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.

December 11, 2009

Dependency model with valence

(Klein and Manning, ACL 2004)

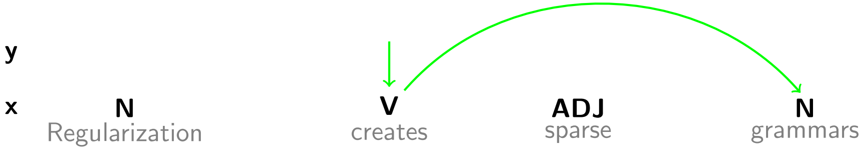
y

x **N** Regularization **V** creates **ADJ** sparse **N** grammars

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

Dependency model with valence

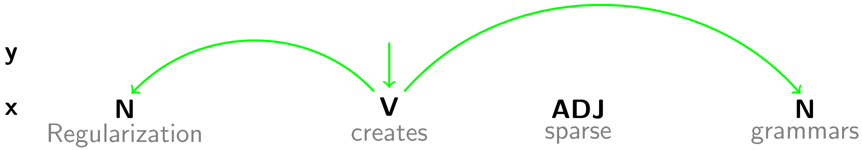
(Klein and Manning, ACL 2004)



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

Dependency model with valence

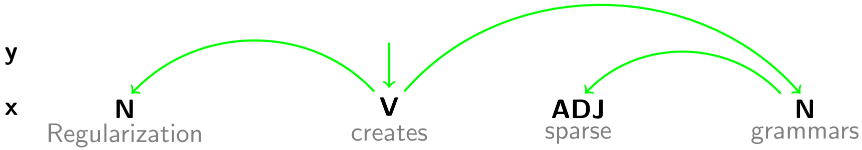
(Klein and Manning, ACL 2004)



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

Dependency model with valence

(Klein and Manning, ACL 2004)



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

Outline

- A problem this model faces
- A measure of parent-child pair sparsity
- A modification to the objective
- How this modification improves parsing accuracy

Traditional objective optimization

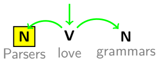
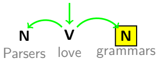
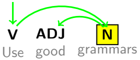
- **Traditional objective**: marginal log likelihood

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

- **Optimization method**: expectation maximization (EM)
- **Problem**: grammar is very permissive; EM may learn a grammar that is not concise
- Can we precisely define "concise", so that we can incorporate it into the objective?

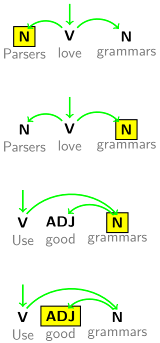
A measure of sparsity

Intuition: True # of unique (parent, child) POS tag pairs is small

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

A measure of sparsity

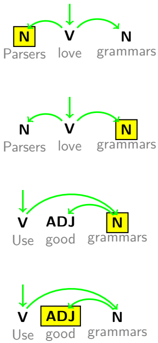
Intuition: True # of unique (parent, child) POS tag pairs is small



	N→N	V→N	ADJ→N	N→ADJ	V→ADJ	ADJ→ADJ
Parsers love grammars (N)	0	1	0			
Parsers love grammars (N)	0	1	0			
Use good grammars (N)	0	1	0			
Use good grammars (ADJ)				1	0	0
max ↓	0	1	0	1	0	0

A measure of sparsity

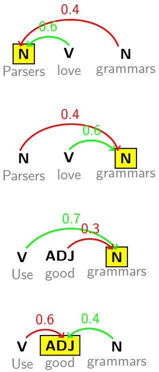
Intuition: True # of unique (parent, child) POS tag pairs is small



	N→N	V→N	ADJ→N	N→ADJ	V→ADJ	ADJ→ADJ
	0	1	0			
	0	1	0			
	0	1	0			
				1	0	0
max		↓		↓		
sum = 2 ←	0	1	0	1	0	0

Measuring sparsity on distributions over trees

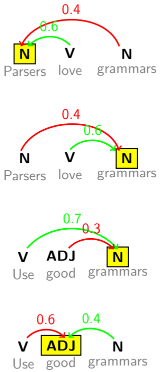
For a distribution $p_\theta(\mathbf{y} \mid \mathbf{x})$ instead of gold trees: Restate sparsity measure over edge expectations (*posterior* probabilities)



N→N	V→N	ADJ→N	N→ADJ	V→ADJ	ADJ→ADJ
0.4	0.6	0			
0.4	0.6	0			
0	0.7	0.3			
			0.4	0.6	0

Measuring sparsity on distributions over trees

For a distribution $p_\theta(\mathbf{y} \mid \mathbf{x})$ instead of gold trees: Restate sparsity measure over edge expectations (*posterior* probabilities)

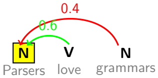


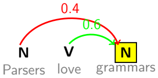
N→N	V→N	ADJ→N	N→ADJ	V→ADJ	ADJ→ADJ
0.4	0.6	0			
0.4	0.6	0			
0	0.7	0.3			
			0.4	0.6	0

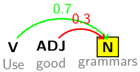
max ↓ ↓

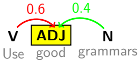
0.4	0.7	0.3	0.4	0.6	0

For a distribution $p_\theta(\mathbf{y} \mid \mathbf{x})$ instead of gold trees: Restate sparsity measure over edge expectations (*posterior* probabilities)





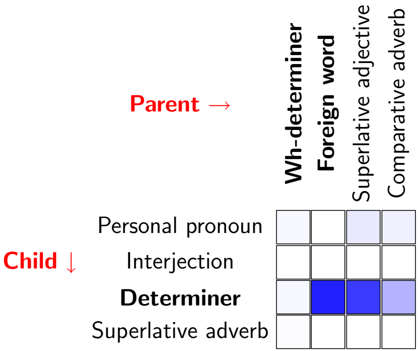




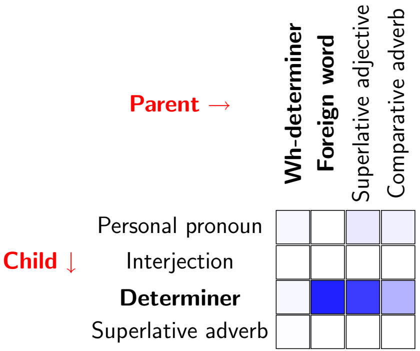
	N→N	V→N	ADJ→N	N→ADJ	V→ADJ	ADJ→ADJ
Parsers love grammars (1)	0.4	0.6	0			
Parsers love grammars (2)	0.4	0.6	0			
Use good grammars (1)	0	0.7	0.3			
Use good grammars (2)				0.4	0.6	0
max ↓	0.4	0.7	0.3	0.4	0.6	0

sum = 2.4 ←

Example partial edge types table



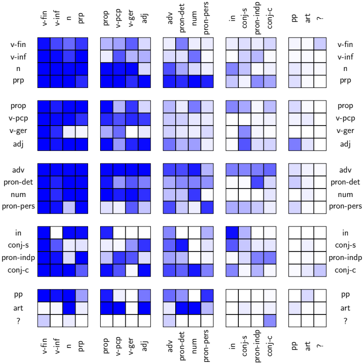
Example partial edge types table



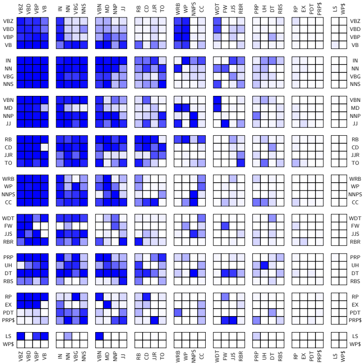
- In at least one sentence, foreign word → determiner has high posterior probability
- Wh-determiners never dominate determiners

Edge type tables for supervised initialization

White $\rightarrow$ max $= 0$



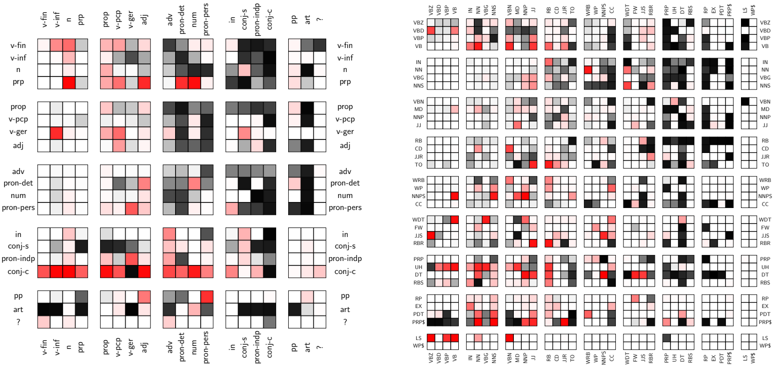
Portuguese



English

Edge type tables for EM

- Red → max posterior < supervised
- Black → max > supervised; much black implies model assigns non-zero probability to too many pair types



Portuguese **English**

Previous approaches to improving performance

- Structural annealing to constrain dependency lengths (Smith and Eisner, ACL 2006)
- Model extension (Headden et al., NAACL 2009): $\mathcal{L}(\theta')$
- Parameter regularization: $\mathcal{L}(\theta) + \log p(\theta)$

Priors

- Discounting Dirichlet prior (Headden et al., ACL 2009)
- Logistic normal prior (Cohen et al., NIPS 2008; Cohen and Smith, NAACL 2009)
- Hierarchical Dirichlet processes (Liang et al., EMNLP 2007; Johnson et al., NIPS 2007)
- All of the above cut down on # of children, but we really want to cut down on # of parent-child pairs

$$\begin{aligned} \theta_{\mathit{child}|\mathit{parent}} &\neq \max(\mathit{posterior}_{\mathit{parent},\mathit{child}}) \\ \text{parameters} &\neq \text{posteriors} \end{aligned}$$

(Graca et al., NIPS 2007 & 2009)

Posterior regularization (PR): Minimize number of unique parent-child pairs directly through E-step penalty term on the posteriors $q(\mathbf{y} \mid \mathbf{x})$

Direct approach to sparsity problem

(Graca et al., NIPS 2007 & 2009)

Posterior regularization (PR): Minimize number of unique parent-child pairs directly through E-step penalty term on the posteriors $q(\mathbf{y} \mid \mathbf{x})$

M-Step $$\theta^{t+1} = \arg\max_{\theta} E_X \left[\sum_{\mathbf{y}} q^t(\mathbf{y} \mid \mathbf{x}) \log p_\theta(\mathbf{x}, \mathbf{y}) \right]$$

Direct approach to sparsity problem

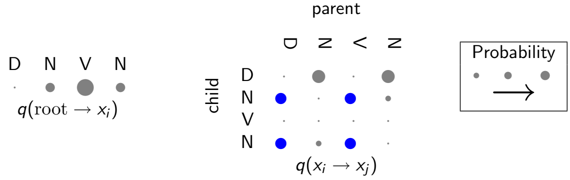
(Graca et al., NIPS 2007 & 2009)

Posterior regularization (PR): Minimize number of unique parent-child pairs directly through E-step penalty term on the posteriors $q(\mathbf{y} \mid \mathbf{x})$

M-Step $$\theta^{t+1} = \arg\max_{\theta} E_X \left[\sum_{\mathbf{y}} q^t(\mathbf{y} \mid \mathbf{x}) \log p_\theta(\mathbf{x}, \mathbf{y}) \right]$$

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

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



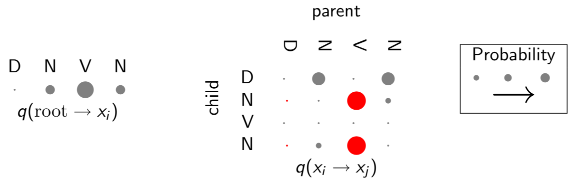
(Graca et al., NIPS 2007 & 2009)

Posterior regularization (PR): Minimize number of unique parent-child pairs directly through E-step penalty term on the posteriors $q(\mathbf{y} \mid \mathbf{x})$

M-Step $$\theta^{t+1} = \arg\max_{\theta} E_X \left[\sum_{\mathbf{y}} q^t(\mathbf{y} \mid \mathbf{x}) \log p_\theta(\mathbf{x}, \mathbf{y}) \right]$$

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

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

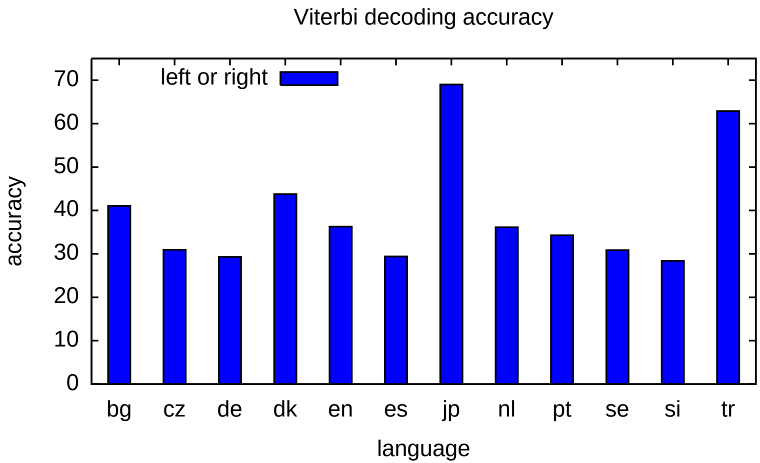


Experimental setup

- 12 languages: 11 from CoNLL-X shared task, English from Penn Treebank
- Processing of train and test sets: strip punctuation, consider only sentences of length ≤ 10
- For training, also eliminate sentences of length ≤ 3 to increase model stability
- Assume POS tags given (but no parse trees)
- Initialize model as in Klein and Manning, ACL 2004

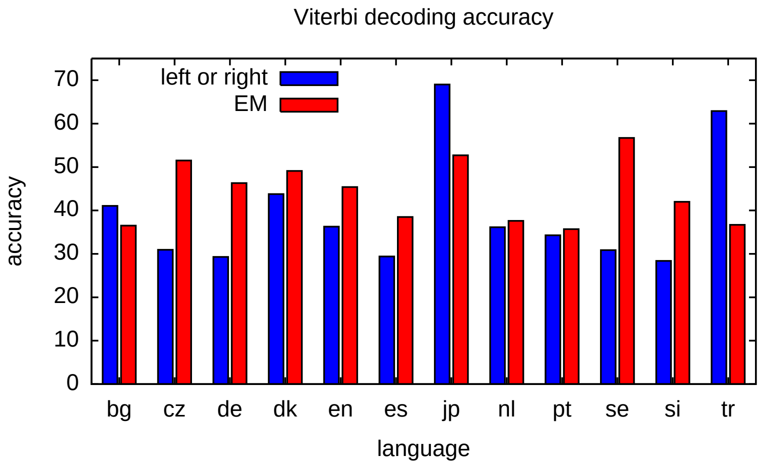
Baseline

Baseline: best of link-left, link-right



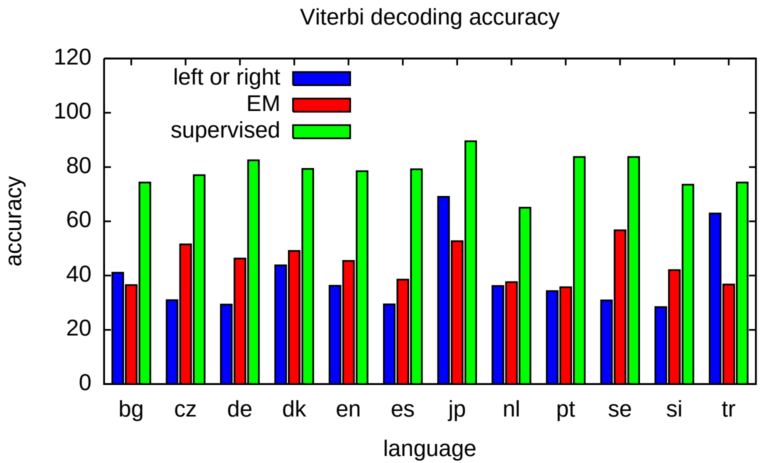
Baseline vs. EM

Baseline wins by a lot on the verb-final languages



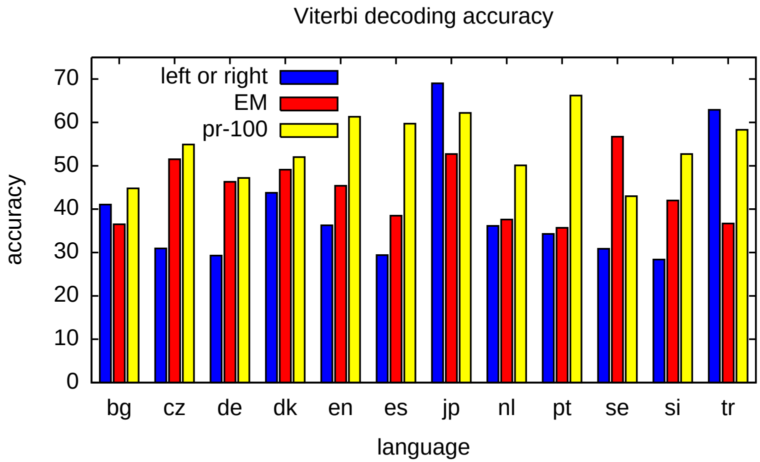
Baseline vs. EM vs. Supervised

And all models are well below supervised performance

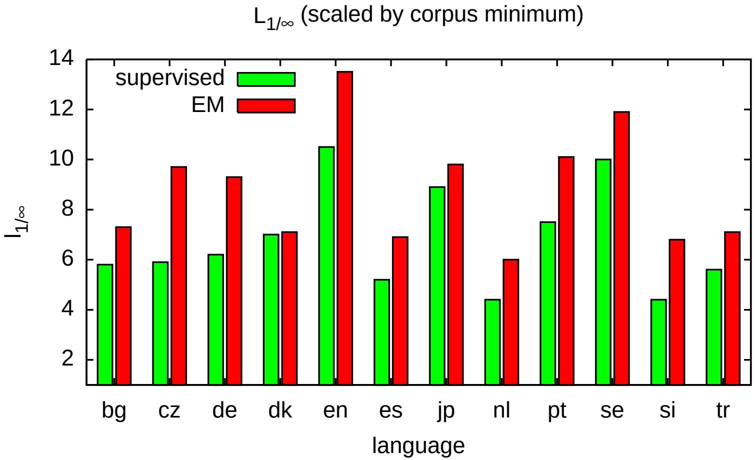


Sparsity's impact on accuracy

- Improve over EM in 11/12 cases
- Average of 10.3% accuracy increase

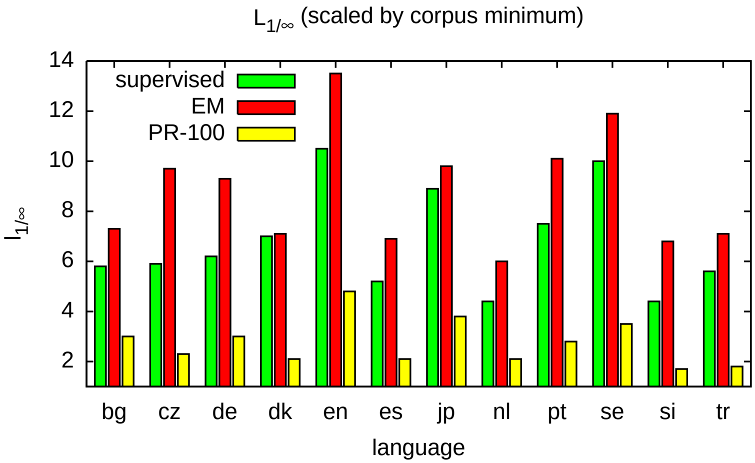


Sparsity measure for supervised vs. EM

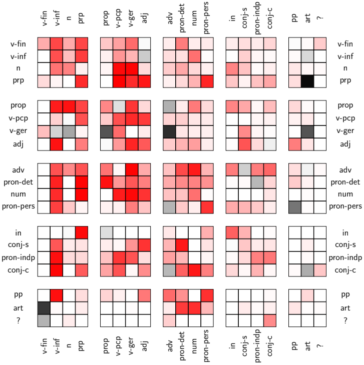


Sparsity measure for PR

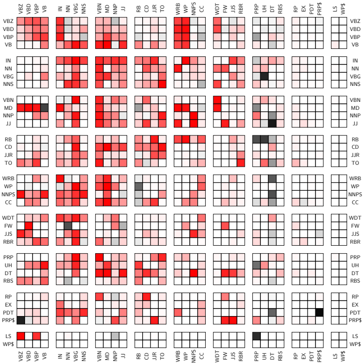
Regularization strength $\sigma = 100$



Mostly red $\longrightarrow$ oversparsification

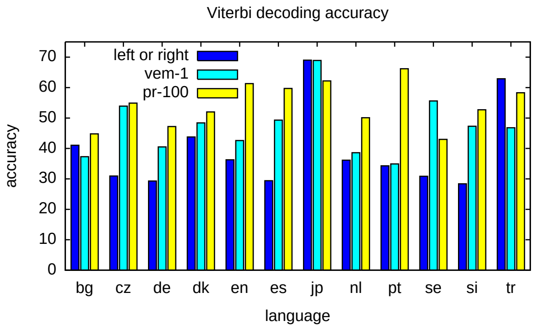


Portuguese



English

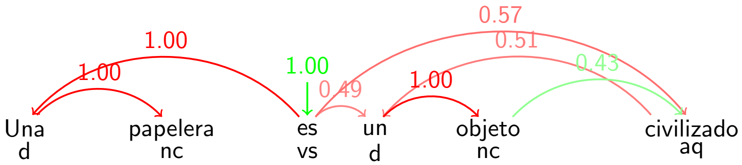
Comparison to discounting Dirichlet prior



- PR outperforms discounting Dirichlet prior in 10/12 cases
- Dirichlet prior has higher number of unique parent-child pairs in expectation than supervised, for all languages
- PR performance comparable to shared logistic normal prior (Cohen and Smith, NAACL 2009) on English; 61.3% for logistic vs. 62% for PR

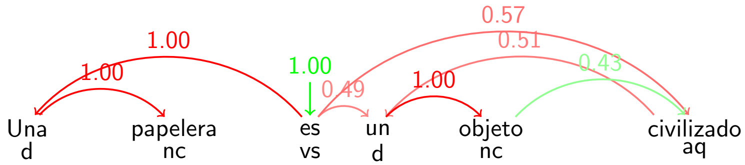
Spanish parse analysis

EM:

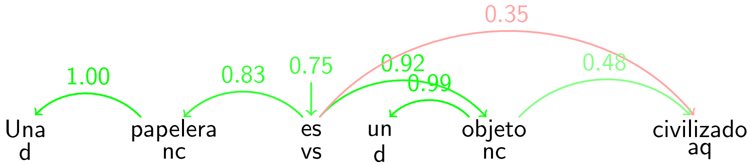


Spanish parse analysis

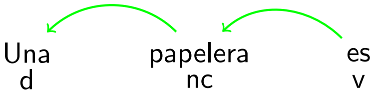
EM:

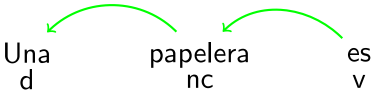
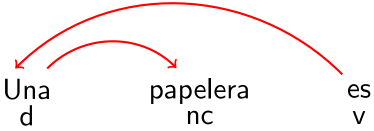


PR:



Spanish parse analysis

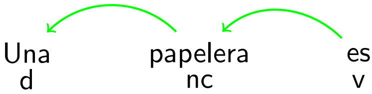
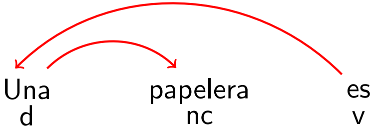
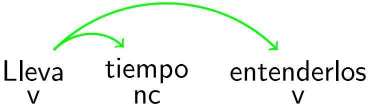
Parse	Unique parent-child pairs
	(v, nc); (nc, d)

Parse	Unique parent-child pairs
	(v, nc); (nc, d)
	(v, d); (d, nc)

Spanish parse analysis

Parse	Unique parent-child pairs
Una (d) papelera (nc) es (v)	(v, nc); (nc, d)
Una (d) papelera (nc) es (v)	(v, d); (d, nc)
Lleva (v) tiempo (nc) entenderlos (v)	(v, nc); (v, v)

Spanish parse analysis

Parse	Unique parent-child pairs
	(v, nc); (nc, d)
	(v, d); (d, nc)
	(v, nc); (v, v)

- Parses 1 and 3 $\longrightarrow$ 3 unique pairs total
- Parses 2 and 3 $\longrightarrow$ 4 unique pairs total

Summary

- **Problem**: Supervised model exhibits fewer unique parent-child pairs than EM model
- **Proposed solution**: Use posterior regularization to decrease expected number of such pairs through an E-step penalty term
- **Result**: Positive impact on accuracy in 11/12 cases

Future work

- Tendency to oversparsify, but reducing σ too much has negative impact on accuracy
- More sparsity in a different aspect of the grammar?
- Sparsity constraint may provide enough guidance to allow for much more complicated models
- Joint induction of POS and parse trees