Bayesian Modeling of Dependency Trees Using Hierarchical Pitman-Yor Priors

Hanna Wallach WALLACH@CS.UMASS.EDU
Department of Computer Science, University of Massachusetts, Amherst, MA 01003 USA

Charles Sutton SUTTON@CS.BERKELEY.EDU
Computer Science Division, University of California, Berkeley, CA 94720 USA

Andrew McCallum MCCALLUM@CS.UMASS.EDU
Department of Computer Science, University of Massachusetts, Amherst, MA 01003 USA

Recent work in hierarchical priors for language modeling [MacKay and Peto, 1994, Teh, 2006, Goldwater et al., 2006] has shown significant advantages to Bayesian methods in NLP. But the issue of sparse conditioning contexts is ubiquitous in NLP, and these smoothing ideas can be applied more broadly to extend the reach of Bayesian modeling in natural language. For example, a useful representation of higher-level syntactic structure is given by dependency graphs are one such representation of this kind of higher-level structure. Specifically, dependency graphs encode relationships between words and their sentence-level, syntactic modifiers by representing each sentence in a corpus as a directed graph with nodes consisting of the part-of-speech-tagged words in that sentence.

In this paper, we describe two Bayesian models over dependency trees. First, we show that a classic generative dependency model can be substantially improved by (a) using a hierarchical Pitman-Yor process as a prior over the distribution over dependents of a word, and (b) sampling the hyperparameters of the prior. Remarkably, these changes alone yield a significant increase in parse accuracy over the standard model. Second, we present a Bayesian dependency parsing model in which latent state variables mediate the relationships between words and their dependents. The model clusters bilexical dependencies into states using a similar approach to that employed by Bayesian topic models when clustering words into topics. It discovers word clusters with a fine-grained syntactic character.

1. Supervised Bayesian Dependency Parsing

The best-known generative modelling framework for dependency trees is that of Eisner [1996]. This model generates a tagged sentence and its corresponding dependency graph using a parent-outward process. In this model, each parent generates a sequence of children starting in the centre and moving outward to the left and then similarly to the right. Generation of each child is conditioned upon the identity of the tagged parent, the direction of the child in relation to the parent (left or right) and the most recently generated sibling child. That is, conditioned on the parent, the sequence of children in each direction is a first order Markov chain.

The probability of a sentence $\mathbf{w}$ with corresponding part-of-speech tags $\mathbf{s}$,and tree $\mathbf{t}$, generated according to this process, is

$$P(\mathbf{w}, \mathbf{s}, \mathbf{t}) = \prod_n P(w_n, s_n, |w_{\pi(n)}, s_{\pi(n)}, s_{\sigma(n)}, d_n)$$

where d_n is the direction of w_n with respect to its parent, $\pi(n)$ is the position of w_n's parent, and $\sigma(n)$ the position of w_n's immediately preceding sibling (moving outward from w_n's parent in direction d_n).

Estimating each of the parent-child distributions can be difficult because of the sparse conditioning context. Therefore, the estimates are interpolated with estimates of probabilities that depend on a reduced conditioning context, in a manner similar to language modeling.

The interpolation method used by Eisner can be interpreted as a hierarchical Dirichlet Bayesian model, analogously to MacKay and Peto [1994]. A Bayesian reinterpretation has three advantages: firstly, the concentration parameters may be sampled, rather than fixed to some particular value, as is done by Eisner. Secondly, the counts need not correspond to the raw observation counts, as is the case when using the max-

imal restaurant assumption; the minimal restaurant assumption and Gibbs sampling both give rise to other count values.[1] Finally, it is also possible to use priors other than the hierarchical Dirichlet distribution. In this work, we show that a large improvement in performance can be obtained by using a Pitman-Yor prior instead.

We measure the parse accuracy when trained on parse trees generated from the Penn Treebank. Inference consists of two tasks: sampling model hyperparameters given the training data, and inferring trees for unseen test sentences. For the trees, the parents for all words in a sentence can be jointly sampled using an algorithm that combines dynamic programming with the Metropolis-Hastings method. The algorithm is similar to that of Johnson et al. [2007a,b] for unlexicalised probabilistic context-free grammars. The concentration and discount parameters of the Pitman-Yor priors are sampled using slice sampling.

Results are shown in table 1. Using a hierarchical Pitman-Yor prior and sampling hyperparameters both give considerable improvements over a hierarchical Dirichlet model with fixed concentration parameters and the maximal restaurant assumption (equivalent to Eisner's original model). In the hierarchical Dirichlet variant of the model, sampling hyperparameters gives an accuracy improvement of approximately 4%. Using a hierarchical Pitman-Yor prior improves accuracy over the hierarchical Dirichlet variant by approximately 3%. Sampling the hyperparameters of the Pitman-Yor prior gives an accuracy improvement of 5% over the Eisner-equivalent hierarchical Dirichlet model. This corresponds to a 26% reduction in error.

2. "Syntactic Topic" Dependency Models

One advantage of a generative approach to modelling dependency trees is that other latent variables may be incorporated into the model. To demonstrate this, we also present a dependency parsing model with latent variables that mediate the relationships between words and their modifiers, resulting in a clustering of bilexical dependencies.

This model can be considered to be a dependency-based analogue of the syntactic component from the syntax-based topic model of Griffiths et al. [2005]. The models differ in their underlying structure, however: In the model presented in this section, the underlying structure is a tree that combines both words and unobserved syntactic states; in Griffiths et al.'s model, the structure is a simply a chain over latent states. This difference means that there are two kinds of latent information that must be inferred in the dependency-based model: The structure of each dependency tree and the identities of the latent states. In Griffiths et al.'s model, only the latter need be inferred.

The generative process underlying the model in this section is similar to that of the model presented in the previous section, with the key difference that latent state variables $\mathbf{s}$ mediate the relationships between parents and children. The probability of an untagged sentence $\mathbf{w}$ with latent states $\mathbf{s}$ and tree $\mathbf{t}$ is therefore given by

$$P(\mathbf{w}, \mathbf{s}, \mathbf{t}) = \prod_n \theta_{s_n|w_{\pi(n)}} \, \phi_{w_n|s_n}$$

where the vector $\boldsymbol{\theta}_{w'}$ is the distribution over latent states for parent word w' and the vector $\boldsymbol{\phi}_s$ is the distribution over child words for latent state s. (Note that sibling words are ignored in this model, making it a first order dependency model.) In other words, parent words are collapsed down to the latent state space and children are generated on the basis of these states. As a result, the clusters induced by the latent states are expected to exhibit syntactic properties and can be thought of as "syntactic topics"—specialised distributions over words with a syntactic flavour.

Penn Treebank sections 2–21 were used as training data. The true dependency trees and words were used to obtain a single sample of states. These states, trees and words were then used to sample states and trees for Penn Treebank section 23.

Some example states or "syntactic topics" are shown in table 3. Each column in each row consists of the ten words most likely to be generated by a particular state. The states exhibit a good correspondence with parts-of-speech, but are more finely grained. For example, the states in the first and third columns in the top row both correspond to nouns. However, the first contains job titles, while the third contains place names. Similarly, the states in the fourth and fifth columns in the top row both correspond to verbs. However, the fourth contains transitive past-tense verbs, while the fifth contains present-tense verbs. The state shown in the final column in the bottom row is particularly interesting because the top words are entirely plural nouns. This kind of specificity indicates that these states are likely to be beneficial in other tasks where part-of-speech tags are typically used, such as named entity recognition.

[1] In the chinese restaurant metaphor, the restaurant assumptions deal with the issue of how many tables at each level serve the same dish. For a discussion of this issue, see ????.

A more detailed description of this work is available online at `http://www.cs.umass.edu/~wallach/parsing.pdf`.

References

Jason Eisner. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 340–345, Copenhagen, August 1996.

Sharon Goldwater, Tom Griffiths, and Mark Johnson. Interpolating between types and tokens by estimating power-law generators. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 459–466. MIT Press, Cambridge, MA, 2006.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press, Cambridge, MA, 2005.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. Bayesian inference for pcfgs via Markov chain monte carlo. In *HLT/NAACL*, 2007a.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press, Cambridge, MA, 2007b.

D. J. C. MacKay and L. Peto. A hierarchical Dirichlet language model. *Natural Language Engineering*, 1 (3):1–19, 1994. URL `http://citeseer.nj.nec.com/mackay94hierarchical.html`.

Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, 2006. URL `http://www.aclweb.org/anthology/P/P06/P06-1124`.

		Restaurant assumption	
		Maximal	Minimal
Dirichlet	fixed α values [Eisner, 1996]	80.7	80.2
Dirichlet	sampled α values	84.3	84.1
Pitman-Yor	fixed α and d values	83.6	83.7
Pitman-Yor	sampled α and d values	85.4	**85.7**

Table 1. Parse accuracy of the hierarchical Pitman-Yor dependency model on Penn Treebank data. Results are computed using the maximum probability tree.

Model	Accuracy (sampled trees)	Accuracy (most probable tree)
50 states	59.2	63.8
100 states	60.0	64.1
150 states	60.5	64.7
200 states	60.4	64.5
POS tags	55.3	63.1

Table 2. Parse accuracy of the "syntactic topic" dependency model on the Penn Treebank (standard train/test split). As a baseline, the latent states are fixed to part-of-speech tags. Results for sampled trees are averaged over 10 samples.

president	year	u.s.	made	is	in
director	years	california	offered	are	on
officer	months	washington	filed	was	,
chairman	quarter	texas	put	has	for
executive	example	york	asked	have	at
head	days	london	approved	were	with
attorney	time	japan	announced	will	and
manager	weeks	canada	left	had	as
chief	period	france	held	's	by
secretary	week	britain	bought	would	up
10	would	more	his	ms.	sales
8	will	most	their	mrs.	issues
1	could	very	's	who	prices
50	should	so	her	van	earnings
2	can	too	and	mary	results
15	might	than	my	lee	stocks
20	had	less	your	dorrance	rates
30	may	and	own	linda	costs
25	must	enough	'	carol	terms
3	owns	about	old	hart	figures

Table 3. The top ten words most likely to be generated as children by twelve of the states inferred from the true dependency trees for Penn Treebank sections 2–21. These examples are taken from a model with 150 states.