A Bayesian Model of Syntax-Directed Tree to String Grammar Induction
Trevor Cohn and Phil Blunsom

Presented by: Jeff Flanigan
The Problem

- Lots of heuristics for grammar extraction
- Word alignments could be better
- EM for extracting rules fails
  - See “Why Generative Phrase Models Underperform Surface Heuristics” DeNero et al. 2006 (VERY good read!)
The Problem

• Lots of heuristics for grammar extraction
• Word alignments could be better
• EM for extracting rules fails
  – See “Why Generative Phrase Models Underperform Surface Heuristics” DeNero et al. 2006 (VERY good read!)

Learn rules directly

Use Bayesian Methods
The Idea

Learn alignments between target trees and source spans directly

Each node in tree has a latent variable (the alignment)
Nodes can be unaligned
The Idea

One rule extracted for each aligned node
Alignment span ⇔ Rule

Use Bayesian Nonparametrics to Prevent Degeneracy
Bayesian Learning

\[ p(\theta \mid D, \alpha) = \frac{p(D \mid \theta)p(\theta \mid \alpha)}{p(D \mid \alpha)} \propto p(D \mid \theta)p(\theta \mid \alpha) \]

In Bayesian grammar induction, \( \theta \) is distribution over grammars (usually write \( G \) instead of \( \theta \))
Learning a grammar is a “draw” from \( G \)
Bayesian Learning

\[ p(\theta | D, \alpha) = \frac{p(D | \theta)p(\theta | \alpha)}{p(D | \alpha)} \propto p(D | \theta)p(\theta | \alpha) \]

In Bayesian grammar induction, \( \theta \) is distribution over grammars (usually write \( G \) instead of \( \theta \))
Learning a grammar is a “draw” from \( G \)

Use Gibbs Sampling to get a draw (i.e. grammar)
Gibbs Sampling

- The GS algorithm:
  1. Suppose the graphical model contains variables $x_1, \ldots, x_n$
  2. Initialize starting values for $x_1, \ldots, x_n$
  3. Do until convergence:
     1. Pick an ordering of the $n$ variables (can be fixed or random)
     2. For each variable $x_i$ in order:
        1. Sample $x$ from $P(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, i.e. the conditional distribution of $x_i$ given the current values of all other variables
        2. Update $x_i \leftarrow x$

- When we update $x_i$, we **immediately** use its new value for sampling other variables $x_j$
Bayesian Learning

Use Gibbs Sampling

\[ p(\theta \mid D, \alpha) = \frac{p(D \mid \theta)p(\theta \mid \alpha)}{p(D \mid \alpha)} \propto p(D \mid \theta)p(\theta \mid \alpha) \]

Need: prior over the space of grammars
Prior over Grammars

Consitituent $c$
Rewrite $c$ using rule $r$ with probability:

$$r|c \sim G_c$$

$G_c$ is a distribution over grammars, drawn from a Dirichlet Process:

$$G_c|\alpha_c, P_0 \sim \text{DP}(\alpha_c, P_0(\cdot|c))$$
Dirichlet Process

Draws from $\text{DP}(\alpha, H(\Omega))$ are distributions (i.e. $\geq 0$, sum to 1)
$H$ is the base distribution over $\Omega$
$\alpha$ is the concentration parameter – determines sparsity

Base distribution
A single draw from the DP

In our case $\Omega = \text{space of rules}$

Picture source: http://www.cs.cmu.edu/~epxing/Class/10708/lecture/lecture24-DP.pdf
Base Distribution $H(\Omega)$

Base distribution probability of rewriting $c$ with RHS $r$

$$P_0(r | c) = P_0(e, w | c) = P(e | c) \cdot P(w | e)$$

- $e = \text{elementry tree}$
- $w = \text{words in source}$

$P(e | c)$
- Expand $c$ recursively
- Number of child nodes $\sim \text{Geom}(p_{\text{child}})$
- Pre-terminals have one child
- Draw non-terminals and terminals uniformly from $N$ and $T$

$P(w | e)$
- Number of terminals $\sim \text{Geom}(p_{\text{term}})$
- Draw source terminals (i.e. source phrases) uniformly from possible phrases
- Arrange variables, source phrases randomly

Example:

$$\langle (\text{NP NP}_1 \ (\text{PP (IN of) NP}_2)), \ 2 \ \text{的} \ 1 \rangle$$
The Gibbs Sampler

Visit sentences/nodes randomly, and resample a rule.

Resample using $P(r | \text{everything else})$
The Gibbs Sampler

Dirichlet Process Prior

\[ r | c \sim G_c \]
\[ G_c | \alpha_c, P_0 \sim \text{DP}(\alpha_c, P_0(\cdot | c)) \]
The Gibbs Sampler

Dirichlet Process Prior

\[ r|c \sim G_c \]
\[ G_c|\alpha_c, P_0 \sim \text{DP}(\alpha_c, P_0(\cdot|c)) \]

Integrate out \( G_c \) (see Neal 2000 “Markov Chain Sampling Methods for Dirichlet Process Mixture Models”)


The Gibbs Sampler

Dirichlet Process Prior

\[ r|c \sim G_c \]
\[ G_c|\alpha_c, P_0 \sim \text{DP}(\alpha_c, P_0(\cdot|c)) \]

Integrate out \( G_c \) (see Neal 2000 “Markov Chain Sampling Methods for Dirichlet Process Mixture Models”)

Gibbs Sampler is simple!

\[
p(r_i|r^{-i}, c, \alpha_c, P_0) = \frac{n_{r_i}^{-i} + \alpha_c P_0(r_i|c)}{n_c^{-i} + \alpha_c}
\]

- \( n_{r_i}^{-i} \) = # times rule \( r_i \) used everywhere else
- \( n_c^{-i} \) = # times \( c \) used everywhere else
The Gibbs Sampler

Dirichlet Process Prior

\[ r|c \sim G_c \]
\[ G_c|\alpha_c, P_0 \sim DP(\alpha_c, P_0(·|c)) \]

Integrate out \( G_c \) (see Neal 2000 “Markov Chain Sampling Methods for Dirichlet Process Mixture Models”)

Gibbs Sampler is simple!

\[ p(r_i|r^{-i}, c, \alpha_c, P_0) = \frac{n_{r_i}^{-i} + \alpha_c P_0(r_i|c)}{n_{c}^{-i} + \alpha_c} \]

- \( n_{r_i}^{-i} \): \# times rule \( r_i \) used everywhere else
- \( n_{c}^{-i} \): \# times \( c \) used everywhere else

Small \( \alpha \) means do what’s popular

“Rich get richer”
Resampling Operators

- 1st EXPAND
  - Resample the alignment subject to constrains
  - Can't go outside closest aligned parent
  - Must include descendants
  - Can't overlap siblings
Resampling Operators

• 2\textsuperscript{nd} SWAP
  – Swap alignments of two nodes
  – Only allowed for nodes w/ unaligned descendants
Summary

$$p(\theta | D, \alpha) = \frac{p(D | \theta) p(\theta | \alpha)}{p(D | \alpha)} \propto p(D | \theta) p(\theta | \alpha)$$
Training Set

• FBIS + 100k Sinorama

|                      | English ← Chinese |
|----------------------|------------------|
| Sentences            | 300k             |
| Words or Segments    | 11.0M            |
| Avg. Sent. Length    | 36               |
| Longest Sent.        | 80               |
|                      | 8.6M             |
|                      | 28               |
|                      | 80               |
Training

- \( \alpha = 10^6 \)
- \( p_{\text{child}} = p_{\text{expand}} = p_{\text{term}} = .5 \)
- 300 iterations of Gibbs sampler
- \( \frac{1}{2} \) hour per iteration on single core 2.3 Ghz
Extracted Grammar

- Number of rules for maximum tree depth
- Number of rules for source terminals
- Number of rules for target terminals
Extracted Grammar

Top 10 Rules not in GHKM

\[
\begin{align*}
&\langle (\text{TOP} \ (\text{S} \ \text{NP}1 \ \text{VP}2 \ .3)), \ 1 \ 2 \ 3 \rangle \\
&\langle (\text{S} \ (\text{VP} \ (\text{TO} \ \text{to}) \ \text{VP}1)), \ 1 \rangle \\
&\langle (\text{NP} \ \text{NP}1 \ (\text{PP} \ (\text{IN} \ \text{of}) \ \text{NP}2)), \ 2 \ 1 \rangle \\
&\langle (\text{PP} \ (\text{IN} \ \text{in}) \ \text{NP}1), \ \text{在} \ 1 \rangle \\
&\langle (\text{NP} \ \text{NP}1 \ (\text{PP} \ (\text{IN} \ \text{of}) \ \text{NP}2)), \ 1 \ 2 \rangle \\
&\langle (\text{NP} \ (\text{DT} \ \text{the}) \ \text{NN}1), \ \text{的} \ 1 \rangle \\
&\langle (\text{S} \ (\text{VP} \ \text{TO}1 \ \text{VP}2)), \ 1 \ 2 \rangle \\
&\langle (\text{VP} \ (\text{VBZ} \ \text{is}) \ \text{NP}1), \ \text{是} \ 1 \rangle \\
&\langle (\text{NP} \ (\text{NP} \ (\text{DT} \ \text{the}) \ \text{NN}1) \ (\text{PP} \ (\text{IN} \ \text{of}) \ \text{NP}2)), \ 2 \ 1 \rangle \\
\end{align*}
\]

Top 10 Rules not in New Model

\[
\begin{align*}
&\langle (\text{PP} \ (\text{IN} \ \text{at}) \ (\text{NP} \ \text{DT}1 \ (\text{NNS \ levels}))), \ \text{1 级} \rangle \\
&\langle (\text{NP} \ \text{NP}1 \ 2 \ \text{NP}3 \ (\), \ \text{CC} \ 4 \ \text{NP}5), \ 1 \ 2 \ 3 \ 4 \ 5 \rangle \\
&\langle (\text{NP} \ \text{NP}1 \ 2 \ \text{NP}3 \ 4 \ \text{NP}5 \ (\), \ (\text{CC} \ \text{and}) \ \text{NP}6), \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \rangle \\
&\langle (\text{S} \ \text{S}1 \ (\text{NP} \ (\text{PRP They})) \ \text{VP}2 \ .3), \ 1 \ 2 \ 3 \rangle \\
&\langle (\text{S} \ \text{PP}1 \ 2 \ \text{NP}3 \ \text{VP}4 \ .5 \ 6), \ 1 \ 2 \ 3 \ 4 \ 6 \ 5 \rangle \\
&\langle (\text{S} \ \text{PP}1 \ 2 \ \text{NP}3 \ \text{VP}4 \ .5), \ 1 \ 2 \ 3 \ 4 \ 5 \rangle \\
&\langle (\text{NP} \ (\text{NNP Foreign}) \ (\text{NNP Ministry}) \ \text{NN1} \ (\text{NNP Zhu}) \ (\text{NNP Bangzao}), \ \text{外交部 1 朱邦造} \rangle \\
&\langle (\text{S} \ \text{S}1 \ \text{S}2), \ 1 \ 2 \rangle \\
&\langle (\text{S} \ \text{S}1 \ (\text{NP} \ (\text{PRP We})) \ \text{VP}2 \ .3), \ 1 \ 2 \ 3 \rangle \\
&\langle (\text{NP} \ (\text{DT} \ \text{the}) \ (\text{NNS people}) \ \text{POS1}), \ \text{人民} \ 1 \rangle \\
\end{align*}
\]

Sampled Grammar

\[
\begin{align*}
&\langle (\text{NP} \ (\text{DT} \ \text{the}) \ \text{NN1}), \ \text{的} \ 1 \rangle \\
&\langle (\text{NP} \ (\text{NP} \ (\text{DT} \ \text{the}) \ \text{NN1}) \ (\text{PP} \ (\text{IN} \ \text{of}) \ \text{NP2})), \ 2 \ \text{的} \ 1 \rangle \\
&\langle (\text{NP} \ (\text{DT} \ \text{the}) \ \text{NN1}), \ 1 \ \text{的} \rangle \\
&\langle (\text{NP} \ (\text{NP} \ (\text{DT} \ \text{the}) \ \text{NN2}) \ (\text{PP} \ (\text{IN} \ \text{of}) \ \text{NP3})), \ 3 \ \text{的} \ 1 \ 2 \rangle \\
&\langle (\text{PP} \ (\text{IN} \ \text{of}) \ \text{NP1}), \ 1 \ \text{的} \rangle \\
\end{align*}
\]

GHKM Grammar

\[
\begin{align*}
&\langle (\text{NP} \ \text{JJ1} \ \text{NNS}2), \ 1 \ \text{的} \ 2 \rangle \\
&\langle (\text{NP} \ \text{JJ1} \ \text{NN2}), \ 1 \ \text{的} \ 2 \rangle \\
&\langle (\text{NP} \ \text{DT}1 \ \text{JJ2} \ \text{NN3}), \ 1 \ 2 \ \text{的} \ 3 \rangle \\
&\langle (\text{NP} \ \text{PRP$1$} \ \text{NN2}), \ 1 \ \text{的} \ 2 \rangle \\
&\langle (\text{NP} \ \text{NP1} \ \text{PP2}), \ 2 \ \text{的} \ 1 \rangle \\
\end{align*}
\]

Table 5: Top five rules which include the possessive particle and at least one variable.
**BLEU Scores**

| Model      | BLEU score |
|------------|------------|
| GHKM       | 26.0       |
| Our model  | 26.6       |

Table 6: Translation results on the NIST test set MT03 for sentences of length ≤ 20.
\[ p(\theta \mid D, \alpha) = \frac{p(D \mid \theta)p(\theta \mid \alpha)}{p(D \mid \alpha)} \propto p(D \mid \theta)p(\theta \mid \alpha) \]