Scalable Variational Inference for Extracting Hierarchical Phrase-based Translation Rules

Baskaran Sankaran
Simon Fraser University
Burnaby BC, Canada
baskaran@cs.sfu.ca

Gholamreza Haffari
Monash University
Clayton VIC, Australia
reza@monash.edu

Anoop Sarkar
Simon Fraser University
Burnaby BC, Canada
anoop@cs.sfu.ca

Abstract

We present a Variational-Bayes model for learning rules for the Hierarchical phrase-based model directly from the phrasal alignments. Our model is an alternative to heuristic rule extraction in hierarchical phrase-based translation (Chiang, 2007), which uniformly distributes the probability mass to the extracted rules locally. In contrast, in our approach the probability assigned to a rule is globally determined by its contribution towards all phrase pairs and results in a sparser rule set. We also propose a distributed framework for efficiently running inference for realistic MT corpora. Our experiments translating Korean, Arabic and Chinese into English demonstrate that they are able to exceed or retain the performance of baseline hierarchical phrase-based models.

1 Introduction

Hierarchical phrase-based translation (Hiero) as described in (Chiang, 2005; Chiang, 2007) uses a synchronous context-free grammar (SCFG) derived from heuristically extracted phrase pairs obtained by symmetrizing bidirectional many-to-many word alignments (Och and Ney, 2004). The phrase-pairs are constrained by the source-target alignments such that all the alignment links from the source (target) words are connected to the target (source) words within the phrase. Given a word-aligned sentence pair \( \langle f_i^j, e_i^j, A \rangle \), where \( A \) indicate the alignments, the source-target sequence pair \( \langle f_i^j, e_i^j \rangle \) can be a phrase-pair iff the following alignment constraint is satisfied.

\[
(k, k') \in A : k \in [i, j] \Leftrightarrow k' \in [i', j']
\]

Given the phrase-pairs, SCFG rules are extracted by replacing aligned sequences of words in source and target sides by co-indexed non-terminals and rewriting the replaced source-target word sequences as separate rules. Consider a rule \( X \rightarrow \langle \beta, \gamma \rangle \), where \( \beta \) and \( \gamma \) are sequences of terminals and non-terminals. Now, given another rule \( X \rightarrow \langle f_i^j, e_i^j \rangle \), such that \( f_i^j \) and \( e_i^j \) are contained fully within \( \beta \) and \( \gamma \) as sub-phrases, the larger rule could be rewritten to create a new rule.

\[
X \rightarrow \langle \beta_p, X_k, \beta_s, \gamma_p, X_k, \gamma_s \rangle
\]  

And target sides by co-indexed non-terminals and rewriting the replaced source-target word sequences as separate rules. Consider a rule \( X \rightarrow \langle \beta, \gamma \rangle \), where \( \beta \) and \( \gamma \) are sequences of terminals and non-terminals. Now, given another rule \( X \rightarrow \langle f_i^j, e_i^j \rangle \), such that \( f_i^j \) and \( e_i^j \) are contained fully within \( \beta \) and \( \gamma \) as sub-phrases, the larger rule could be rewritten to create a new rule.

Here \( \beta_p \) (\( \beta_s \)) refers to any prefix (suffix) of \( \beta \) that precedes (follows) \( f_i^j \). Note that the non-terminals are co-indexed with a unique index so that they are rewritten simultaneously.

Figure 1: Chinese-English phrase-pair with alignments

As a concrete example, consider the word aligned Chinese-English phrase pair shown in Figure 1. Notice that the phrase 联合国 (united nations) is incorrectly aligned to English determiner the, even though in ideal case the entire Chinese phrase 联合国 难民 专员 公署 should be aligned on the English side to the unhcr. The heuristic approach extracts 32 rules, some of which are shown in Figure 2.

The distribution of the rules is unknown, as the different derivations of the sentences are not explicitly observed. Thus, Chiang (2007) follows an approach similar to that of Bod (1998) and hypothesizes a distribution over the rules. Under this each phrase-pair is assumed to have a unit count, which is uniformly distributed to all the rules extracted from this phrase-pair. The locally assigned rule counts are then aggregated across the entire set of phrase-pairs. The probability for each phrase pair

International Joint Conference on Natural Language Processing, pages 438–446, Nagoya, Japan, 14-18 October 2013.

438
Figure 3: Rules extracted for translating the Chinese phrase 在中南海. The probabilities are shown for grammar extracted from heuristic as well our proposed method. The least preferred translations are shown with *. Our Variational-Bayes method extracts a grammar having a peaked distribution as shown.

* X \rightarrow \langle在 中南海 | at chong nan hai\rangle
\begin{align*}
P_{Heu}(e|f) & = 0.333 \\
P_{VB}(e|f) & = 0.003
\end{align*}

\begin{align*}
P_{Heu}(e|f) & = 0.333 \\
P_{VB}(e|f) & = 0.008
\end{align*}

\begin{align*}
P_{Heu}(e|f) & = 0.333 \\
P_{VB}(e|f) & = 0.988
\end{align*}

\begin{align*}
X & \rightarrow \langle在 中南海 | at zhongnanhai\rangle
\end{align*}

Figure 2: Rules extracted for the example phrase-pair in Figure 1. The rules encoding incorrect translations are marked with *.

is then estimated using relative frequency estimation.

1.1 Motivation

A major problem with this heuristic rule extraction method is the lack of global re-weighting of the pseudo-counts beyond their local assignments. By assigning uniform weight to the rules, Chiang (2007) assumes all the rules extracted from a given phrase-pair to be equally good. However, some rules might be better than others in terms of generalization, for capturing a syntactic phrase-pair, or being a semantically coherent unit of translation.

Due to this uniform treatment of good and poor translations, probability mass is wasted on poor translation candidates. For example the phrase-pair in Fig 1 would generate several poor translation rules (shown with * in Fig. 2). This is due to the incorrect alignment link between 联合国 and the (note that the word the is typically aligned with a large number of words due to its frequency). The heuristic extraction method simply assigns uniform count to all translations and as a result the first translation in Fig. 2 becomes the fourth best translation for this source phrase.

In Chiang (2007) the rule extraction algorithm produces a fairly flat distribution over rules. For example the different translation options of the Chinese phrase 在中南海 (at zhongnanhai) all have the same $p(e|f)$ probability as shown in Figure 3. In contrast, our method produces a peaked distribution and shifts the probability mass towards at zhongnanhai, which is the preferred translation.

In this paper, we propose a method which distributes the probability mass among the rules (generated from a phrase-pair) based on their contribution in explaining the collection of all phrase-pairs in a global manner. This difference in estimation methods can lead to a peaked distribution of rule probabilities. Secondly we also present a distributed framework that enables rule extraction on large datasets that are typical in SMT. Our Variational-Bayes approach for rule extraction improves/retains the translation quality for the three different language pairs. Finally, we also present a detailed analysis comparing the extracted SCFG with the heuristically extracted SCFG.

2 Model

Our model uses the notion of a derivation: the set of rules that fully derive an aligned phrase pair, and learns the estimates for the rules contained in the derivations through Variational inference. Setting the notations, we denote the set of derivations for a given phrase pair $x$ as $\phi_x$ and the set of all rules as $G$. Given the set of initial phrase pairs $X'$ and a prior over the grammars $G$, we formulate Hiero grammar extraction as the task of inferring a posterior distribution over Hiero grammars. Using Bayes’ rule, we can express the posterior over the grammar $G$ given the set of bilingual phrases $X'$ as: $P(G|X') \propto P(G)P(X'|G)$.

As mentioned earlier, our model replaces the heuristic rule extraction step in Hiero pipeline. Consequently our model assumes the existence of initial phrase-pairs obtained from bidirectional symmetrization of word alignments. We use the following two-step generative story to create an aligned phrase pair from the Hiero rules.
1. First decide the derivation type \( z_d \) for generating the aligned phrase pair \( x \). It can either be a terminal derivation or hierarchical derivation with one/two gaps.\(^1\) i.e., \( z_d = \{\text{TERM, HIER-A1, HIER-A2}\} \).

2. Then identify the constituent rules \( r \) in the derivation to generate the phrase pair.

Under this model the probability of a particular derivation \( d \in \phi_x \) for a given phrase pair \( x \) can be expressed as:

\[
p(d) \propto p(z_d) \prod_{r \in d} p(r|G, \theta)
\]

where \( r \) is a rule in grammar \( G \) and \( \theta \) is the grammar parameter.

Figure 4 depicts the generative story of our generative model. The derivation-type \( z_d \) is sampled from a multinomial distribution parameterized by \( \phi^z \), where \( \phi^z \) is distributed itself by a Dirichlet distribution with hyper-parameter \( \alpha_z \). The grammar rules are generated from a multinomial distribution parameterized by \( \theta \), where \( \theta \) itself is distributed according to a Dirichlet distribution parameterized by a concentration parameter \( \alpha_h \) and a base distribution \( p_0 \). For the base distribution, we use a simple but yet informative prior based on geometric mean of the bidirectional alignment scores. This allows us to only explore the rules that would be consistent with the underlying word alignments.\(^2\) Thus our setting closely resembles that of the Hiero heuristic rule extraction.

Our goal is thus to infer the joint posterior \( p(\theta, \Phi|\alpha_h, p_0, \alpha_z, \lambda) \), where \( \theta \) are the model parameters and \( \Phi \) the latent derivations over all the phrase pairs.

\(^1\)This refers to the maximum arity of a rule involved in the derivation.

\(^2\)While a non-parametric prior would be better from a Bayesian perspective, we leave it for future consideration.

3 Training

For inference we resort to a variational approximation and factorize the posterior distributions over grammar parameters \( \theta \) and latent derivations \( \Phi \) as:

\[
p(\theta, \Phi|\alpha_h, p_0, \alpha_z, \lambda) \approx q(\theta|u)q(\Phi|\pi)
\]

where \( u \) and \( \pi \) are the parameters of the variational distributions.

The inference is then performed in an EM-style algorithm iteratively updating the parameters \( u \) and \( \pi \). We initialize \( u^t := \alpha_h p_0 \), which is then updated with expected rule counts in subsequent iterations. The expected count for a rule \( r \) at time-step \( t \) can be written as:

\[
\mathbb{E}[r^t] = \sum_{d \in \phi_x} p(d|\pi^{t-1}, x) f_d(r)
\]

where \( p(d|\pi^{t-1}, x) \) is the probability of the derivation \( d \) for the phrase pair \( x \) and \( f_d(r) \) is the frequency of the rule \( r \) in derivation \( d \). The \( p(d|.) \) term in Equation 3 can then be written in terms of \( \pi \) as:

\[
p(d|\pi^{t-1}, x) \propto p(z_d) \prod_{r \in d} \pi_r^{t-1}
\]

The \( p(d|.) \) are normalized across all the derivations of a given phrase pair to yield probabilities. For each derivation \( z_d \), its expected count (at time \( t \)) is the sum of the probabilities of all the derivations of its type.

\[
\mathbb{E}[z^t_d] = \sum_x \sum_{z_d = z_d'} p(d'|\pi^{t-1}, x)
\]

We initialize the Dirichlet hyperparameters \( \alpha_{zd} \) using a Gamma prior ranging between \( 10^{-1} \) and \( 10^3 \); \( \alpha_{zd} \sim \text{Gamma}(10^{-1}, 10^3) \).

\(^3\)In initial experiments we used an initial prior of \( \alpha_z = [10^2, 10, 10^2] \) to compensate for the smaller probabilities for arity-2 derivation resulting from two multiplications. However, our later experiments showed it to be unnecessary and so we used an initial prior that does not prefer any particular outcome.
Algorithm 1 Variational-Bayes for Hiero Rules

Input: Set of aligned phrase-pairs \( X \)

Get prior distribution \( \mathbf{u} = \{ u_r = \alpha_{h_0} | r \in G \} \)

Set \( \mathbf{u}^t = \mathbf{u} \)

for time-step \( t = 1, 2, \ldots \) do

for \( z_d \in Z \) do

\[ p(z_d) \leftarrow \exp \left( \psi(\alpha_{z_d}^{t-1}) - \psi(\sum_{z_d} \alpha_{z_d}^{t-1}) \right) \]

end for

for \( r \in G \) do

\[ \pi_r^{t-1} \leftarrow \exp \left( \psi(u_r^{t-1}) - \psi(\sum_r u_r^{t-1}) \right) \]

end for

for \( x \in X \) do

for \( d \in G \) do

Compute \( p(d|\pi^{t-1}, x) \) as in (4)

\[ E[z_d] \leftarrow E[z_d] + p(d|\pi^{t-1}, x) \]

end for

end for

for \( z_d \in Z \) do

Estimate \( \alpha_{z_d}^t \leftarrow \alpha_{z_d}^{t-1} + E[z_d] \)

end for

for \( r \in G \) do

Estimate posterior \( \mathbf{u}^t: u_r^t \leftarrow u_r^{t-1} + E[r] \)

end for

end for

Output: Posterior distribution \( \mathbf{u}^t \)

---

We run inference for a fixed number of iterations\(^4\) and use the grammar along with their posterior counts from the last iteration for the translation table. Following (Sankaran et al., 2011), we use the shift-reduce style algorithm to efficiently encode the word aligned phrase-pair as a normalized decomposition tree (Zhang et al., 2008). The possible derivations (that are consistent with the word alignments) could then be enumerated by simply traversing every node in the decomposition tree and replacing its span by a non-terminal \( X \).

### 3.1 Distributing Inference

While the above training procedure works well for smaller datasets, it does not scale well for the realistic MT datasets (which have millions of sentence pairs) due to greater memory and time requirements. To address this shortcoming, we distribute the training using a Map-Reduce style framework, where each node works on the local dataset in computing the required statistics and then communicates the statistics to a central aggregator reduce node.

Distributed inference for Expectation Maximization algorithm was studied in (Wolfe et al., 2008). They used three different topologies in terms of computation time, bandwidth requirement and so on. While Map-Reduce is substantially slower than the All-pairs and Junction-tree topologies, it takes much lesser bandwidth than the other two apart from being much easier to implement. Furthermore our choice of the Variational inference naturally lends itself to distributed training.

We simply shard the set of aligned phrase pairs and parallelize the training steps for the shards across different nodes. At the end of local computation of the statistics (expected rule counts for example), we need to aggregate the statistics to get a global view, which will then be used in the next iteration/training step. We parallelize this aggregation across several nodes in one or two reduce steps as required. At the end of aggregation we communicate the updated statistics to each node on a need basis.\(^5\)

### 4 Experiments

We experiment with three datasets of varying sizes. We use the University of Rochester Korean-English dataset consisting of almost 60K sentence pairs for the small data setting. For moderate and large datasets we use Arabic-English (ISI parallel corpus) and Chinese-English (ISI parallel text and GALE phase-1) corpora. We use the MTC dataset having 4 references for tuning and testing for our Chinese-English experiments. The statistics of the corpora used in our experiments are summarized in Table 1.

<table>
<thead>
<tr>
<th>Lang</th>
<th>Training Corpus</th>
<th>Train/ Tune/ Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ko-En</td>
<td>URochester data</td>
<td>59218/ 1118/ 1118</td>
</tr>
<tr>
<td>Ar-En</td>
<td>ISI Ar-En corpus</td>
<td>1.1 M/ 1982/ 987</td>
</tr>
<tr>
<td>Cn-En</td>
<td>HK + GALE ph-1</td>
<td>2.3 M/ 1928/ 919</td>
</tr>
</tbody>
</table>

Table 1: Corpus Statistics in # of sentences

We follow the standard MT practice and use GIZA++ (Och and Ney, 2003) for word aligning the parallel corpus. We then use the heuristic step that symmetrizes the bidirectional alignments (Och et al., 1999) to extract the initial phrase-pairs up to a certain length, consistent with the word alignments. Finally we employ our proposed Variational-Bayes training to learn rules for

---

\(^4\)In our experiments, we set the number of iterations to 10.

\(^5\)We simulate the Map-Reduce style of computation using a regular high-performance cluster using a mounted filesystem rather than a Hadoop cluster with a distributed filesystem.
Table 2: BLEU scores for baseline heuristic extraction and the proposed Variational-Bayes model. Best scores are in boldface and statistically significant differences are italicized.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ko-En</th>
<th>Ar-En</th>
<th>Cn-En</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>7.18</td>
<td>37.82</td>
<td>28.58</td>
</tr>
<tr>
<td>Variational-Bayes</td>
<td>7.68</td>
<td>37.76</td>
<td>28.40</td>
</tr>
</tbody>
</table>

Hiero. As a baseline Hiero model, we use the heuristic rule extraction (Chiang, 2007) approach to extract the rules. In both cases the parameters are estimated by the relative frequency estimation.

For decoding we use our in-house hierarchical phrase-based system Kriya6 (Sankaran et al., 2012b). We use the following 8 standard features for the log-linear model: translation probabilities \( p(e|f) \) and \( p(f|e) \), lexical probabilities \( p_l(e|f) \) and \( p_l(f|e) \), phrase and word penalties, language model and glue rule penalty.

4.1 Results

The main BLEU score results are summarized in Table 2 and the key aspects are summarized below.

- **Higher BLEU scores**: Our Bayesian model performs better than the baseline heuristic rule extractor for Korean-English. Furthermore, the improvement of 0.5 BLEU is statistically significant at \( p \)-value of 0.01.

- **Large corpora**: Our distributed inference model easily scales to the large corpora and the inference completes in less than a day for Chinese-English. It also retains BLEU scores in the same level as the baseline models for both Arabic-English and Chinese-English.

4.2 Compact Models

Some earlier research on Hiero have explored model size reduction as a means of reducing the time and space complexity of the Hiero decoder as well as for mitigating issues such as overgeneration (Setiawan et al., 2009). These approaches use a variety of compression strategies, viz. threshold pruning (Zollmann et al., 2008), pattern-based filtering (He et al., 2009; Iglesias et al., 2009) and significance pruning (Yang and Zheng, 2009).

While compact models is not the central idea of our work, we nevertheless explore the effect of a simple threshold pruning strategy on the grammar learned from our proposed model. Table 3 shows the results for the pruned grammars, where we prune the rules having expected count below a mincount threshold. We present the results for specific mincount settings based on our experiments on the held-out tuning set for each language pair.

<table>
<thead>
<tr>
<th>Model</th>
<th>Ko-En</th>
<th>Ar-En</th>
<th>Cn-En</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU: VB</td>
<td>7.68</td>
<td>37.76</td>
<td>28.40</td>
</tr>
<tr>
<td>Pruning mincount</td>
<td>0.25</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Model size: pruned</td>
<td>1.65</td>
<td>58.9</td>
<td>87.3</td>
</tr>
<tr>
<td>Reduction:</td>
<td>38.2%</td>
<td>82.2%</td>
<td>81.5%</td>
</tr>
<tr>
<td>BLEU: pruned</td>
<td>7.64</td>
<td>37.58</td>
<td>28.45</td>
</tr>
</tbody>
</table>

Table 3: Model sizes (in millions) and BLEU scores of the full VB and pruned VB grammars. Mincount implies the expected rule count threshold used for pruning the full VB grammar. Best/indistinguishable BLEU scores are shown in boldface.

- **Retains score with smaller grammar**: The pruned grammars retain the performance of the full grammar, even while using just 18% of the complete model.

- **Higher reduction for large dataset**: Variational inference reduces the model size over 80% for the large corpora. While this is similar to the findings of Johnson et al. (2007) and that of the pruning strategies mentioned above; the question of whether an intelligent model selection strategy can yield higher BLEU scores is still open.

- **Faster decoding**: The compact grammars naturally result in faster decoding and we observed up to 20-30% speedup in the translation including the time spent for loading the model.

Sankaran et al. (2012a) proposed a model for extracting compact Hiero grammar with restricted arity (at most 1 non-terminal). In contrast our model is close the classical Hiero model Chiang (2007) having an arity of two. Though our results are not directly comparable to theirs, we nevertheless find our model to yield a better model size reduction than theirs. While they claim up to 57%
reduction, we achieve over 80% for the two large data conditions and about 38% reduction for the Korean-English small data setting.

4.3 Analysis

We now compare the probability distributions of the two grammars at the level of individual rules to understand the differences between them. We considered a set of source phrases that are common in both grammars and analyzed their probability distributions over the translation options.

Specifically we use the Q-Q plot to study the behaviour of two probability distributions as explained below considering the Chinese phrase 联合国 (united nations) as a representative example. The Q-Q plot in Figure 5 plots the $p(e|f)$ probabilities (sorted for the baseline grammar) for different translations of the source phrase. The translations from the baseline grammar are then paired off with the points in the sorted VB curve and the corresponding probabilities are plotted in the same order as the baseline translations. The following conclusions can be drawn from this plot:

- **Penalize poor translations**: Among the low-probability translations, majority of the translations in the VB-grammar have probability less than the corresponding baseline translations. This has the desired property of potentially shifting the probability mass away from poor translations.

- **Reward good translations**: VB-grammar rewards some translations that were deemed to be poor by the heuristic method, by assigning a slightly higher probability than the heuristic grammar. A manual inspection showed that the rules with higher probabilities were objectively better translation rules. For example Table 4 contrasts the probabilities assigned by the two methods for the first four translation options in Fig. 5.

- **Uniform probability is not informative**: The heuristic extraction method tends to assign an uniform probability for groups of translations and this is evident in the flat segments of the baseline curve and is especially dominant in the low probability region. In contrast, the VB-grammar is more peaked (in Fig. 5 the probabilities are sorted for the VB grammar).

| Translations   | Heu $p(e|f)$ | VB $p(e|f)$ |
|----------------|-------------|-------------|
| alert the united nations | 4.28e-04 | 3.46e-04 |
| during          | 4.28e-04 | 3.20e-04 |
| for un          | 4.28e-04 | 5.02e-04 |
| human           | 4.28e-04 | 3.20e-04 |

Table 4: Probabilities assigned by the two methods for the first four translations in Fig. 5. The better translation among four and the higher probability assigned by our model are italicized.

We also observe similar trend for several source phrases in both Arabic-English and Chinese-English corpora.

At the macro level, we compare the sizes of the different types of rules in the heuristic and the Variational-Bayes grammar. The baseline grammar extract slightly more rules with arity-1 than the grammar extracted by our model (see Figure 6). Our model extracts rules used in a derivation of a phrase-pair, only if all its constituent rules are consistent with the Hiero rule constraints (such as restriction on the total number of terminals and non-terminals in the rule). However the heuristic method extracts all the consistent rules and does not consider the derivations. While this is a more stricter constraint, the VB model extracts slightly more (about 170K) arity-2 rules as we allow the unaligned words to be attached to different levels of hierarchical rules during the construction of the decomposition tree. This extracts translation rules that are beyond the purview of the heuristic method, since the Viterbi alignments cannot capture them.

We earlier examined the effect of pruning the VB grammar in Section 4.2 and noted that the grammar could be substantially reduced for different language pairs without sacrificing translation quality. In this context, we compare the effects of pruning the heuristic and VB grammars in Figure 6 for Chinese-English. For the same mincount threshold of 1 as the best performing VB setting, the BLEU score of the heuristic grammar drops by over 1 point. However this setting prunes over 99% of the arity-1 and arity-2 rules even while it retains all the terminal rules. This is primarily because of the way the heuristic method estimates rule counts by uniformly distributing the weight among all the rules. The terminal rules are sufficient for coverage but does not capture long distance movements; and the lack of arity-1 and arity-2 rules further restrict the reordering ability of the model. We have to substantially lower the
mincount threshold to 0.05, in order to get performance comparable to the pruned VB grammar setting. Interestingly, this uses about 7M more rules (13M more arity-1 rules, but 6M fewer arity-2 rules) than VB-Pr (1.0), but its BLEU score is marginally lower than the latter. This could be ascribed to the missing arity-2 rules, which could be crucial for certain long-distance reordering.

As the final part of the analysis, we also present the 100 high probability lexical phrases extracted by both rule extraction methods for the Arabic-English corpus in Figure 7. As seen, the heuristic grammar assigns high probability to the rules translating proper nouns and short phrases, whereas the VB method assigns high probability to more generic translations.

5 Related Research

Most of the research on learning Hiero SCFG rules has been focussed on inducing phrasal alignments between source and target using Bayesian models (Blunsom et al., 2008; Blunsom et al., 2009; Levenberg et al., 2012; Cohn and Haffari, 2013). Broadly speaking, these generative approaches learn a posterior over parallel tree structures on the sentence pairs. While these methods extract hierarchical rules, they do not conform to Hiero-style rules. Consequently the hierarchical rules are used only for learning an alignment model and cannot be used directly in the Hiero decoder. Instead, these approaches employ the standard Hiero heuristics to extract rules to be used by the decoder from the alignments predicted by their model. In this sense, these are similar to Bayesian models for learning alignments using stochastic Inversion transduction grammars (ITG) (Wu, 1997) or linear...
This paper introduced a novel Bayesian model for learning Hiero SCFG translation rules which is an alternative to the commonly used heuristic rule extraction approach. For inference, we use Variational-EM along with a Map-Reduce style framework for distributing the training process. This allowed us to efficiently train the model for very large corpuses. We provided quantitative results and also a detailed qualitative analysis to demonstrate the superiority of the model trained by our approach. In future work, we would like to extend our model for inference directly on full sentence pairs as has been applied for the syntax-based model (Galley et al., 2006).
References


