A Bayesian Model of Syntax-Directed Tree to String Grammar Induction

Trevor Cohn and Phil Blunsom

Presented by Kevin Gimpel
4/7/2010

Overview

- Problem: rule extraction for syntax-based SMT systems
  - Usually done by word alignment followed by heuristics
  - In some early work, rule weights were trained via EM, but this is also problematic

- Solution: Bayesian model with nonparametric priors on rule distributions
  - Avoids separate word alignment step
  - Nonparametric priors allow sets of rules to be unbounded
  - Dirichlet process (DP) priors favor power law effects among rules, avoiding degenerate solutions typically found by EM

- Continuing a line of research into Bayesian models for phrase/rule extraction in MT and parsing
  - DeNero et al. (2008), Blunsom et al. (2009), Cohn et al. (2009), etc.

Formalism

- Synchronous Tree Substitution Grammar (STSG)
- Generalization of SCFG in which RHS of rules can contain trees
- Example rule:

$$\langle (\text{NP NP}_{\boxed{1}} \text{ (PP (IN of) NP}_{\boxed{2}})), \boxed{2} \text{ 的 } \boxed{1} \rangle$$

- They use a standard model parameterization: collection of conditional distributions, one for each LHS nonterminal

Model

- They use a Dirichlet process prior for each of these distributions:

RHS of rule

LHS of rule

Base distribution for c

$$
\begin{aligned}
r|c &\sim G_c \\
G_c|\alpha_c, P_0 &\sim \mathrm{DP}(\alpha_c, P_0(\cdot|c))
\end{aligned}
$$

  - The set of rules for each nonterminal c is unbounded
  - They use the standard approach of integrating out G_c during inference via collapsed Gibbs sampling
- The base distribution factors the generation of the RHS of a rule into a simple generative story

Base Distribution

- Simple generative process:
  - Generates each nonterminal and terminal in the target tree, then each terminal and variable placement in the source string
  - Favors small rules

rule (elementary tree, source string)

$$r = (\mathbf{e}, \mathbf{w})$$

$$P_0(\mathbf{e}, \mathbf{w}|c) = P(\mathbf{e}|c)P(\mathbf{w}|\mathbf{e})$$

Root of rule

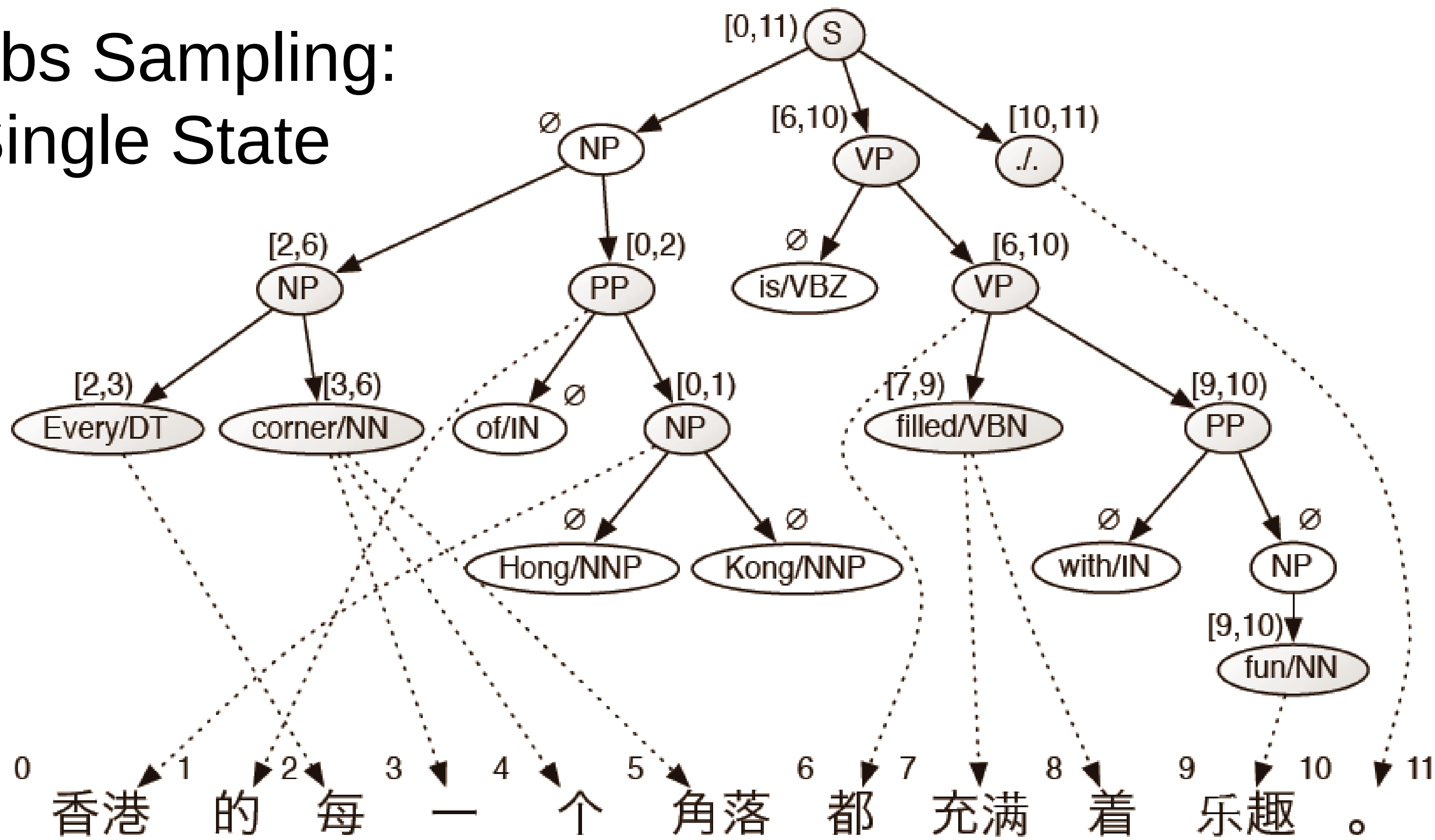
Aside: Modeling Extensions

- Their model captures power law effects among rules within each distribution
- But these distributions are independent
  - The rules for a VP have no effect on the rules for an S
- Possible extension: hierarchical Dirichlet process (HDP)
  - Shares power law effects among different distributions (e.g., among the VP rule distribution and the S rule distribution)
  - Has been used frequently when models contain a large number of conditional distributions that should share characteristics (e.g., n-gram language models)
  - Could have a separate HDP for each “family” of rule distributions

Inference

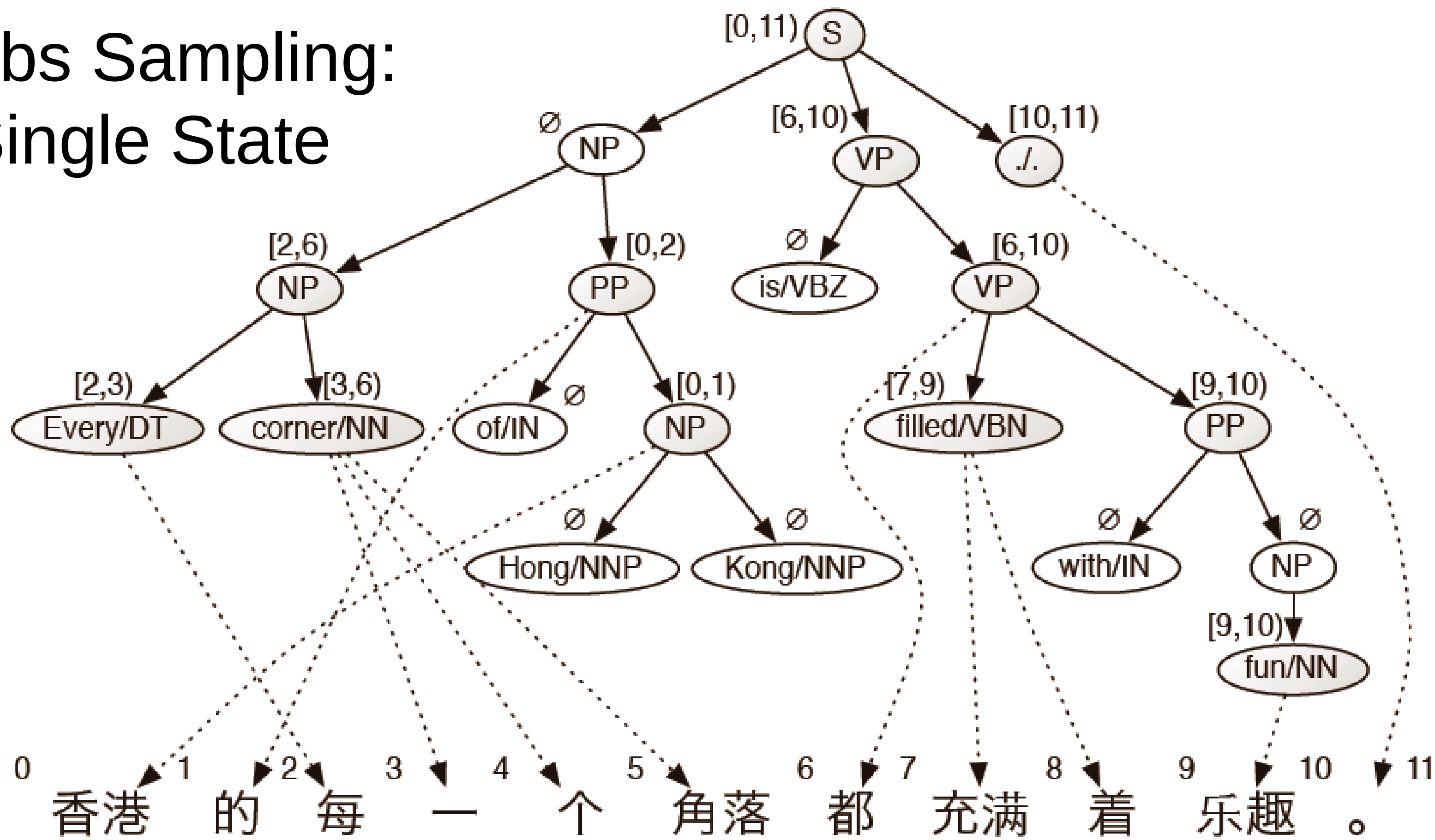
- They want to avoid doing word alignment as a preprocessing step followed by heuristic rule extraction
- Instead, they use Gibbs sampling to sample from the posterior distribution over grammars
- They extract rules from a single final sample

Gibbs Sampling: Single State



Shaded nodes are roots of rules that get extracted

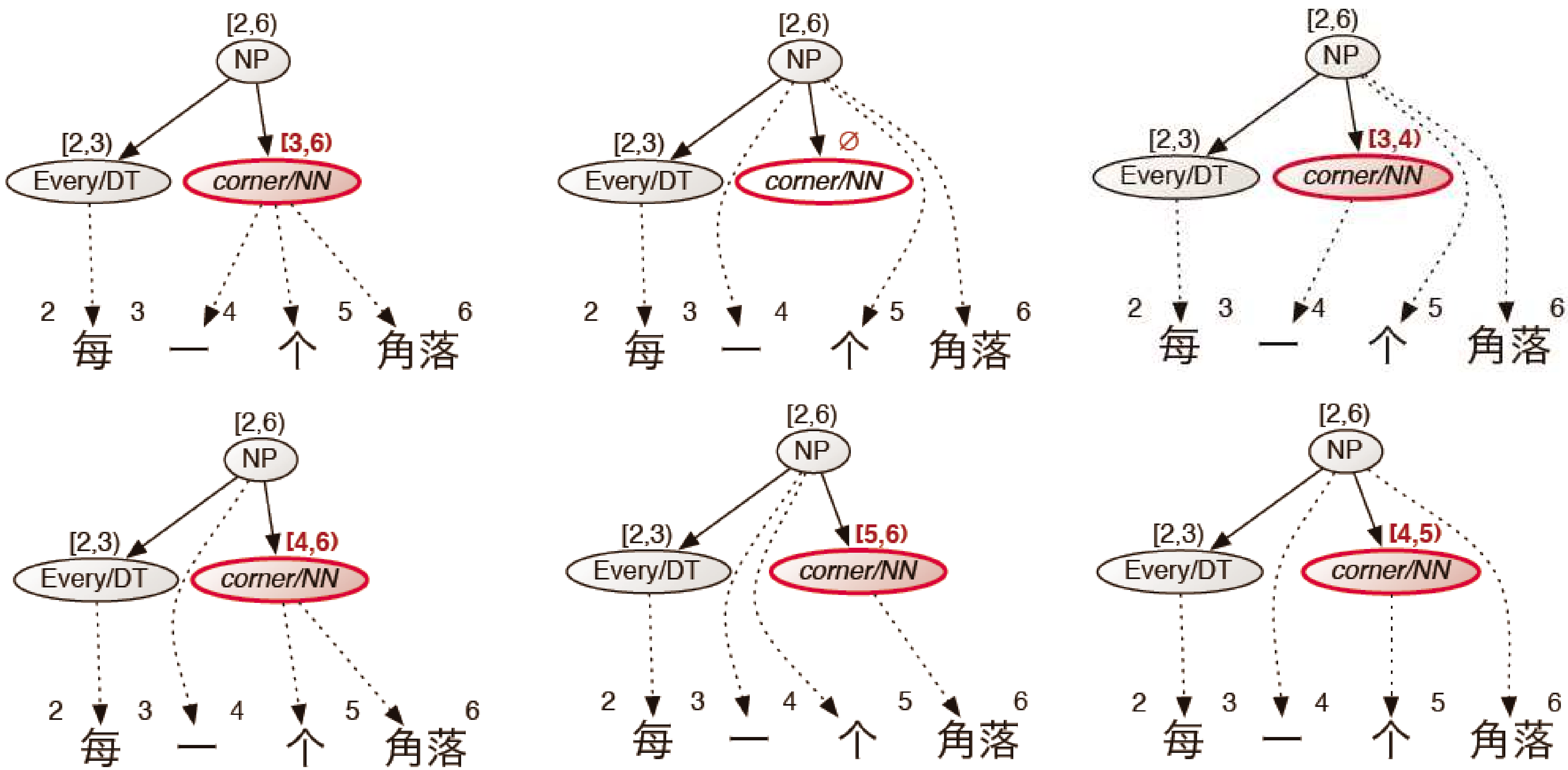
Gibbs Sampling: Single State



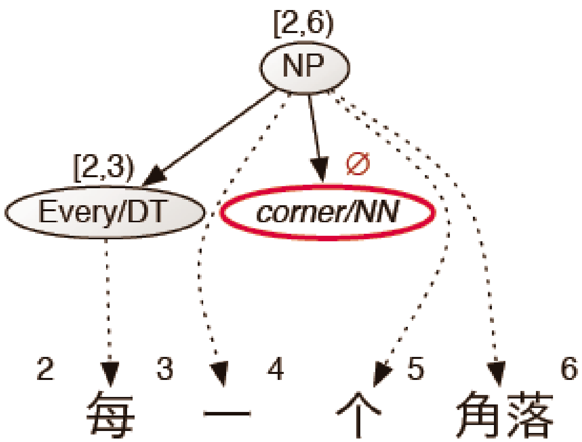
Rules extracted:

⟨(S (NP NP$_{\boxed{1}}$ PP$_{\boxed{2}}$) VP$_{\boxed{3}}$.$_{\boxed{4}}$), $\boxed{2}$ $\boxed{1}$ $\boxed{3}$ $\boxed{4}$⟩
⟨(NP DT$_{\boxed{1}}$ NN$_{\boxed{2}}$), $\boxed{1}$ $\boxed{2}$⟩
⟨(DT Every), 每⟩
⟨(NN corner), 一 个 角落⟩
⟨(PP (IN of) NP$_{\boxed{1}}$), $\boxed{1}$ 的⟩
⟨(NP (NNP Hong) (NNP Kong)), 香港⟩
⟨(VP (VBZ is) VP$_{\boxed{1}}$), $\boxed{1}$⟩
⟨(VP VBN$_{\boxed{1}}$ PP$_{\boxed{2}}$), 都 $\boxed{1}$ $\boxed{2}$⟩
⟨(VBN filled), 充 着⟩
⟨(PP (IN with) (NP NN$_{\boxed{1}}$)), $\boxed{1}$⟩
⟨(NN fun), 趣⟩
⟨(. .), 。⟩

Gibbs Sampling: Expand Operator

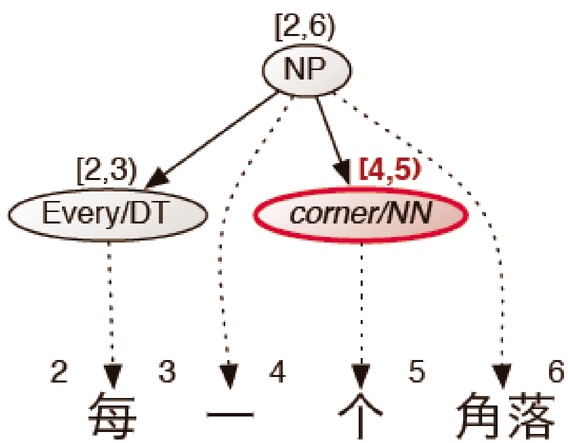


Gibbs Sampling: Expand Operator



$$r_p = \langle (\text{NP DT}_{\boxed{1}}\ (\text{NN corner})), \boxed{1}\ 一\ 个\ 角落 \rangle$$

Gibbs Sampling: Expand Operator



$$r_{p'} = \langle (\text{NP DT}_{\boxed{1}}\ \text{NN}_{\boxed{2}}), \boxed{1}\ \text{一}\ \boxed{2}\ \text{角落} \rangle$$

$$r_v = \langle (\text{NN corner}), \text{个} \rangle$$

Gibbs Sampling

- Also one other operator (Swap)
- A single iteration of Gibbs sampling consists of visiting every sentence pair and:
  - (1) Applying the Expand operator to every node in the tree
  - (2) Then, applying the Swap operator to every applicable pair of nodes in the tree

Experimental Setup

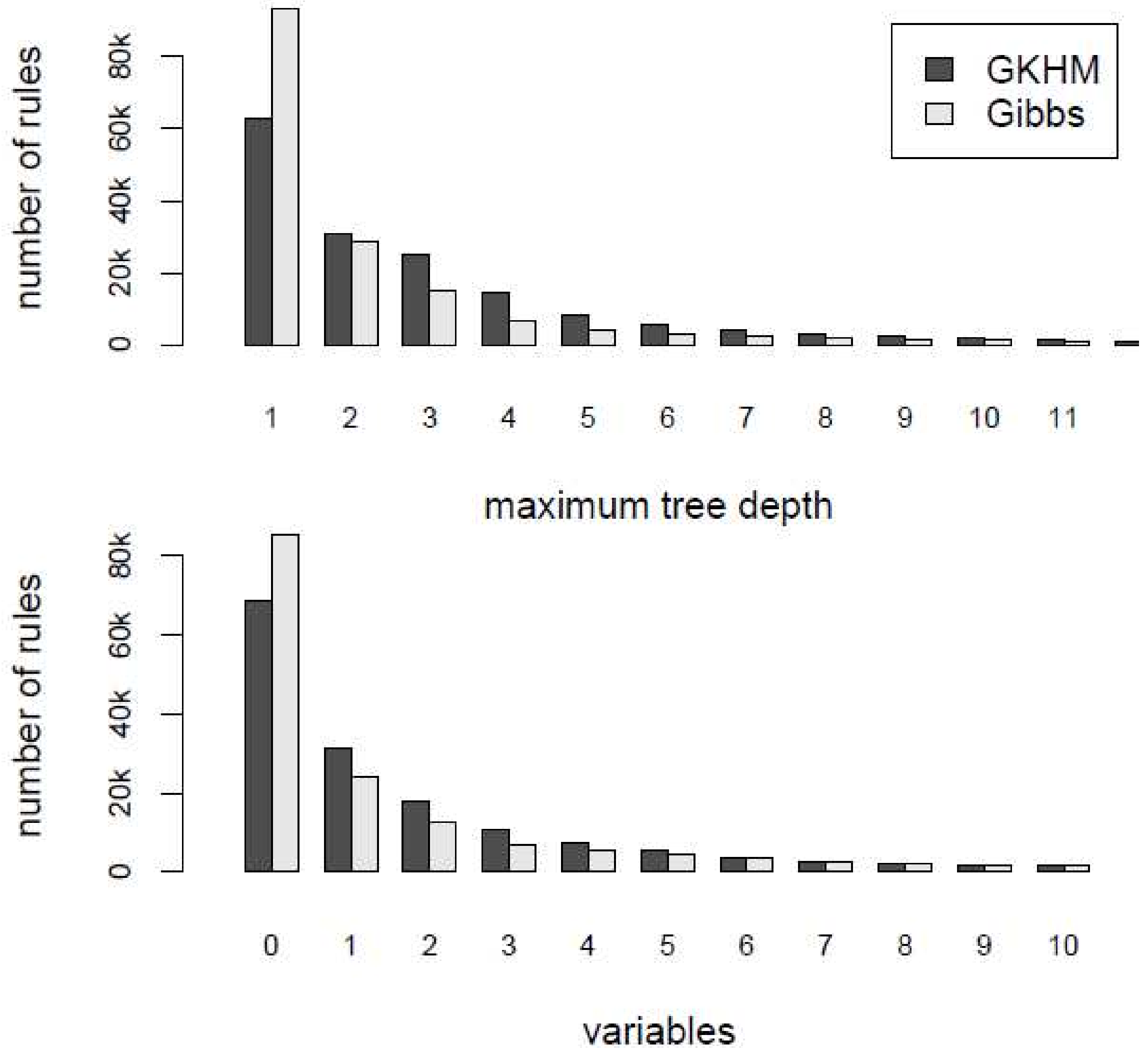
- 300k sentence pairs of Chinese-English
  - FBIS and 100k sentences of Sinorama
- GHKM rule extraction as baseline
- Gibbs sampling run for 300 iterations
  - Initialized using GHKM
  - Took one week
  - Grammar taken from final sample

Results

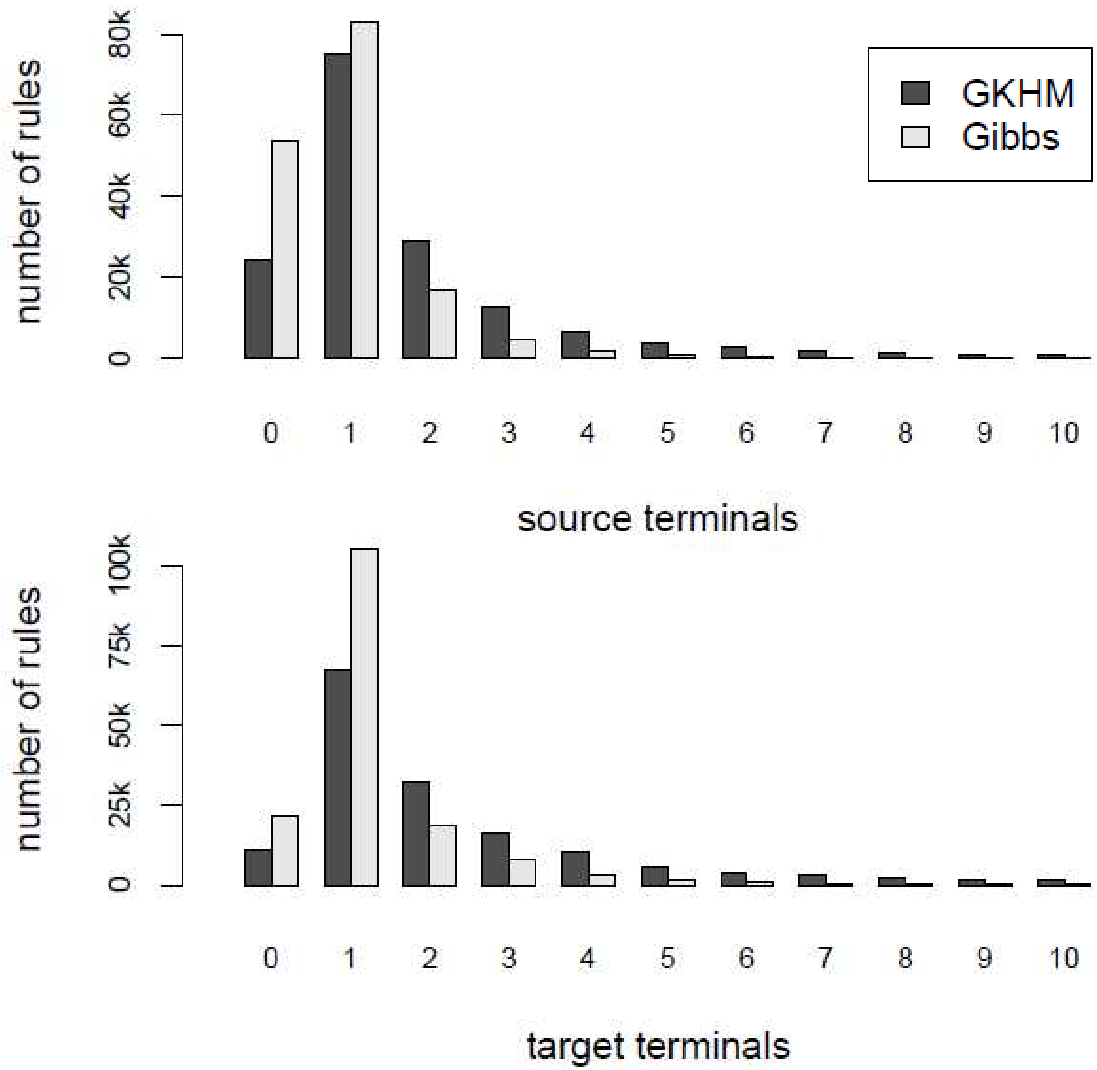
Model	BLEU score
GHKM	26.0
Our model	26.6

- GHKM and sampled grammar have roughly the same number of rules (~1.62 million)
- GHKM has more large rules, sampled grammar has smaller and simpler rules

Results


GKHM
Gibbs
number of rules
0
20k
40k
60k
80k
1 2 3 4 5 6 7 8 9 10 11
maximum tree depth
number of rules
0
20k
40k
60k
80k
0 1 2 3 4 5 6 7 8 9 10
variables

Results


GKHM
Gibbs
number of rules
0
20k
40k
60k
80k
0 1 2 3 4 5 6 7 8 9 10
source terminals
number of rules
0
25k
50k
75k
100k
0 1 2 3 4 5 6 7 8 9 10
target terminals

Example Grammar Rules

Only in sampled grammar
(all appear frequently):

⟨(TOP (S NP$\boxed{1}$ VP$\boxed{2}$.$\boxed{3}$)), $\boxed{1}\boxed{2}\boxed{3}$⟩
⟨(S (VP (TO to) VP$\boxed{1}$)), $\boxed{1}$⟩
⟨(NP NP$\boxed{1}$ (PP (IN of) NP$\boxed{2}$)), $\boxed{2}\boxed{1}$⟩
⟨(PP (IN in) NP$\boxed{1}$), 在 $\boxed{1}$⟩
⟨(NP NP$\boxed{1}$ (PP (IN of) NP$\boxed{2}$)), $\boxed{1}\boxed{2}$⟩
⟨(NP (DT the) NN$\boxed{1}$), 的 $\boxed{1}$⟩
⟨(S (VP TO$\boxed{1}$ VP$\boxed{2}$)), $\boxed{1}\boxed{2}$⟩
⟨(VP (VBZ is) NP$\boxed{1}$), 是 $\boxed{1}$⟩
⟨(NP (NP (DT the) NN$\boxed{1}$) (PP (IN of) NP$\boxed{2}$)), $\boxed{2}\boxed{1}$⟩

Only in GHKM grammar
(all appear very infrequently):

⟨(PP (IN at) (NP DT$\boxed{1}$ (NNS levels))), $\boxed{1}$ 級⟩
⟨(NP NP$\boxed{1}$,$\boxed{2}$ NP$\boxed{3}$ (, ,) CC$\boxed{4}$ NP$\boxed{5}$), $\boxed{1}\boxed{2}\boxed{3}\boxed{4}\boxed{5}$⟩
⟨(NP NP$\boxed{1}$,$\boxed{2}$ NP$\boxed{3}$,$\boxed{4}$ NP$\boxed{5}$ (, ,) (CC and) NP$\boxed{6}$), $\boxed{1}\boxed{2}\boxed{3}\boxed{4}\boxed{5}$, $\boxed{6}$⟩
⟨(S S$\boxed{1}$ (NP (PRP They)) VP$\boxed{2}$.$\boxed{3}$), $\boxed{1}\boxed{2}\boxed{3}$⟩
⟨(S PP$\boxed{1}$,$\boxed{2}$ NP$\boxed{3}$ VP$\boxed{4}$.$\boxed{5}$ “$\boxed{6}$), $\boxed{1}\boxed{2}\boxed{3}\boxed{4}\boxed{6}\boxed{5}$⟩
⟨(S PP$\boxed{1}$,$\boxed{2}$ NP$\boxed{3}$ VP$\boxed{4}$.$\boxed{5}$), $\boxed{1}$ 中 $\boxed{2}\boxed{3}\boxed{4}\boxed{5}$⟩
⟨(NP (NNP Foreign) (NNP Ministry) NN$\boxed{1}$ (NNP Zhu) (NNP Bangzao)), 外交部 $\boxed{1}$ 朱邦造⟩
⟨(S S$\boxed{1}$ S$\boxed{2}$), $\boxed{1}\boxed{2}$⟩
⟨(S S$\boxed{1}$ (NP (PRP We)) VP$\boxed{2}$.$\boxed{3}$), $\boxed{1}\boxed{2}\boxed{3}$⟩
⟨(NP (DT the) (NNS people) POS$\boxed{1}$), 人民 $\boxed{1}$⟩

The GHKM grammar misses many common and useful rules that the sampled grammar finds