# Sparsity in Grammar Induction

**Jennifer Gillenwater**
Computer & Information Science
University of Pennsylvania

**Kuzman Ganchev**
Computer & Information Science
University of Pennsylvania

**João Graça**
$L^2F$ INESC-ID
Lisboa, Portugal

**Ben Taskar**
Computer & Information Science
University of Pennsylvania

**Fernando Pereira**
Google, Inc.

## Abstract

We explore the role of sparsity in unsupervised dependency parser grammar induction by exploiting a common trend in many languages: the number of unique combinations of pairs of part-of-speech (POS) tags for child-parent relationships is relatively small. We express this bias using the posterior regularization (PR) framework [6] and experiment with English, Spanish and Bulgarian. We show that our system achieves significant improvement in unlabeled accuracy over the standard expectation maximization (EM) baseline.

## 1 Introduction

In this paper we explore the problem of biasing unsupervised dependency parsing models to favor a novel kind of sparsity of dependencies: the number of unique combinations of POS tags of child-parent relationships is small. Recent work [7] has shown that a similar kind of sparsity is very effective for POS induction. In this paper we build on those ideas, and apply the concept of sparsity to dependency parsing.

We consider the problem of unsupervised grammar induction from a POS-tagged corpus. For concreteness, define the *type* of a dependency edge to be the pair of parent-child POS tags that it connects. A property shared by many languages and annotation styles is that many such parent-child pairs never occur. For example, it is ungrammatical for nouns to dominate verbs, adjectives to dominate adverbs, and so forth.

These hard constraints are central to grammatically, but very difficult to learn using latent variable models in the absence of labeled data. Previous work in the context of POS tag induction has tried to learn such sparsity structure by introducing an improper prior on model parameters, with limited success. One potential explanation for the limited success of this approach is that, in general, linguistic phenomena often display heavy-tailed distributions. Consequently, we must be careful not to destroy the heavy tail of the distribution in our attempt to learn the sparsity structure of the grammar.

To do so, instead of encouraging sparsity of the *parameters*, we will try to encourage sparsity in the *posteriors*: the model should try to explain the data using only a small number of edge types. We will do this by augmenting the maximum likelihood objective of a latent variable model by a penalty term designed to encourage sparsity of the posteriors. The penalty term has the property that if a particular edge type is very strongly supported by the data in some instance, then we will not pay an additional cost for hypothesizing it elsewhere. By contrast, when we have a choice between explaining some observations using an edge type we have observed before and one we have never observed before, we will strongly prefer to use the edge type we have observed before.

This formulation is more flexibile because it allows the data to guide which parameter values we change to achieve the sparsity structure.

Section 2 defines the generative model, for dependency parsing with valence. Section 3 explains the general theory behind learning with PR constraints and how to encode posterior sparsity under the PR framework for this model. Section 4 describes the results of dependency parsing experiments across 3 languages. Section 5 discusses related work, and Section 6 concludes.

## 2 Parsing Model

We use a generative parsing model, the dependency model with valence of Klein and Manning [9]. Under this model, the probability of a particular parse $\mathbf{y}$ and a sentence with POS tags $\mathbf{x}$ is given by

$$p_\theta(\mathbf{y},\mathbf{x}) = p_{\text{root}}(r(\mathbf{x})) \prod_{y\in\mathbf{y}} p_{\neg\text{stop}}(y_p, y_d, v_y) p_{\text{child}}(y_p, y_d, y_c) \prod_{x\in\mathbf{x}} p_{\text{stop}}(x, \text{left}, v_l)\, p_{\text{stop}}(x, \text{right}, v_r) \tag{1}$$

where $r(\mathbf{x})$ is the POS tag of the root of the parse tree $\mathbf{y}$, $y$ is an edge from parent $y_p$ to child $y_c$ in direction $y_d$, either left or right, and $v_y$ indicates valency—false if $y_p$ has no other children further from it in direction $y_d$ than $y_c$, true otherwise. The valencies $v_r/v_l$ are marked as true if $x$ has any children further to the left/right than $x$ in $\mathbf{y}$, false otherwise.

## 3 Learning with Posterior Regularization

In order to express the desired preference for posterior sparsity, we use the posterior regularization (PR) framework [6], which incorporates side information into parameter estimation in the form of linear constraints on posterior expectations. This allows tractable learning and inference even when the constraints would be intractable to encode directly in the model, for instance to enforce that each hidden state in an HMM is used only once in expectation. Moreover, PR can represent prior knowledge that cannot be easily expressed as priors over model parameters, like the constraint used in this paper. PR can be seen as a penalty on the standard marginal likelihood objective, which we define first:

$$\textbf{Neg. Marginal Log Likelihood:}\;\; \mathcal{L}(\theta) = \widehat{\mathbf{E}}[-\log p_\theta(\mathbf{x})] = \widehat{\mathbf{E}}[-\log \sum_{\mathbf{z}} p_\theta(\mathbf{z},\mathbf{x})]$$

over the parameters $\theta$, where $\widehat{\mathbf{E}}$ is the empirical expectation over the unlabeled sample $\mathbf{x}$, and $\mathbf{z}$ represents the hidden states. This standard objective may also be regularized with a parameter prior $-\log p(\theta) = C(\theta)$, for example a Dirichlet.

Posterior information in PR is specified with sets $\mathcal{Q}_\mathbf{x}$ of distributions over the hidden variables $\mathbf{z}$ defined by linear constraints on feature expectations:

$$\mathcal{Q}_\mathbf{x} = \{q(\mathbf{z} \mid \mathbf{x}) : \mathbf{E}_q[\mathbf{f}(\mathbf{x},\mathbf{z})] \le \mathbf{b}\}. \tag{2}$$

The marginal log-likelihood of a model is then penalized with the KL-divergence between the desired distributions $\mathcal{Q}_\mathbf{x}$ and the model, $\text{KL}(\mathcal{Q}_\mathbf{x} \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) = \min_{q\in\mathcal{Q}_\mathbf{x}} \text{KL}(q(\mathbf{z}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x}))$. The revised learning objective minimizes:

$$\textbf{PR Objective:}\;\; \mathcal{L}(\theta) + C(\theta) + \widehat{\mathbf{E}}[\text{KL}(\mathcal{Q}_\mathbf{x} \parallel p_\theta(\mathbf{z} \mid \mathbf{x}))]. \tag{3}$$

It will be convenient to use a soft version of this constrained objective for those constraints that we introduce in the next section. This just requires replacing the constraint set $\mathcal{Q}_\mathbf{x} : \mathbf{E}_q[\mathbf{f}] \le \mathbf{b}$, with a penalty term $R(\mathbf{b})$ and a soft constraint $\mathbf{E}_q[\mathbf{f}] \le \mathbf{b}$. For dependency parsing, $R(\mathbf{b})$ encourages the number of edge types encountered in the entire corpus to be small in the projected posteriors $q$. The overall objective is then:

$$\arg\min_{\theta,q,\mathbf{b}}\;\; \mathcal{L}(\theta) + \widehat{\mathbf{E}}[\text{KL}(q \parallel p_\theta) + R(\mathbf{b})] \quad \text{s.t.} \quad \mathbf{E}_q[\mathbf{f}(\mathbf{x},\mathbf{z})] \le \mathbf{b}. \tag{4}$$

See Graça et al. [7] for details on the soft version of the PR framework as well as for the optimization algorithm we use.

### 3.1 $\ell_1/\ell_\infty$ regularization

We now explain how we choose the posterior constraint regularizer $R(\mathbf{b})$ to encourage sparsity in the types of edges used by our model. Consider numbering the potential parse edges in the following way: identify the POS tag $c$ of a child and POS tag $p$ of a parent, and fix an arbitrary ordering of all $p \rightarrow c$ edges. Each possible edge can then be uniquely identified by $c$, $p$ and its index $i$ in the ordering. Let the feature $f_{cpi}$ have value 1 whenever edge of type $p \rightarrow c$ at index $i$ is included in a parse tree.

We would like our model to use only a few different types of edges $(c, p)$. This can be achieved if it "costs" a lot to predict an edge of type $(c, p)$ for the first time but once that happens, it should be "free" for other times we predict an edge of the same type. More precisely, we would like the sum ($\ell_1$ norm) over edge types $(c, p)$ of the maxima ($\ell_\infty$ norm) of the expectation of using such an edge to be small. Formally, this is expressed by the objective:

$$\min_{q,\xi_{cp}} \; \mathrm{KL}(q \parallel p_\theta) + \sigma \sum_{cp} \xi_{cp} \quad \mathrm{s.\,t.} \quad \mathbf{E}_q[f_{cpi}] \le \xi_{cp} \tag{5}$$

where $\sigma$ is the strength of the regularization. The dual of this objective has a very simple form:

$$\max_{\lambda \ge 0} \; -\log\left(\sum_z p_\theta(\mathbf{z}) \exp(-\boldsymbol{\lambda} \cdot \mathbf{f}(\mathbf{z}))\right) \quad \mathrm{s.\,t.} \quad \sum_i \lambda_{cpi} \le \sigma \tag{6}$$

where $\mathbf{z}$ ranges over sets of parse trees for the entire corpus, $\mathbf{f}(\mathbf{z})$ is the vector of $f_{cpi}$ feature values for assignment $\mathbf{z}$, $\boldsymbol{\lambda}$ is the vector of dual parameters $\lambda_{cpi}$, and the primal parameters are $q(\mathbf{z}) \propto p_\theta(\mathbf{z}) \exp\left(-\boldsymbol{\lambda} \cdot \mathbf{f}(\mathbf{z})\right)$. This can be computed via projected gradient, as described by Bertsekas [2].

When $\sigma$ is zero, the projection is an identity mapping and the algorithm reduces to EM. As $\sigma \rightarrow \infty$, the constraints force each occurrence of an edge of type $(c, p)$ to have the same probability of being predicted. For intermediate values of $\sigma$, we prefer to lower the confidence of the highest probability edges of each type. For types that are supported by many examples, this pressure is distributed among many edges and has little effect. For types that are supported by even a single instance with probability 1 of occurrence, the other edges of that same type do not feel pressure. Finally for edge types that are supported only weakly and only by a few instances, we will prefer to lower their probability.

## 4 Experiments

We evaluate our models on the English, Bulgarian and Spanish corpora from the CoNLL X shared task. Following the example of [15], we strip punctuation from the sentences and keep only those sentences that are of length $\le 10$. Longer sentences and punctuation tend to confuse the model more. Some statistics about the corpora are given in Table 1.

| | English | Bulgarian | Spanish |
|---|---|---|---|
| tags | 34 | 11 | 17 |
| word types | 7501 | 9982 | 1366 |
| word tokens | 37746 | 27878 | 2722 |
| sentences | 5458 | 4811 | 476 |

Table 1: Basic statistics showing the relative sizes of the corpora and their tagsets. This includes all sentences that are of length $\le 10$ after punctuation is stripped.

Models are judged based on attachment accuracy—the fraction of words assigned the correct parent. As smoothing we add a very small backoff probability of $4.5 \times 10^{-5}$ to each learned parameter. Figure 1 (a) compares the accuracy of training the model using normal EM vs using PR with sparse constraints. Using sparse constraints greatly improves the model accuracy for Bulgarian and Spanish In fact in the Spanish corpus it brings the accuracy halfway closer to the fully supervised model accuracy. Enforcing sparsity doesn't improve over standard EM in the case of English. We're still investigating why this is the case. Figure1 (b) shows the $\ell_1/\ell_\infty$for the same models. Using sparsity constraints brings this score much closer to the supervised model, in all cases.

|  | Accuracy |  |  | $\ell_1/\ell_\infty$ |  |  |
|---|---|---|---|---|---|---|
|  | English | Bulgarian | Spanish | English | Bulgarian | Spanish |
| EM | 48.4 | 42.6 | 37.2 | 19.43 | 9.35 | 9.97 |
| PR | 45.8 | 54.8 | 62.8 | 9.56 | 6.62 | 3.91 |
| Supervised | 80.6 | 75.8 | 79.7 | 10.96 | 6.36 | 5.54 |

Figure 1: Preliminary results comparing three different training scenarios by attachment accuracy and $\ell_1/\ell_\infty$. **EM**: the EM algorithm. **PR**: our method with $\sigma = 100$. **Supervised**: using max likelihood parameter estimates based on the gold labels. EM and PR use the initialization heuristic from [9] and were run for 100 iterations.

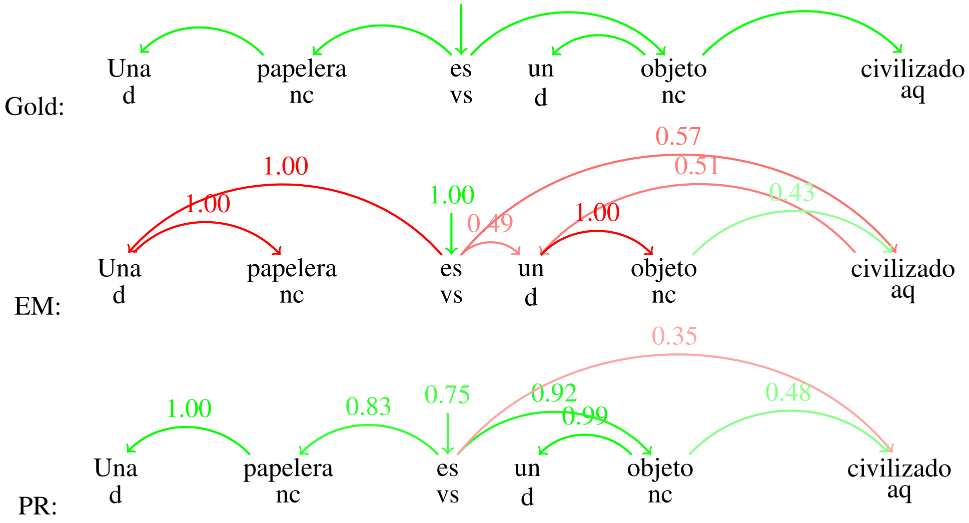

Figure 2: Parses for an example sentence from the Spanish corpus. The number on an edge indicates its posterior probability. Top: Gold parse. Middle: Standard EM posteriors. Bottom: PR sparsity posteriors.

[**JE:** TODO: All supervised experiments need to be re-run; will be slightly higher. Also, all experiments need to be run for the test sets, and statistics for these sets need to be added to the corpus stats table.]

Figure 2 shows an example of a tree where PR significantly outperforms standard EM. As is evidenced in this example, Viterbi parses for standard EM frequently contain the error that determiners are made parents of nouns, instead of the reverse. PR tends not to make this error. One explanation for this improvement is that it is a result of the fact that nouns can sometimes appear without determiners. For example, consider the sentence “Lleva tiempo entenderlos” (translation: “It takes time to understand”) with tags “main-verb common-noun main-verb”. In this situation EM must assign the noun to a parent that is not a determiner. In contrast, when PR sees that sometimes nouns can appear without determiners but that the opposite situation does not occur, it shifts the model parameters to make nouns the parent of determiners instead of the reverse, since then it does not have to pay the cost of assigning a parent with a new tag to cover each noun that doesn’t come with a determiner.

## 5 Related Work

Our learning method (PR) is very closely related to generalized expectation constraints [11, 12], see [1] for details, and is also motivated by a Bayesian view of learning with constraints on posteriors as described in Liang *et al.* [10].

Much work as been dedicated to the task of unsupervised dependency parsing. Departing from the dependency parsing model with valence, several improvements have been porposed: constraining the length of dependencies in a parse tree [14], adding prior over the parameters to encourage sparsity [3, 4, 5], extending the model by better modelling valency [13], and better modelling the categories of the children generated by each parent [8].

## 6 Conclusion

In this paper we presented a method to encourage sparsity in the types of edges learned for dependency grammar induction. The method is encoded as an $\ell_1/\ell_\infty$ penalty using the posterior regularization framework, which does not depend on the particular parametrization of the model. We presented preliminary experiments that show this kind of penalty can result in models that greatly outperform the baseline EM models for two different langauges. These promising results encourage future work in this direction. In particular, we intend to explore defining edge types in terms of parent word and child word rather than tags, to enforce the sparsity constraints at a finer level. Furthermore, we would like to investigate constraints on edge length, either by encouraging locality to prefer short dependency edges, or by introducing some limited language-specific linguistic knowledge.

## References

[1] K. Bellare, G. Druck, and A. McCallum. Alternating projections for learning with expectation constraints. In *In Proc. UAI*, 2009.

[2] D.P. Bertsekas, M.L. Homer, D.A. Logan, and S.D. Patek. Nonlinear programming. *Athena scientific*, 1995.

[3] S.B. Cohen, K. Gimpel, and N.A. Smith. Logistic normal priors for unsupervised probabilistic grammar induction. In *Proc. NIPS*, 2008.

[4] S.B. Cohen and N.A. Smith. The shared logistic normal distribution for grammar induction. In *Proc. NAACL*, 2009.

[5] Shay Cohen and Noah A Smith. Variational inference for grammar induction with prior knowledge. In *Proc. ACL-IJCNLP*, 2009.

[6] J. Graça, K. Ganchev, and B. Taskar. Expectation maximization and posterior constraints. In *Proc. NIPS*, 2008.

[7] J. Graça, K. Ganchev, B. Taskar, and F. Pereira. Posterior sparsity vs parameter sparsity. In *Proc. NIPS*, 2010.

[8] W.P. Headden III, M. Johnson, and D. McClosky. Improving unsupervised dependency parsing with richer contexts and smoothing. In *Proc. NAACL*, 2009.

[9] D. Klein and C. Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proc. ACL*, 2004.

[10] P. Liang, M. I. Jordan, and D. Klein. Learning from measurements in exponential families. In *Proc. ICML*, 2009.

[11] G. Mann and A. McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In *Proc. ICML*, 2007.

[12] G. Mann and A. McCallum. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *In Proc. ACL*, 2008.

[13] D. McClosky. Modeling valence effects in unsupervised grammar induction. Technical report, Technical Report CS-09-01, Brown University, Providence, RI, USA, 2008.

[14] N. Smith and J. Eisner. Annealing structural bias in multilingual weighted grammar induction. In *Proc. ACL*, 2006.

[15] N.A. Smith and J. Eisner. Annealing structural bias in multilingual weighted grammar induction. In *Proc. ACL*, 2006.