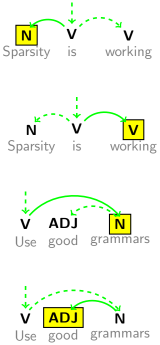


Ambiguity measure using posteriors: $L_{1/\infty}$



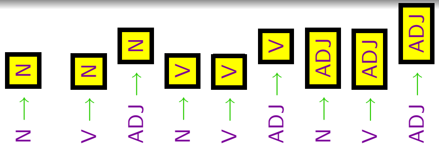
N → N	V → N	ADJ → N	N → V	V → V	ADJ → V	N → ADJ	V → ADJ	ADJ → ADJ
0	1	0						
			0	1	0			
0	1	0						

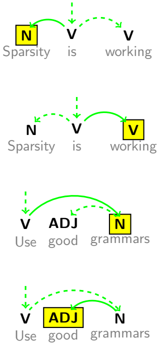
Ambiguity measure using posteriors: $L_{1/\infty}$



N → N	V → N	ADJ → N	N → V	V → V	ADJ → V	N → ADJ	V → ADJ	ADJ → ADJ
0	1	0						
			0	1	0			
0	1	0						
						1	0	0

Ambiguity measure using posteriors: $L_{1/\infty}$





0	1	0						
			0	1	0			
0	1	0						
						1	0	0

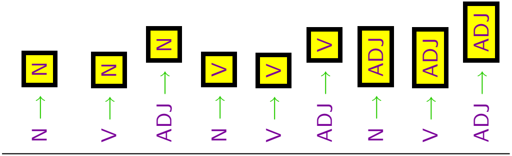
max ↓

0	1	0	0	1	0	1	0	0

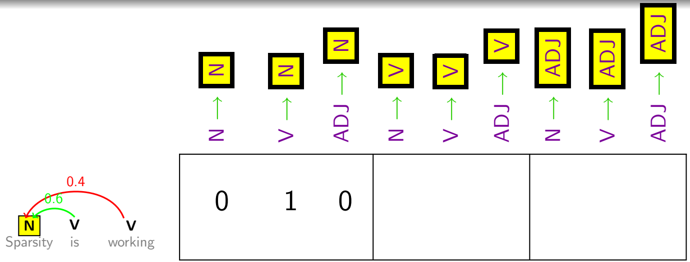
sum $= 3 \leftarrow$

Measuring ambiguity on distributions over trees

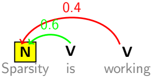
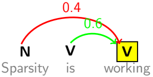
For a distribution $p_\theta(\mathbf{y} \mid \mathbf{x})$ instead of gold trees:



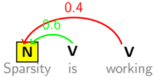
Measuring ambiguity on distributions over trees

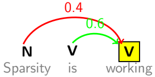


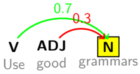
Measuring ambiguity on distributions over trees

	N → N	V → N	ADJ → N	N → V	V → V	ADJ → V	N → ADJ	V → ADJ	ADJ → ADJ
	0	1	0						
				.4	.6	0			

Measuring ambiguity on distributions over trees

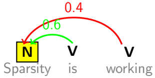


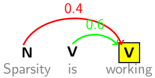


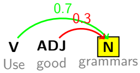


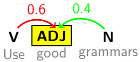
	N → N	V → N	ADJ → N	N → V	V → V	ADJ → V	N → ADJ	V → ADJ	ADJ → ADJ
	0	1	0						
				.4	.6	0			
	0	.7	.3						

Measuring ambiguity on distributions over trees



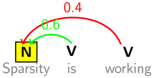
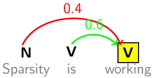
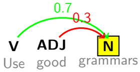






N → N	V → N	ADJ → N	N → V	V → V	ADJ → V	N → ADJ	V → ADJ	ADJ → ADJ
0	1	0						
			.4	.6	0			
0	.7	.3						
						.4	.6	0

Measuring ambiguity on distributions over trees

	N → N	V → N	ADJ → N	N → V	V → V	ADJ → V	N → ADJ	V → ADJ	ADJ → ADJ
	0	1	0						
				.4	.6	0			
	0	.7	.3						
							.4	.6	0

max ↓

sum = 3.3 ←	0	1	.3	.4	.6	0	.4	.6	0

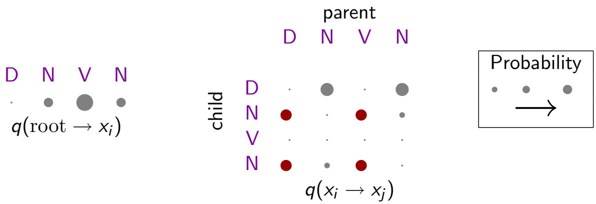
Minimizing ambiguity through posterior regularization

E-Step $$q^t(\mathbf{y} \mid \mathbf{x}) = \arg\min_{q(\mathbf{y}|\mathbf{x})} KL(q \parallel p_{\theta^t})$$

Minimizing ambiguity through posterior regularization

E-Step $q^t(\mathbf{y} \mid \mathbf{x}) = \underset{q(\mathbf{y}|\mathbf{x})}{\arg\min}\, KL(q \parallel p_{\theta^t})$

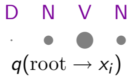
$q(\mathbf{y} \mid \mathbf{x}) =$

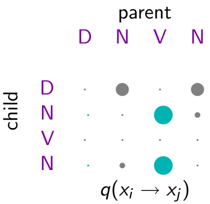


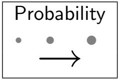
Apply E-step penalty $L_{1/\infty}$ on posteriors $q(\mathbf{y} \mid \mathbf{x})$ to induce sparsity (Graca et al., NIPS 2007 & 2009)

$$\textbf{E-Step} \qquad q^t(\mathbf{y} \mid \mathbf{x}) = \underset{q(\mathbf{y}|\mathbf{x})}{\arg\min} \, KL(q \parallel p_{\theta^t}) + \sigma L_{1/\infty}(q(\mathbf{y} \mid \mathbf{x}))$$

$q(\mathbf{y} \mid \mathbf{x}) =$







- English from Penn Treebank: state-of-the-art accuracy

Learning Method	Accuracy		
	≤ 10	≤ 20	all
PR ($\sigma = 140$)	**62.1**	**53.8**	**49.1**
LN families	59.3	45.1	39.0
SLN TieV & N	61.3	47.4	41.4
PR ($\sigma = 140$, $\lambda = 1/3$)	64.4	55.2	50.5
DD ($\alpha = 1$, λ learned)	**65.0 (**$\pm$5.7**)**		

Experimental results

- English from Penn Treebank: state-of-the-art accuracy

Learning Method	Accuracy		
	≤ 10	≤ 20	all
PR ($\sigma = 140$)	**62.1**	**53.8**	**49.1**
LN families	59.3	45.1	39.0
SLN TieV & N	61.3	47.4	41.4
PR ($\sigma = 140$, $\lambda = 1/3$)	64.4	55.2	50.5
DD ($\alpha = 1$, λ learned)	**65.0** (**$\pm$5.7**)		

- 11 other languages from CoNLL-X:
  - Dirichlet prior baseline: 1.5% average gain over EM
  - Posterior regularization: 6.5% average gain over EM

Experimental results

- English from Penn Treebank: state-of-the-art accuracy

Learning Method	Accuracy		
	≤ 10	≤ 20	all
PR ($\sigma = 140$)	**62.1**	**53.8**	**49.1**
LN families	59.3	45.1	39.0
SLN TieV & N	61.3	47.4	41.4
PR ($\sigma = 140$, $\lambda = 1/3$)	64.4	55.2	50.5
DD ($\alpha = 1$, λ learned)	**65.0** (±5.7)		

- 11 other languages from CoNLL-X:
  - Dirichlet prior baseline: 1.5% average gain over EM
  - Posterior regularization: 6.5% average gain over EM
- **Come see the poster for more details**