A Bayesian Model for Learning SCFGs with Discontiguous Rules

Abby Levenberg
Dept. of Computer Science
University of Oxford
ablev@cs.ox.ac.uk

Chris Dyer
School of Computer Science
Carnegie Mellon Univeristy
cdyer@cs.cmu.edu

Phil Blunsom
Dept. of Computer Science
University of Oxford
pblunsom@cs.ox.ac.uk

Abstract

We describe a nonparametric model and corresponding inference algorithm for learning Synchronous Context Free Grammar derivations for parallel text. The model employs a Pitman-Yor Process prior which uses a novel base distribution over synchronous grammar rules. Through both synthetic grammar induction and statistical machine translation experiments, we show that our model learns complex translational correspondences— including discontiguous, many-to-many alignments—and produces competitive translation results. Further, inference is efficient and we present results on significantly larger corpora than prior work.

1 Introduction

In the twenty years since Brown et al. (1992) pioneered the first word-based statistical machine translation (SMT) models substantially more expressive models of translational equivalence have been developed. The prevalence of complex phrasal, discontiguous, and non-monotonic translation phenomena in real-world applications of machine translation has driven the development of hierarchical and syntactic models based on synchronous context-free grammars (SCFGs). Such models are now widely used in translation and represent the state-of-the-art in most language pairs (Galley et al., 2004; Chiang, 2007). However, while the models used for translation have evolved, the way in which they are learnt has not: naïve word-based models are still used to infer translational correspondences from parallel corpora.

In this work we bring the learning of the minimal units of translation in step with the representational power of modern translation models. We present a nonparametric Bayesian model of translation based on SCFGs, and we use its posterior distribution to infer synchronous derivations for a parallel corpus using a novel Gibbs sampler. Our model is able to: *1*) directly model many-to-many alignments, thereby capturing non-compositional and idiomatic translations; *2*) align discontiguous phrases in both the source and target languages; *3*) have no restrictions on the length of a rule, the number of nonterminal symbols per rule, or their configuration.

Learning synchronous grammars is hard due to the high polynomial complexity of dynamic programming and the exponential space of possible rules. As such most prior work for learning SCFGs has relied on inference algorithms that were heuristically constrained or biased by word-based alignment models and small experiments (Wu, 1997; Zhang et al., 2008; Blunsom et al., 2009; Neubig et al., 2011). In contrast to these previous attempts, our SCFG model scales to large datasets (over 1.3M sentence pairs) without imposing restrictions on the form of the grammar rules or otherwise constraining the set of learnable rules (e.g., with a word alignment).

We validate our sampler by demonstrating its ability to recover grammars used to generate synthetic datasets. We then evaluate our model by inducing word alignments for SMT experiments in several typologically diverse language pairs and across a range of corpora sizes. Our results attest to our model's ability to learn synchronous grammars encoding complex translation phenomena.

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 223–232, Jeju Island, Korea, 12–14 July 2012.

2 Prior Work

The goal of directly inducing phrasal translation models from parallel corpora has received a lot of attention in the NLP and SMT literature. Marcu and Wong (2002) presented an ambitious maximum likelihood model and EM inference algorithm for learning phrasal translation representations. The first issue this model faced was a massive parameter space and intractable inference. However a more subtle issue is that likelihood based models of this form suffer from a degenerate solution, resulting in the model learning whole sentences as phrases rather than minimal units of translation. DeNero et al. (2008) recognised this problem and proposed a nonparametric Bayesian prior for contiguous phrases. This had the dual benefits of biasing the model towards learning minimal translation units, and integrating out the parameters such that a much smaller set of statistics would suffice for inference with a Gibbs sampler. However this work fell short by not evaluating the model independently, instead only presenting results in which it was combined with a standard word-alignment initialisation, thus leaving open the question of its efficacy.

The fact that flat phrasal models lack a structured approach to reordering has led many researchers to pursue SCFG induction instead (Wu, 1997; Cherry and Lin, 2007; Zhang et al., 2008; Blunsom et al., 2009). The asymptotic time complexity of the inside algorithm for even the simplest SCFG models is $O(|\mathbf{s}|^3|\mathbf{t}|^3)$, too high to be practical for most real translation data. A popular solution to this problem is to heuristically restrict inference to derivations which agree with an independent alignment model (Cherry and Lin, 2007; Zhang et al., 2008). However this may have the unintended effect of biasing the model back towards the initial alignments that they attempt to improve upon. More recently Neubig et al. (2011) reported a novel Bayesian model for phrasal alignment and extraction that was able to model phrases of multiple granularities via a synchronous Adaptor Grammar. However this model suffered from the common problem of intractable inference and results were presented for a very small number of samples from a heuristically pruned beam, making interpreting the results difficult.

Blunsom et al. (2009) presented an approach similar to ours that implemented a Gibbs sampler for a nonparametric Bayesian model of ITG. While that work managed to scale to a non-trivially sized corpus, like other works it relied on a state-of-the-art word alignment model for initialisation. Our model goes further by allowing discontiguous phrasal translation units. Surprisingly, the freedom that this extra power affords allows the Gibbs sampler we propose to mix more quickly, allowing state-of-the-art results from a simple initialiser.

3 Model

We use a nonparametric generative model based on the 2-parameter Pitman-Yor process (PYP) (Pitman and Yor, 1997), a generalisation of the Dirichlet Process, which has been used for various NLP modeling tasks with state-of-the-art results such as language modeling, word segmentation, text compression and part of speech induction (Teh, 2006; Goldwater et al., 2006; Wood et al., 2011; Blunsom and Cohn, 2011). In this section we first provide a brief definition of the SCFG formalism and then describe our PYP prior for them.

3.1 Synchronous Context-Free Grammar

An synchronous context-free grammar (SCFG) is a 5-tuple $\langle \Sigma, \Delta, V, S, R \rangle$ that generalises context-free grammar to generate strings concurrently in two languages (Lewis and Stearns, 1968). Σ is a finite set of source language terminal symbols, Δ is a finite set of target language terminal symbols, V is a set of nonterminal symbols, with a designated start symbol S, and R is a set of synchronous rewrite rules. A string pair is generated by starting with the pair $\langle S_1 \mid S_1 \rangle$ and recursively applying rewrite rules of the form $X \rightarrow \langle \mathbf{s}, \mathbf{t}, \mathbf{a} \rangle$ where the left hand side (LHS) X is a nonterminal in V, **s** is a string in $(\Sigma \cup V)^*$, **t** is a string in $(\Delta \cup V)^*$ and **a** specifies a one-to-one mapping (bijection) between nonterminal symbols in **s** and **t**. The following are examples:[1]

VP $\rightarrow \langle$ *schlage* NP_1 NP_2 *vor* $\mid$ *suggest* NP_2 *to* NP_1 $\rangle$
NP $\rightarrow \langle$ *die Kommission* $\mid$ *the commission* $\rangle$

[1]The nonterminal alignment **a** is indicated through subscripts on the nonterminals.

In a probabilistic SCFG, rules are associated with probabilities such that the probabilities of all rewrites of a particular LHS category sum to 1.

Translation with SCFGs is carried out by parsing the source language with the monolingual source language projection of the grammar (using standard monolingual parsing algorithms), which induces a parallel tree structure and translation in the target language (Chiang, 2007). Alignment or synchronous parsing is the process of concurrently parsing both the source and target sentences, uncovering the derivation or derivations that give rise to a string pair (Wu, 1997; Dyer, 2010).

Our goal is to infer the most probable SCFG derivations that explain a corpus of parallel sentences, given a nonparametric prior over probabilistic SCFGs. In this work we will consider grammars with a single nonterminal category X.

3.2 Pitman-Yor Process SCFG

Before training we have no way of knowing how many rules will be needed in our grammar to adequately represent the data. By using the Pitman-Yor process as a prior on the parameters of a synchronous grammar we can formulate a model which prefers smaller numbers of rules that are reused often, thereby avoiding degenerate grammars consisting of large, overly specific rules. However, as the data being fit grows, the model can become more complex. The PYP is parameterised by a *discount* parameter d, a *strength* parameter θ, and the base distribution $\mathcal{G}_0$, which gives the prior probability of an event (in our case, events are rules) before any observations have occurred. The discount is subtracted from each positive rule count and dampens the rich get richer effect where frequent rules are given higher probability compared to infrequent ones. The strength parameter controls the variance, or concentration, about the base distribution.

In our model, a draw from a PYP is a distribution over SCFG rules with a particular LHS (in fact, it is a distribution over all well-formed rules). From this distribution we can in turn draw individual rules:

$$\begin{aligned} \mathcal{G}_X &\sim \mathrm{PY}(d, \theta, \mathcal{G}_0), \\ \mathrm{X} \rightarrow \langle \mathbf{s}, \mathbf{t}, \mathbf{a} \rangle &\sim \mathcal{G}_X. \end{aligned}$$

Although the PYP has no known analytical form, we can marginalise out the $\mathcal{G}_X$'s and reason about individual rules directly using the process described by Teh (2006). In this process, at time n a rule r_n is generated by stochastically deciding whether to make another copy of a previously generated rule or to draw a new one from the base distribution, $\mathcal{G}_0$. Let $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots)$ be the sequence of draws from $\mathcal{G}_0$; thus $|\boldsymbol{\varphi}|$ is the total number of draws from $\mathcal{G}_0$. A rule r_n corresponds to a selection of a φ_k. Let c_k be a counter indicating the number of times φ_k has been selected. In particular, we set r_n to φ_k with probability

$$\frac{c_k - d}{\theta + n},$$

and increment c_k, or with probability

$$\frac{\theta + d \cdot |\boldsymbol{\varphi}|}{\theta + n},$$

we draw a new rule from $\mathcal{G}_0$, append it to $\boldsymbol{\varphi}$, and use it for r_n.

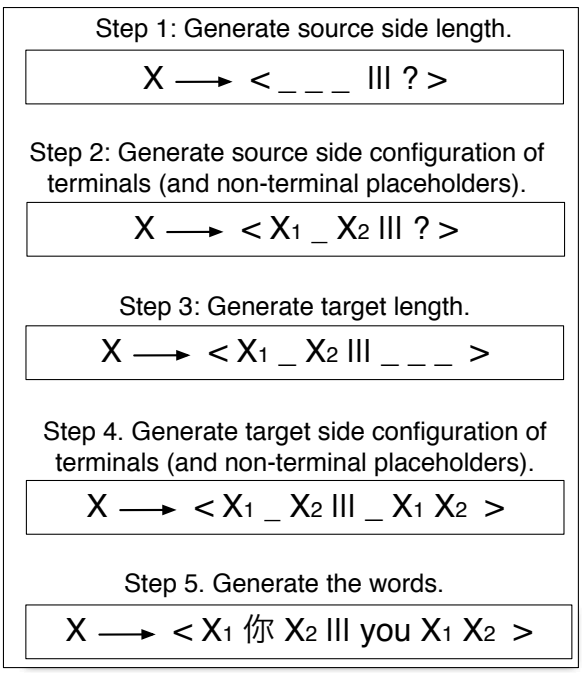


Figure 1: Example generation of a synchronous grammar rule in our $\mathcal{G}_0$.

3.3 Base Distribution

The base distribution $\mathcal{G}_0$ for the PYP assigns probability to a rule based our belief about what constitutes a good rule independent of observing any of

the data. We describe a novel generative process for all rules $X \rightarrow \langle \mathbf{s}, \mathbf{t}, \mathbf{a} \rangle$ that encodes these beliefs.

We describe the generative process generally here in text, and readers may refer to the example in Figure 1. The process begins by generating the source length (total number of terminal and nonterminal symbols, written $|\mathbf{s}|$) by drawing from a Poisson distribution with mean 1:

$$|\mathbf{s}| \sim \text{Poisson}(1) \ .$$

This assigns high probability to shorter rules, but arbitrarily long rules are possible with a low probability. Then, for every position in **s**, we decide whether it will contain a terminal or nonterminal symbol by repeated, independent draws from a Bernoulli distribution. Since we believe that shorter rules should be relatively more likely to contain terminal symbols than longer rules, we define the probability of a terminal symbol to be $\phi^{|\mathbf{s}|}$ where $0 < \phi < 1$ is a hyperparameter.

$$s_i \sim \text{Bernoulli}(\phi^{|\mathbf{s}|}) \quad \forall\, i \in [1, |\mathbf{s}|] \ .$$

We next generate the length of the target side of the rule. Let $\#_{\text{NT}}(\mathbf{s})$ denote the number of nonterminal symbols we generated in **s**, i.e., the arity of the rule. Our intuition here is that source and target lengths should be similar. However, to ensure that the rule is well-formed, **t** must contain exactly as many nonterminal symbols as the source does. We therefore draw the number of target terminal symbols from a Poisson whose mean is the number of terminal symbols in the source, plus a small constant λ_0 to ensure that it is greater than zero:

$$|\mathbf{t}| - \#_{\text{NT}}(\mathbf{s}) \sim \text{Poisson}\left(|\mathbf{s}| - \#_{\text{NT}}(\mathbf{s}) + \lambda_0\right) \ .$$

We then determine whether each position in **t** is a terminal or nonterminal symbol by drawing uniformly from the bag of $\#_{\text{NT}}(\mathbf{s})$ source nonterminals and $|\mathbf{t}| - \#_{\text{NT}}(\mathbf{s})$ terminal indicators, without replacement. At this point we have created a rule template which indicates how large the rule is, whether each position contains a terminal or nonterminal symbol, and the reordering of the source nonterminals **a**. To conclude the process we must select the terminal types from the source and target vocabularies. To do so, we use the following distribution:

$$P_{\text{terminals}}(\mathbf{s}, \mathbf{t}) = \frac{P_{M1\leftarrow}(\mathbf{s}, \mathbf{t}) + P_{M1\rightarrow}(\mathbf{s}, \mathbf{t})}{2}$$

where $P_{M1\leftarrow}(\mathbf{s}, \mathbf{t})$ ($P_{M1\rightarrow}(\mathbf{s}, \mathbf{t})$) first generates the source (target) terminals from uniform draws from the vocabulary, then generates the string in the other language according to IBM MODEL 1, marginalizing over the alignments (Brown et al., 1993).

4 Gibbs Sampler

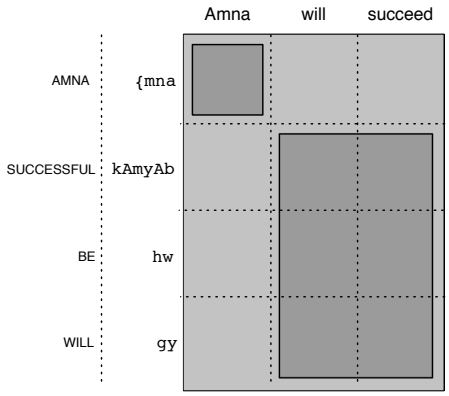
In this section we introduce a Gibbs sampler that enables us to perform posterior inference given a corpus of sentence pairs. Our innovation is to represent the synchronous derivation of a sentence pair in a hierarchical 4-dimensional binary alignment grid, with elements $z_{[s,t,u,v]} \in \{0, 1\}$.

The settings of the grid variables completely determine the SCFG rules in the current derivation. A setting of a binary variable $z_{[s,t,u,v]} = 1$ represents a constituent linking the source span $[s, t]$ and the target span $[u, v]$ in the current derivation; variables with a value of 0 indicate no link between spans $[s, t]$ and $[u, v]$.[2] This relationship from our grid representation is illustrated in Figure 2a.

Our Gibbs sampler operates over the space of all the random variables $z_{[s,t,u,v]}$, resampling one at a time. Changes to a single variable imply that at most *two* additional rules must be generated, as illustrated in Figure 2b. The probability of choosing a binary setting of 0 or 1 for a variable is proportional to the probability of generating the two derivations under the model described in the previous section. Note that for a given sentence, most of the bispan variables must be set to 0 otherwise they would violate the *strict nesting constraint* required for valid SCFG derivations. We discuss below how to exploit this fact to limit the number of binary variables that must be resampled for each sentence.

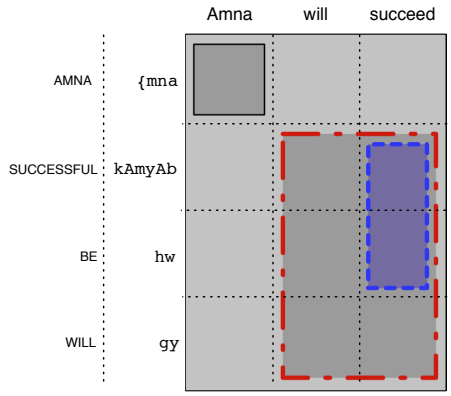
To be valid, a Gibbs sampler must be *ergodic* and satisfy *detailed balance*. Ergodicity requires that there is non-zero probability that any state in the sampler be reachable from any other state. Clearly

[2] Our grid representation is the synchronous generalisation of the well-known correspondence between CFG derivations and Boolean matrices; see Lee (2002) for an overview.



(a) An example grid representation of a synchronous derivation. The SCFG rules (annotated with their bispans) that correspond to this setting of the grid are:

$X_{[0,4,0,3]} \rightarrow$
$\langle\ X_{[0,1,0,1]}\ X_{[1,4,1,3]}\ |\ X_{[0,1,0,1]}\ X_{[1,4,1,3]}\ \rangle$
$X_{[0,1,0,1]} \rightarrow \langle\ \{\texttt{mna}\ |\ \textit{Amna}\ \rangle$
$X_{[1,4,1,3]} \rightarrow \langle\ \texttt{kAmyAb hw gy}\ |\ \textit{will succeed}\ \rangle$



(b) The toggle operator resamples a bispan variable (here, $z_{[1,3,2,3]}$, shown in blue) to determine whether it should be subtracted from the immediately dominating rule (bispan in red) and made into a child rule in the derivation. This would require the addition of the following two rules:

$X_{[1,4,1,3]} \rightarrow \langle\ X_{[1,3,2,3]}\ \texttt{gy}\ |\ \textit{will}\ X_{[1,3,2,3]}\rangle$
$X_{[1,3,2,3]} \rightarrow \langle\ \texttt{kAmyAb hw}\ |\ \textit{succeed}\ \rangle$

Alternatively, the active bispan variable can be set so it is *not* a constituent, which would require the single rule:

$X_{[1,4,1,3]} \rightarrow \langle\ \texttt{kAmyAb hw gy}\ |\ \textit{will succeed}\ \rangle$

Figure 2: A single operation of the Gibbs sampler for a binary alignment grid.

our operator satisfies this since given any configuration of the alignment grid we can use the toggle operator to flatten the derivation to a single rule and then break it back down to reach any derivation.

Detailed balance requires that the probability of transitioning between two possible adjacent sampler states respects their joint probabilities in the stationary distribution. One way to ensure this is to make the order in which bispan variables are visited deterministic and independent of the variables' current settings. Then, the probability of the sampler targeting any bispan in the grid is equal regardless of the current configuration of the alignment grid.

A naive instantiation of this strategy is to visit all $|\mathbf{s}|^2|\mathbf{t}|^2$ bispans in some order. However, since we wish to be able to draw many samples, this is not computationally feasible. A much more efficient approach avoids resampling variables that would result in violations without visiting each of them individually. However, to ensure detailed balanced is maintained, the order that we resample bispans has to match the order we would sample them using any exhaustive approach. We achieve this by always checking a derivation top-down, from largest to smallest bispan. Under this ordering, whether or not a smaller bispan is visited will be independent of how the larger ones were resampled. Furthermore, the set of variables that may be resampled is fixed given this ordering. Therefore, the probability of sampling any possible bispan in the sentence pair is still uniform (ensuring detailed balance), while our sampler remains fast.

5 Evaluation

The preceding sections have introduced a model, and accompanying inference technique, designed to induce a posterior distribution over SCFG derivations containing discontiguous and phrasal translation rules. The evaluation that follows aims to determine our models ability to meet these design goals, and to do so in a range of translation scenarios.

In order to validate both the model and the sampler's ability to learn an SCFG we first conduct a synthetic experiment in which the true grammar is

known. Subsequently we conduct a series of experiments on real parallel corpora of increasing sizes to explore the empirical properties of our model.

5.1 Synthetic Data Experiments

Prior work on SCFG induction for SMT has validated modeling claims by reporting BLEU scores on real translation tasks. However, the combination of noisy data and the complexity of SMT pipelines conspire to obscure whether models actually achieve their design goals, normally stated in terms of an ability to induce SCFGs with particular properties.

Here we include a small synthetic data experiment to clearly validate our models ability to learn an SCFG that includes discontiguous and phrasal translation rules with non-monotonic word order.

Using the probabilistic SCFG shown in the top half of Table 1 we stochastically generated three thousand parallel sentence pairs as training data for our model. We then ran the Gibbs sampler for fifty iterations through the data.

The bottom half of Table 1 lists the five rules with the highest marginal probability estimated by the sampler. Encouragingly our model was able to recover a grammar very close to the original. Even for such a small grammar the space of derivations is enormous and the task of recovering it from a data sample is non-trivial. The divergence from the true probabilities is due to the effect of the prior assigning shorter rules higher probability. With a larger data sample we would expect the influence of the prior in the posterior to diminish.

GRAMMAR RULE	TRUE PROBABILITY
$X \rightarrow \langle\ X_1\ a\ X_2\ \mid X_1\ X_2\ 1\ \rangle$	0.2
$X \rightarrow \langle\ b\ c\ d\ \mid 3\ 2\ \rangle$	0.2
$X \rightarrow \langle\ b\ d\ \mid 3\ \rangle$	0.2
$X \rightarrow \langle\ d\ \mid 3\ \rangle$	0.2
$X \rightarrow \langle\ c\ d\ \mid 3\ 1\ \rangle$	0.2
SAMPLED RULE	**SAMPLED PROBABILITY**
$X \rightarrow \langle\ d\ \mid 3\ \rangle$	0.25
$X \rightarrow \langle\ b\ d\ \mid 3\ \rangle$	0.24
$X \rightarrow \langle\ c\ d\ \mid 3\ 1\ \rangle$	0.24
$X \rightarrow \langle\ b\ c\ d\ \mid 3\ 2\ \rangle$	0.211
$X \rightarrow \langle\ X_1\ a\ X_2\ \mid X_1\ X_2\ 1\ \rangle$	0.012

Table 1: Manually created SCFG used to generate synthetic data, and the five most probable inferred rules by our model.

	ZH-EN NIST	UR-EN NIST	DE-EN EUROPARL
TRAIN (SRC)	8.6M	1.2M	34M
TRAIN (TRG)	9.5M	1.0M	36M
DEV (SRC)	22K	18K	26K
DEV (TRG)	27K	16K	28K

Table 2: Corpora statistics (in words).

5.2 Machine Translation Evaluation

Ultimately the efficacy of a model for SCFG induction will be judged on its ability to underpin a state-of-the-art SMT system. Here we evaluate our model by applying it to learning word alignments for parallel corpora from which SMT systems are induced. We train models across a range of corpora sizes and for language pairs that exhibit the type of complex alignment phenomena that we are interested in modeling: Chinese → English (ZH-EN), Urdu → English (UR-EN) and German → English (DE-EN).

Data and Baselines

The UR-EN corpus is the smallest of those used in our experiments and is taken from the NIST 2009 translation evaluation.[3] The ZH-EN data is of a medium scale and comes from the FBIS corpus. The DE-EN pair constitutes the largest corpus and is taken from Europarl, the proceedings of the European Parliament (Koehn, 2003). Statistics for the data are shown in Table 2. We measure translation quality via the BLEU score (Papineni et al., 2001).

All translation systems employ a Hiero translation model during decoding. Baseline word alignments were obtained by running GIZA++ in both directions and symmetrizing using the `grow-diag-final-and` heuristic (Och and Ney, 2003; Koehn et al., 2003). Decoding was performed with the `cdec` decoder (Dyer et al., 2010) with the synchronous grammar extracted using the techniques developed by Lopez (2008). All translation systems include a 5-gram language model built from a five hundred million token subset

[3] `http://www.itl.nist.gov/iad/mig/tests/mt/2009/`

Language Pair	Test Set	Model 4 Baseline	Model 1 Initialisation	PYP-SCFG Weak M1 Init.	PYP-SCFG Strong HMM Init.
ur-en	mt09	23.1	18.5	23.7	**24.0**
zh-en	mt03-08	29.4	19.8	28.3	**29.8**
de-en	EuroParl	28.4	25.5	27.8	**29.2**

Table 3: Results for the SMT experiments in Bleu . The baseline is produced using a full Giza++ run. The Model 1 Initialisation column is from the initialisation alignments using Model 1 and no sampling. The PYP-SCFG columns show results for the 500th sample for both Model 1 and HMM initialisations.

of all the English data made available for the NIST 2009 shared task (Graff, 2003).

Experimental Setup

To obtain the PYP-SCFG word alignments we ran the sampler for five hundred iterations for each of the language pairs and experimental conditions described below. We used the approach of Newman et al. (2007) to distribute the sampler across multiple threads. The strength θ and discount d hyperparameters of the Pitman-Yor Processes, and the terminal penalty ϕ (Section 3.3), were inferred using slice sampling (Neal, 2000).

The Gibbs sampler requires an initial set of derivations from which to commence sampling. In our experiments we investigated both *weak* and a *strong* initialisations, the former based on word alignments from IBM Model 1 and the latter on alignments from an HMM model (Vogel et al., 1996). For decoding we used the word alignments implied by the derivations in the final sample to extract a Hiero grammar with the same standard set of relative frequency, length, and language model features used for the baseline.

Weak Initialisation

Our first translation experiments ascertain the degree to which our proposed Gibbs sampling inference algorithm is able to learn good synchronous derivations for the PYP-SCFG model. A number of prior works on alignment with Gibbs samplers have only evaluated models initialised with the more complex Giza++ alignment models (Blunsom et al., 2009; DeNero et al., 2008), as a result it can be difficult to separate the performance of the sampler from that of the initialisation. In order to do this, we initialise the sampler using just the Model 1 distribution used in the PYP-SCFG model's base distribution. We denote this a weak initialisation as no alignment models outside of those included in the PYP-SCFG model influence the resulting word alignments. The Bleu scores for translation systems built from the five hundredth sample are show in the Weak M1 Init. column of Table 3. Additionally we build a translation system from the Model 1 alignment used to initialise the sampler without using using our PYP-SCFG model or sampling. Bleu scores are shown in the Model 1 Initialisation column of Table 3. Firstly it is clear Model 1 is indeed a weak initialiser as the resulting translation systems achieve uniformly low Bleu scores. In contrast, the models built from the output of the Gibbs sampler for the PYP-SCFG model achieve Bleu scores comparable to those of the Model 4 Baseline. Thus the sampler has moved a good distance from its initialisation, and done so in a direction that results in better synchronous derivations.

Language Pair	PYP-SCFG Model 1 Init.	PYP-SCFG HMM Init.
ur-en	1.93/2.08	1.45/1.58
zh-en	3.47/4.28	1.69/2.37
de-en	4.05/4.77	1.50/2.04

Table 4: Average source/target rule lengths in the PYP-SCFG models after the 500th sample for the different initialisations.

Strong Initialisation

Given we have established that the sampler can produce state-of-the-art translation results from a

weak initialisation, it is instructive to investigate whether initialising the model with a strong alignment system, the GIZA++ HMM (Vogel et al., 1996), leads to further improvements. Column HMM INIT. of Table 3 shows the results for initialising with the HMM word alignments and sampling for 500 iterations. Starting with a stronger initial sample results in both quicker mixing and better translation quality for the same number of sampling iterations.

Table 4 compares the average lengths of the rules produced by the sampler with both the strong and weak initialisers. As the size of the training corpora increases (UR-EN → ZH-EN → DE-EN) we see that the average size of the rules produced by the weakly initialised sampler also increases, while that of the strongly initialised model stays relatively uniform. Initially both samplers start out with a large number of long rules and as the sampling progresses the rules are broken down into smaller, more generalisable, pieces. As such we conclude from these metrics that after five hundred samples the strongly initialised model has converged to sampling from a mode of the distribution while the weakly initialised model converges more slowly and on the longer corpora is still travelling towards a mode. This suggests that longer sampling runs, and Gibbs operators that make simultaneous updates to multiple parts of a derivation, would enable the weakly initialised model to obtain better translation results.

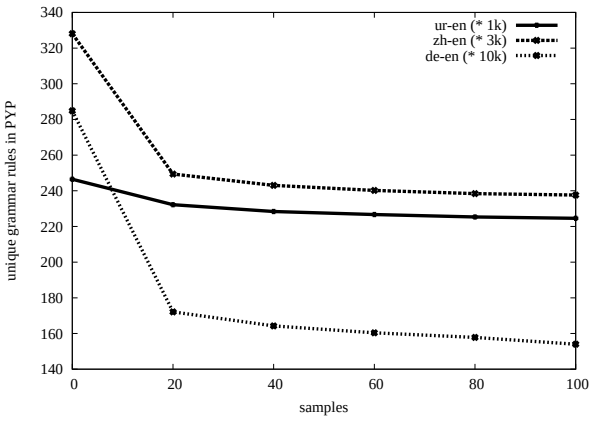


Figure 3: Unique grammar rules for each language pair as a function of the number of samples. The number of rule types decreases monotonically as sampling continues. Rule counts are displayed by normalised corpus size (see Table 2).

X→ ⟨ 底 | end of ⟩
X→ ⟨ 届 全 | ninth ⟩*
X→ ⟨ 运作 X | charter X ⟩
X→ ⟨ 信心 | confidence in ⟩
X→ ⟨ 中国 政府 X | the chinese government X ⟩
X→ ⟨ 都 是 | are ⟩
X→ ⟨ 新华社 北京 X | beijing , X ⟩*
X→ ⟨ 有关 部门 | departments concerned ⟩
X→ ⟨ 新华社 华盛顿 X | washington , X ⟩*
X→ ⟨ 鲍威尔 X_1 了 X_2 , | he X_1 X_2 , ⟩*

Table 5: The five highest ZH-EN probability rules in the Hiero grammar built from the PYP-SCFG that are not in the baseline Hiero grammar (top), and the top five rules in the baseline Hiero grammar that are not in the PYP-SCFG grammar (bottom). An * indicates a bad translation rule.

Grammar Analysis

The BLEU scores are informative as a measure of translation quality but we also explored some of the differences in the grammars obtained from the PYP-SCFG model compared to the standard approach. In Figures 3 and 4 we show some basic statistics of the grammars our model produces. From Figure 3 we see that the number of unique rules in the PYP-SCFG grammar decreases steadily as the sampler iterates through the data, so the model is finding an increasingly sparser distribution with fewer but better quality rules as sampling progresses. Note that the gradient of the curves appears to be a function of the size of the corpus and suggests that the model built from the large DE-EN corpus would benefit from a longer sampling run. Figure 4 shows the distribution of rules with a given arity as a percentage of the full grammar after the final sampling iteration. The model prior biases the results to shorter rules as the vast majority of the model probability mass is on rules with zero, one or two nonterminals.

Tables 5 and 6 show the most probable rules in the Hiero translation system obtained using the PYP-SCFG alignments that are not present in the TM from the GIZA++ alignments and visa versa. For both language pairs, four of the top five rules in

X→ ⟨ yh | it is ⟩
X→ ⟨ zmyn | the earth ⟩
X→ ⟨ yhy X | the same X ⟩
X→ ⟨ X_1 nhyN X_2 gy | X_2 not be X_1 ⟩
X→ ⟨ X_1 gY kh X_2 | recommend that X_2 X_1 ⟩*

X→ ⟨ hwN gY | will ⟩
X→ ⟨ Gyr mlky | international ⟩*
X→ ⟨ X_1 *rAye kY X_2 | X_2 to X_1 sources ⟩*
X→ ⟨ nY X_1 nhyN kyA X_2 | did not X_1 X_2 ⟩*
X→ ⟨ xAtwn X_1 ky X_2 | woman X_2 the X_1⟩

Table 6: Five of the top scoring rules in the UR-EN Hiero grammar from sampled PYP-SCFG alignments (top) versus the baseline UR-EN Hiero grammar rules not in the sampled grammar (bottom). An * indicates a bad translation rule.

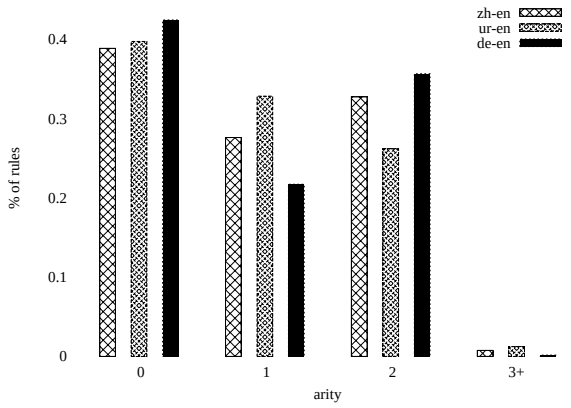


Figure 4: The percentage of rules with a given arity in the final grammar of the PYP-SCFG model.

the PYP-SCFG grammar that are not in the heuristically extracted grammar are correct and minimal phrasal units of translation, whereas only two of the top probability rules in the GIZA++ grammar are of good translation quality.

6 Conclusion and Further Work

In this paper we have presented a nonparametric Bayesian model for learning SCFGs directly from parallel corpora. We have also introduced a novel Gibbs samplller that allows for efficient posterior inference. We show state-of-the-art results and learn complex translation phenomena, including discontiguous and many-to-many phrasal alignments, without applying any heuristic restrictions on the model to make learning tractable. Our evaluation shows that we can use a principled approach to induce SCFGs designed specifically to utilize the full power of grammar based SMT instead of relying on complex word alignment heuristics with inherent bias.

Future work includes the obvious extension to learning SCFGs that contain multiple nonterminals instead of a single nonterminal grammar. We also expect that expanding our sampler beyond strict binary sampling may allow us to explore the space of hierarchical word alignments more quickly allowing for faster mixing. We expect with these extensions our model of grammar induction may further improve translation output.

Acknowledgements

This work was supported by a grant from Google, Inc. and EPRSRC grant no. EP/I010858/1 (Levenberg and Blunsom), the U. S. Army Research Laboratory and U. S. Army Research Office under contract/grant no. W911NF-10-1-0533 (Dyer).

References

P. Blunsom and T. Cohn. 2011. A hierarchical pitman-yor process hmm for unsupervised part of speech induction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 865–874, Portland, Oregon, USA, June. Association for Computational Linguistics.

P. Blunsom, T. Cohn, C. Dyer, and M. Osborne. 2009. A gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 782–790, Suntec, Singapore, August. Association for Computational Linguistics.

P. F. Brown, V. J. D. Pietra, R. L. Mercer, S. A. D. Pietra, and J. C. Lai. 1992. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40.

P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

C. Cherry and D. Lin. 2007. Inversion transduction grammar for joint phrasal translation modeling.

In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 17–24, Rochester, New York, April. Association for Computational Linguistics.

D. Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

J. DeNero, A. Bouchard-Côté, and D. Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 314–323, Honolulu, Hawaii, October. Association for Computational Linguistics.

C. Dyer, A. Lopez, J. Ganitkevitch, J. Weese, F. Ture, P. Blunsom, H. Setiawan, V. Eidelman, and P. Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 7–12.

C. Dyer. 2010. Two monolingual parses are better than one (synchronous parse). In *Proc. of NAACL*.

M. Galley, M. Hopkins, K. Knight, and D. Marcu. 2004. What's in a translation rule? In D. M. Susan Dumais and S. Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 273–280, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

S. Goldwater, T. L. Griffiths, and M. Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Syndney, Australia.

D. Graff. 2003. English Gigaword. Linguistic Data Consortium (LDC-2003T05).

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

P. Koehn. 2003. Europarl: A multilingual corpus for evaluation of machine translation.

L. Lee. 2002. Fast context-free grammar parsing requires fast Boolean matrix multiplication. *Journal of the ACM*, 49(1):1–15.

P. M. Lewis, II and R. E. Stearns. 1968. Syntax-directed transduction. *J. ACM*, 15:465–488, July.

A. Lopez. 2008. *Machine Translation by Pattern Matching.* Ph.D. thesis, University of Maryland.

D. Marcu and D. Wong. 2002. A phrase-based,joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 133–139. Association for Computational Linguistics, July.

R. Neal. 2000. Slice sampling. *Annals of Statistics*, 31:705–767.

G. Neubig, T. Watanabe, E. Sumita, S. Mori, and T. Kawahara. 2011. An unsupervised model for joint phrase alignment and extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 632–641, Portland, Oregon, USA, June. Association for Computational Linguistics.

D. Newman, A. Asuncion, P. Smyth, and M. Welling. 2007. Distributed inference for latent dirichlet allocation. In *NIPS*. MIT Press.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

J. Pitman and M. Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25:855–900.

Y. W. Teh. 2006. 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.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, pages 836–841, Morristown, NJ, USA. Association for Computational Linguistics.

F. Wood, J. Gasthaus, C. Archambeau, L. James, and Y. W. Teh. 2011. The sequence memoizer. *Communications of the Association for Computing Machines*, 54(2):91–98.

D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23:377–403, September.

H. Zhang, C. Quirk, R. C. Moore, and D. Gildea. 2008. Bayesian learning of non-compositional phrases with synchronous parsing. In *Proceedings of ACL-08: HLT*, pages 97–105, Columbus, Ohio, June. Association for Computational Linguistics.