A Provably Correct Learning Algorithm for Latent-Variable PCFGs

Shay Cohen[1] Michael Collins[2]

[1]University of Edinburgh

[2]Columbia University

June 24, 2014

Previous work: spectral algorithm for L-PCFGs

Introduced in Cohen et al. (2012), based on learning for HMMs (Hsu et al., 2009)

An algorithm for latent-variable PCFGs

Unlike EM, no local maxima problem

More efficient than EM

Experimentally, works as well as EM for parsing

Problem with previous work (Cohen et al., 2012)

Parameters are masked by an unknown linear transformation

- Negative probabilities (due to sampling error)
- Parameters cannot be easily interpreted
- Cannot improve parameters using, for example, EM

This talk in a nutshell

Like the spectral algorithm, has theoretical guarantees

Estimates are actual probabilities

More efficient than EM

Can be used to initialize EM, which converges in an iteration or two

Relies heavily on the idea of "pivot features"

- Features that uniquely identify a latent state
- A similar idea is used for topic modeling by Arora et al. (2013)

Outline of this talk

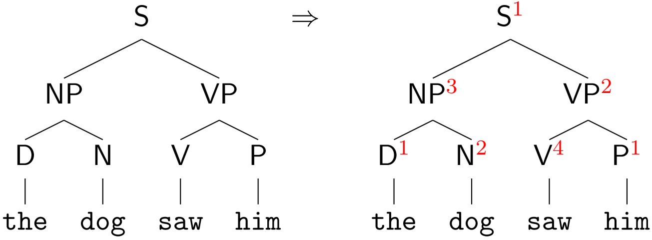
Latent-variable PCFGs (Matsuzaki et al., 2005; Petrov et al., 2006)

Algorithm for L-PCFG estimation

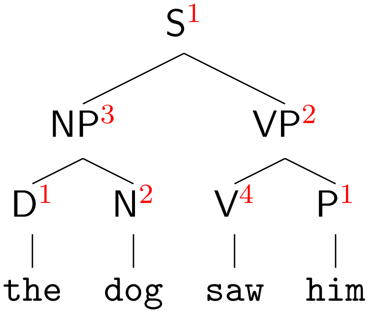
Experiments

Conclusion

L-PCFGs (Matsuzaki et al., 2005; Petrov et al., 2006)



The probability of a tree



$$
\begin{aligned}
& p(\mathsf{tree}, 1\ 3\ 1\ 2\ 2\ 4\ 1) \\
&= \pi(\mathsf{S}^1)\times \\
&\quad t(\mathsf{S}^1 \rightarrow \mathsf{NP}^3\ \ \mathsf{VP}^2|\mathsf{S}^1)\times \\
&\quad t(\mathsf{NP}^3 \rightarrow \mathsf{D}^1\ \ \mathsf{N}^2|\mathsf{NP}^3)\times \\
&\quad t(\mathsf{VP}^2 \rightarrow \mathsf{V}^4\ \ \mathsf{P}^1|\mathsf{VP}^2)\times \\
&\quad q(\mathsf{D}^1 \rightarrow \texttt{the}|\mathsf{D}^1)\times \\
&\quad q(\mathsf{N}^2 \rightarrow \texttt{dog}|\mathsf{N}^2)\times \\
&\quad q(\mathsf{V}^4 \rightarrow \texttt{saw}|\mathsf{V}^4)\times \\
&\quad q(\mathsf{P}^1 \rightarrow \texttt{him}|\mathsf{P}^1)
\end{aligned}
$$

$$
p(\mathsf{tree}) = \sum_{h_1 \ldots h_7} p(tree, h_1\ h_2\ h_3\ h_4\ h_5\ h_6\ h_7)
$$

Outline of this talk

Latent-variable PCFGs (Matsuzaki et al., 2005; Petrov et al., 2006)

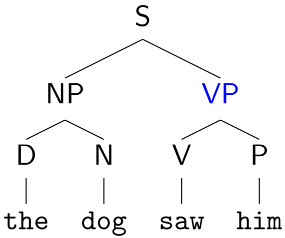
Algorithm for L-PCFG estimation

Experiments

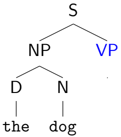
Conclusion

Inside and outside trees

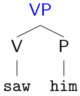
At node VP:



Outside tree $o =$



Inside tree $t =$



Conditionally independent given the label and the hidden state

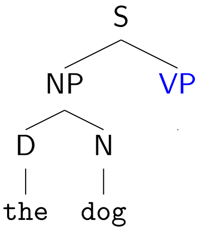
$$p(o, t|\mathsf{VP}, h) = p(o|\mathsf{VP}, h) \times p(t|\mathsf{VP}, h)$$

Designing feature functions

Design functions ψ and ϕ:

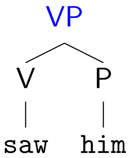
ϕ maps any inside tree to a binary vector of length d

ψ maps any outside tree to a binary vector of length d'



Outside tree $o \Rightarrow$

$\psi(o) = [0, 1, 0, 0, \ldots, 0, 0] \in \mathbb{R}^{d'}$



Inside tree $t \Rightarrow$

$\phi(t) = [1, 0, 0, 0, \ldots, 0, 0] \in \mathbb{R}^{d}$

ψ and ϕ as multinomials $p(f)$ for $f \in [d]$ and $p(g)$ for $g \in [d']$.

Latent state distributions

Think of f and g as representing a whole inside/outside tree

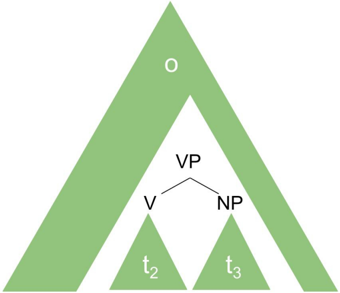
Say we had a way of getting:

- $p(f|h, \text{VP})$ for each h and f inside feature
- $p(g|h, \text{VP})$ for each h and g outside feature

Then we could run EM on a convex problem to find parameters.
How?

Binary rule estimation

Take M samples of nodes with rule VP $\rightarrow$ V NP.



At sample i

- $g^{(i)}$ = outside feature at VP
- $f_2^{(i)}$ = inside feature at V
- $f_3^{(i)}$ = inside feature at NP

$$\{\hat{t}(h_1, h_2, h_3 | \text{VP} \rightarrow \text{V NP}) | h_1, h_2, h_3\}$$
$$= \arg\max_{\hat{t}} \sum_{i=1}^{M} \log \sum_{h_1, h_2, h_3} \Big(\hat{t}(h_1, h_2, h_3 | \text{VP} \rightarrow \text{V NP}) \times$$
$$p(g^{(i)} | h_1, \text{VP}) p(f_2^{(i)} | h_2, \text{V}) p(f_3^{(i)} | h_3, \text{NP})\Big)$$

Binary rule estimation, cont'd

$$\{\hat{t}(h_1, h_2, h_3 | \mathsf{VP} \rightarrow \mathsf{V} \ \ \mathsf{NP}) | h_1, h_2, h_3\}$$
$$= \arg\max_{\hat{t}} \sum_{i=1}^{M} \log \sum_{h_1, h_2, h_3} \Big(\hat{t}(h_1, h_2, h_3 | \mathsf{VP} \rightarrow \mathsf{V} \ \ \mathsf{NP}) \times$$
$$p(g^{(i)} | h_1, \mathsf{VP}) p(f_2^{(i)} | h_2, \mathsf{V}) p(f_3^{(i)} | h_3, \mathsf{NP}) \Big)$$

This objective represents the marginal probability of the corpus

It is a convex objective

Use Bayes' rule to convert to parameters

Main question: how do we get the latent state distributions $p(h|f, \textbf{VP})$ and $p(h|g, \textbf{VP})$?

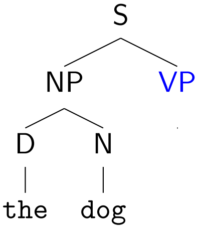
Vector representation of inside and outside trees

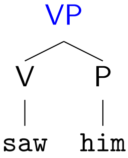
Design functions Z and Y:

Y maps any inside feature value $f \in [d']$ to a vector of length m.

Z maps any outside feature value $g \in [d]$ to a vector of length m.

Convention: m is the number of hidden states under the L-PCFG.





Outside tree $o \Rightarrow$

$Z(g) = [1, 0.4, -5.3, \ldots, 72] \in \mathbb{R}^m$

Inside tree $t \Rightarrow$

$Y(f) = [-3, 17, 2, \ldots, 3.5] \in \mathbb{R}^m$

Z and Y reduce the dimensionality of ϕ and ψ using CCA

Identifying latent state distributions

- For each $f \in [d]$, define: $v(f) = E[Z(g)|f, \text{VP}]$

- $v(f) \in \mathbb{R}^m$ is **"the expected value of an outside tree (representation) given an inside tree (feature)"**

Identifying latent state distributions

- For each $f \in [d]$, define: $v(f) = E[Z(g)|f, \text{VP}]$

- $v(f) \in \mathbb{R}^m$ is **"the expected value of an outside tree (representation) given an inside tree (feature)"**

- By conditional independence:

$$v(f) = \sum_{h=1}^{m} p(h|f, \text{VP}) w(h)$$

where $w(h) \in \mathbb{R}^m$ and
$w(h) = \sum_{g=1}^{d'} p(g|h, \text{VP}) Z(g) = E[Z(g)|h, \text{VP}]$.

- $w(h)$ is **"the expected value of an outside tree (representation) given a latent state"**

Pivot assumption

Reminder: $v(f) = \sum_{h=1}^{m} p(h|f, \mathsf{VP}) w(h) = E[Z(g)|f, \mathsf{VP}]$

Need to solve with respect to $p(h|f, \mathsf{VP})$

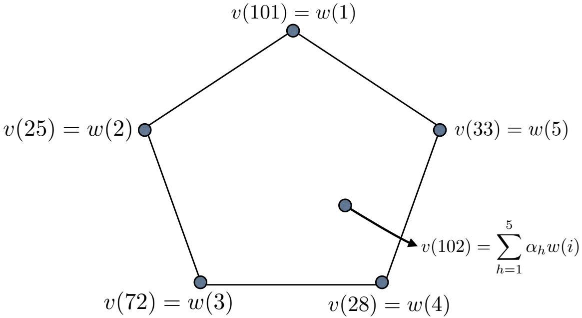
$v(f)$ can be estimated from data

$w(h)$ consist of information about latent states – not observable

Pivot assumption: each h has f such that $p(h|f, \mathsf{VP}) = 1$

Outside representation as convex combination

Reminder: $v(f) = \sum_{h=1}^{m} p(h|f, \text{VP}) w(h) = E[Z(g)|f, \text{VP}]$



Pivot assumption: each h has f such that $p(h|f, \text{VP}) = 1$

Features 101, 25, 72, 28, 33 are "pivot" features for 5 states

α are the latent state distributions!

Summary of algorithm

Calculate $v(f)$ for all f

Identify $w(h)$ for all h by finding the corners of $\mathrm{ConvexHull}(v(1), \ldots, v(d))$

Identify the distribution of latent states by solving

$$v(f) = \sum_{h=1}^{m} p(h|f, \mathsf{VP}) w(h) = E[Z(g)|f, \mathsf{VP}]$$

Repeat the above for outside features g

Solve a convex marginal log-likelihood problem

Outline of this talk

Latent-variable PCFGs (Matsuzaki et al., 2005; Petrov et al., 2006)

Algorithm for L-PCFG estimation

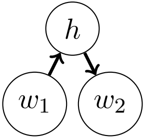
Experiments

Conclusion

Experiments: language modeling

- Saul and Pereira (1997):

$$p(w_2|w_1) = \sum_{h} p(w_2|h)p(h|w_1).$$



This model is a specific case of L-PCFG

- Experimented with bi-gram modeling for two corpora: Brown corpus and Gigaword corpus

Results: perplexity

	Brown			NYT		
m	*128*	*256*	test	*128*	*256*	test
bigram Kneser-Ney	408		415	271		279
trigram Kneser-Ney	386		394	150		158
EM	388	365	364	284	265	267
iterations	9	8		35	32	
pivot	426	597	560	782	886	715

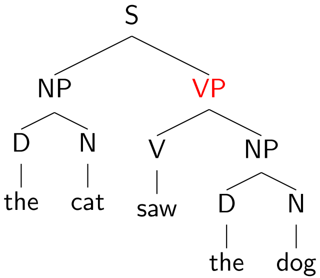
Results: perplexity

	Brown			NYT		
m	*128*	*256*	test	*128*	*256*	test
bigram Kneser-Ney	408		415	271		279
trigram Kneser-Ney	386		394	150		158
EM	388	365	364	284	265	267
iterations	9	8		35	32	
pivot	426	597	560	782	886	715
pivot+EM	**310**	**327**	357	**279**	292	281
iterations	1	1		19	12	

- Initialize EM with pivot algorithm output
- EM converges in much fewer iterations
- Still consistent - called "two-step estimation" (Lehmann and Casella, 1998)

Inside features used

Consider the VP node in the following tree:

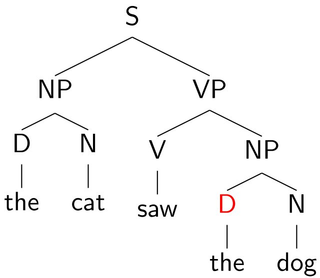


The inside features consist of:

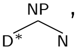
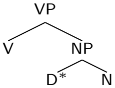
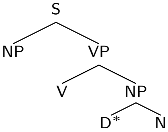
- The pairs `(VP, V)` and `(VP, NP)`
- The rule `VP` $\rightarrow$ `V NP`
- The tree fragment `(VP (V saw) NP)`
- The tree fragment `(VP V (NP D N))`
- The pair of head part-of-speech tag with VP: `(VP, V)`

Outside features used

Consider the D node in the following tree:



The outside features consist of:

- The fragments  ,  and 
- The pair (D, NP) and triplet (D, NP, VP)
- The pair of head part-of-speech tag with D: (D, N)

Results

	sec. 22				sec. 23
m	8	16	24	32	
EM iterations	86.69 40	88.32 30	88.35 30	88.56 20	87.76
Spectral (Cohen et al., 2013)	85.60	87.77	88.53	88.82	88.05
Pivot	83.56	86.00	86.87	86.40	85.83
Pivot+EM iterations	86.83 2	88.14 6	88.64 2	88.55 2	88.03

Again, EM converges in very few iterations

Conclusion

Formal guarantees:

- Statistical consistency
- No problem of local maxima

Advantages over traditional spectral methods:

- No negative probabilities
- More intuitive to understand