Bayesian Iterative-cascade Framework for Hierarchical Phrase-based Translation

Baskaran Sankaran first-name@cs.sfu.ca
Anoop Sarkar first-name@cs.sfu.ca
School of Computing Science, Simon Fraser University, Burnaby, BC. V5A 1S6. Canada

Abstract

The typical training of a hierarchical phrase-based machine translation involves a pipeline of multiple steps where mistakes in early steps of the pipeline are propagated without any scope for rectifying them. Additionally the alignments are trained independent of and without being informed of the end goal and hence are not optimized for translation. We introduce a novel Bayesian *iterative-cascade framework* for training Hiero-style model that learns the alignments together with the synchronous translation grammar in an iterative setting. Our framework addresses the above mentioned issues and provides an elegant and principled alternative to the existing training pipeline. Based on the validation experiments involving two language pairs, our proposed iterative-cascade framework shows consistent gains over the traditional training pipeline for hierarchical translation.

1 Introduction

Hierarchical phrase-based translation, similar to other statistical machine translation (SMT) models are trained in a series of steps that are disparate and often invoke heuristics. The training complexity as well as the modelling deficiencies in learning the translation rules using such multi-step, heuristic ridden pipeline have been documented in many previous publications (Burkett et al., 2010; DeNero and Klein, 2010; Saers et al., 2013a).

Secondly the early steps in the training pipeline, are unaware of and are almost always at odds with, the final goal of training a translation model. As a specific example, the alignment models are trained early in the pipeline, isolated from the step that extracts translations and this could lead to sub-optimal alignments (DeNero and Klein, 2010). This is also true for the syntactic models that rely on word alignments to extract the translation rules that are consistent with those alignments (Galley et al., 2006; Chiang, 2007, *inter alia*).

Consider the example word-aligned phrase pair shown in Figure 1. The baseline Giza++ alignment incorrectly aligns the English *the* to the Chinese word 联合国 (united nations). While the aligner is not going to be perfect, the present serial pipeline does not allow the aligner to *correct* such mistakes or to *adapt* the alignments to yield better translation rules.

Further the serial pipeline results in the propagation of the modelling deficiencies from the

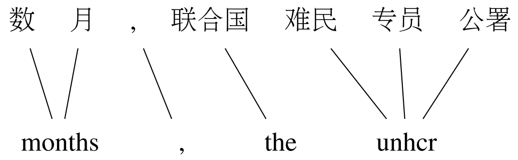

Figure 1: Example Chinese-English *phrase-pair* with word alignments. The aligner incorrectly aligns the Chinese source word 联合国 to *the* in the target side.

early steps to the latter steps. For the above example, the rule extractor of the original Hiero is likely to learn several translation rules some of which are shown in Figure 2. Some of these rules marked by an asterisk (*) encode the incorrect alignment and will lead to patently wrong translations, when applied in a slightly different context.

$^*X \rightarrow \langle$联合国 ||| the$\rangle$
$^*X \rightarrow \langle$数 月 , 联合国 ||| months , the$\rangle$
$^*X \rightarrow \langle$联合国 X_1 ||| the $X_1\rangle$
$X \rightarrow \langle$数 月 , X_1 ||| months , $X_1\rangle$
$X \rightarrow \langle$联合国 难民 专员 公署 ||| the unhcr$\rangle$
$X \rightarrow \langle X_1$ 联合国 X_2 ||| X_1 the $X_2\rangle$
$X \rightarrow \langle X_1$, X_2 难民 专员 公署 ||| X_1 , X_2 unhcr$\rangle$

Figure 2: Some extracted translation rules for the phrase-pair in Figure 1. Rules marked * will lead to wrong translations when used in other contexts.

We present a novel *iterative-cascade* framework as a way to address these issues in the training pipeline of Hiero-style systems. Our framework reduces the disparate multi-step pipeline with a simple two-step cascade model embedded in an iterative setting that allows the individual steps to improve based on some feedback from the other.

2 Iterative-cascade Framework

We now explain the intuitive idea behind our framework. The key idea of the framework is to separate the inference of alignments and hierarchical translation grammar in two successive steps and then enclose the two steps in an iterative setup. Given the dissimilarity between the alignments and SCFG rules this separation makes it easier for the models to handle the two structures at different steps. Thus the first phase reasons over the sentence pairs to find overlapping alignments, yielding a segmentation for sentence pairs, as *phrase-pairs*. Subsequently the second phase, searches over the space of derivations (of the phrase-pairs) in order to learn the optimal ones leading to better grammar.

The framework consists of two steps, viz. i) generating phrase alignments of different granularities and ii) extracting SCFG rules that are consistent with the alignments. The two phases of the iterative-cascade framework are then repeated in an iterative setup. While we could possibly come up with a single model to achieve this, we intend to validate our framework in this work using a simpler approach. We do this by using existing Bayesian models for each step in this paper.

We use the Bayesian hierarchical ITG alignment model (Neubig et al., 2011) for getting the phrasal alignments at the first step. For the phrase extraction step, we use the Bayesian model motivated by a lexical alignment prior employing Variational-Bayesian inference proposed by Sankaran et al. (2013), which operates on the extracted phrasal alignments in the earlier step. We now explain the two models briefly.

2.1 Alignments

The joint model proposed by (Neubig et al., 2011) uses a phrasal-ITG based hierarchical model with a Pitman-Yor Process (PYP) prior. Unlike the earlier models (DeNero et al., 2008; Zhang et al., 2008) that extract minimal many-to-many phrase alignments, Neubig's model extracts phrases of varying granularities. This is achieved by inverting the order to first generate the entire sentence/ phrase pair from a phrase distribution (P_t) and then falls back to ITG derivation to divide the sentence/ phrase pair into shorter phrase-pairs (this effectively avoids the sparsity problem).

Under this model each phrase pair gets some probability distribution $P_{hier}(\langle e, f\rangle : \theta_x, \theta_t)$, where θ_x and θ_t are the parameters of symbol distribution and phrase table respectively. The phrase table parameters θ_t are given by a PYP prior as

$$\theta_t \sim \text{PYP}(\mathsf{d}, \mathsf{s}, P_{dac}) \tag{1}$$

where, d and s are discount and strength parameters. The base measure P_{dac} adopts a "divide-and-conquer" strategy of recursively breaking up a longer phrase-pair into two shorter phrases through an ITG derivation. The entire generative process for begins from the full sentence pair (say s) and follows the script given below.

1. Generate the entire phrase-pair s from the phrase-table distribution P_t. Now fall back to break the phrase-pair through ITG-style derivations employing P_{dac}
2. Decide the ITG derivation type I_d from symbol distribution θ_x which can be BASE, REG or INV
   (a) If I_d = BASE, directly generate a new terminal phrase-pair from P_{base}, based on IBM Model 1 word alignment probabilities, defined similar to (DeNero et al., 2008)
   (b) If I_d = REG, recursively generate smaller biphrases $\langle e_1, f_1\rangle$ and $\langle e_2, f_2\rangle$ from P_{hier} and concatenate them as $\langle e_1 e_2, f_1 f_2\rangle$
   (c) If I_d = INV, recursively generate smaller biphrases $\langle e_1, f_1\rangle$ and $\langle e_2, f_2\rangle$ from P_{hier} and concatenate them as $\langle e_1 e_2, f_2 f_1\rangle$

For inference it uses a sentence-level block sampler exploring the space of ITG-phrase alignments. In order to reduce the time complexity in sampling, it uses a heuristic beam search approximation that prunes the alignment spans based on a probability threshold (see Neubig et al. (2011) for details).

2.2 Grammar Extraction

After extracting the phrasal alignments in the previous step, we now need to learn hierarchical translation grammar along with the rule parameters. We have chosen the Bayesian model proposed by (Sankaran et al., 2013) for this step because, unlike other Bayesian models, their model can infer a Hiero-style grammar[1] that can be used directly by a hierarchical phrase-based decoder.

Their model assumes the existence of *initial* phrase-pairs obtained from bidirectional symmetrization of word alignments (traditional SMT training pipeline). In our case these are obtained by the earlier step. The model generates an aligned phrase pair x from the hierarchical translation rules using the following two-step generative story.

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,[2] i.e. $z_d = \{\textsc{Term}, \textsc{Hier-A}1, \textsc{Hier-A}2\}$. Following Chiang (2007), we allow a maximum of two gaps or two non-terminals in the SCFG because the hierarchical phrase-based decoder becomes prohibitively computationally expensive with more than two non-terminals.
2. Identify the constituent rules $\mathbf{r}$ in the derivation to generate the phrase pair.

$\phi^z \sim \text{Dirichlet}(\boldsymbol{\alpha_z})$	[draw derivation type parameters]
$\theta \sim \text{Dirichlet}(\alpha_h p_0)$	[draw rule parameters]
$z_d \sim \text{Multinomial}(\phi^z)$	[decide the derivation type]
$r \vert \mathbf{r} \in d_x \sim \text{Multinomial}(\theta)$	[generate rules deriving phrase-pair x]

Figure 3: SCFG Extraction model: Definition

Under this model the probability of a particular derivation $d \in \phi_x$ for a given phrase pair x will be given by:

$$p(d) \propto p(z_d) \prod_{r \in d} p(r|\mathcal{G}, \theta) \qquad (2)$$

where r is a rule in grammar $\mathcal{G}$ and θ is the grammar parameter.

Figure 3 depicts the generative story of this model, while its corresponding graphical representation is shown in Figure 4. The derivation-type z_d is sampled from a multinomial distribution parameterized by ϕ^z, where ϕ^z is distributed itself by a Dirichlet distribution with hyper-parameter $\boldsymbol{\alpha_z}$. The grammar rules are generated from a multinomial distribution parameterized by θ, where θ itself is distributed according to a Dirichlet distribution parameterized by a concentration parameter α_h and a base distribution p_0. For the base distribution, we again follow Sankaran et al. (2013) and use an informative prior based on geometric mean of the bidirectional alignment scores. This ensures that the model only considers the derivations that are

[1] As we mentioned earlier most of the other Bayesian models are merely alignment models and employ an additional heuristic step for extracting Hiero-style grammar.

[2] This refers to the maximum arity of a rule involved in the derivation.

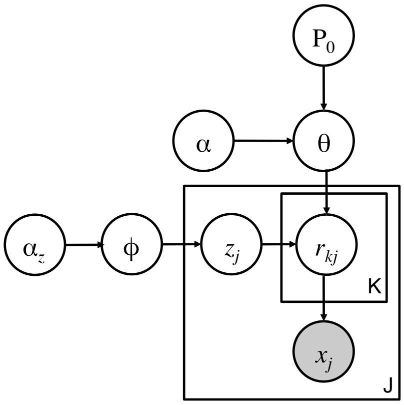

Figure 4: Graphical model depicting SCFG rule extraction in phase-2. The generative process first decides the derivation type z_j from a Multinomial parametrized by ϕ. It then generates the rules r_{kj} in the derivation by using a Dirichlet distribution θ with base measure P_0 and concentration parameter α. There are K rules in the derivation which yields the phrase-pair x_j.

consistent with the underlying word alignments.[3] This setting closely mimics the Hiero heuristic extraction approach (Chiang, 2007), which constrains the rule extraction to be consistent with the alignments.

Our goal in this phase is to infer the joint posterior $p(\theta, \Phi | \alpha_h, p_0, \boldsymbol{\alpha_z}, \mathcal{X})$, where θ are the model parameters and Φ the latent derivations over all the phrase pairs. This could be factorized by using Variational approximation, yielding the posterior distributions θ (over grammar parameters) and Φ (over latent derivations).

$$p(\theta, \Phi | \alpha_h, p_0, \boldsymbol{\alpha_z}, \mathcal{X}) \approx q(\theta | \mathbf{u}) q(\Phi | \pi)$$

where $\mathbf{u}$ and π are the parameters of the variational distributions.

The inference is then performed in an EM-style algorithm, similar to Sankaran et al. (2013) - by iteratively updating the parameters $\mathbf{u}$ and π. We initialize $\mathbf{u}^0 := \alpha_h p_0$, 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:

$$\mathbb{E}[r^t] = \sum_{d \in \phi_x} p(d | \pi^{t-1}, x) f_d(r) \tag{3}$$

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 π as:

$$p(d | \pi^{t-1}, x) \propto p(z_d) \prod_{r \in d} \pi_r^{t-1} \tag{4}$$

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

[3] While a non-parametric prior would be better from a Bayesian perspective, we leave it for future consideration.

the derivations of its type.

$$\mathbb{E}[z_d^t] = \sum_x \sum_{\{z_d = z_{d'} | d' \in \phi_x\}} p(d'|\pi^{t-1}, x) \tag{5}$$

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

2.3 Iterative Cascade Framework

The formulation of the framework with independent modules allow us to easily experiment with existing models for alignment and SCFG rule extraction. This also helps us quickly validate the effectiveness of our framework.

The many-to-many alignments extracted by pialign is directly fed to the rule extraction model in phase-2 for extracting the Hiero-style grammar. In the reverse direction, we could parse the sentences in the training set with the extracted Hiero-style grammar and use the resulting alignments to initialize the aligner in the next iteration. However, we decided to use a simple setting for this paper; hence we iterate the two steps of the iterative framework without the feedback in the reverse direction. Using the existing models as described above, we run the *iterative* cascade framework as below.

1. Run the alignment model described in Section 2.1 for a fixed burn-in iterations (set to 9) and collect alignment samples from the next iteration. (Phase-1)
2. Recompute lexical probabilities based on the current alignments for computing the prior in the phase-2.
3. Extract hierarchical translation grammar from the phrasal alignments that were obtained in phase-1 and using the Variational-Bayesian inference explained in Section 2.2. We run the VB inference for a fixed number (set to 10) of iterations. (Phase-2)
4. Repeat steps 1 through 3 for a small number of times (we use 3 runs) and at each iteration collect the samples independently.

Figures 5 and 6 depict the alignments and the extracted rules at the end of first and second iterations of the cascade framework. Each figure consists of three parts i) word alignments on the left, ii) alignment matrix in the middle and iii) the extracted rules on the right. It can be seen that the incorrect word alignment (marked in red in Figure 5) is correctly aligned at the end of second iteration in Figure 6.

At the end of 3 iterations of the cascade framework, we do model combination to aggregate the resulting Hiero-style grammars. The parameters for the rules are then estimated using relative frequency estimation as is done in the original Hiero rule extraction method.

[4] In our experiments, we set the number of iterations to 10.

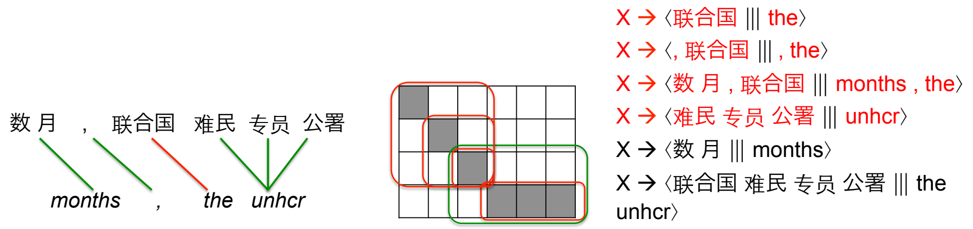
数 月 , 联合国 难民 专员 公署
months , the unhcr
X → 〈联合国 ||| the〉
X → 〈, 联合国 ||| , the〉
X → 〈数 月 , 联合国 ||| months , the〉
X → 〈难民 专员 公署 ||| unhcr〉
X → 〈数 月 ||| months〉
X → 〈联合国 难民 专员 公署 ||| the unhcr〉

Figure 5: Alignments and SCFG rules at the end of first iteration of *iterative-cascade* framework

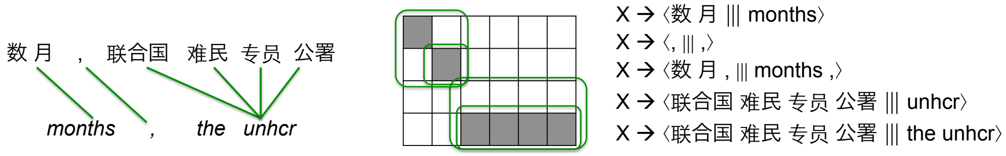
数 月 , 联合国 难民 专员 公署
months , the unhcr
X → 〈数 月 ||| months〉
X → 〈, ||| ,〉
X → 〈数 月 , ||| months ,〉
X → 〈联合国 难民 专员 公署 ||| unhcr〉
X → 〈联合国 难民 专员 公署 ||| the unhcr〉

Figure 6: Alignments and SCFG rules at the end of second iteration of *iterative-cascade* framework

3 Experiments

We evaluate our iterative-cascade framework on two language pairs: Korean-English and Arabic-English. In both language pairs we limit the sentence length of the training set to 60 in order to run the Gibbs sampler in phase-1 efficiently (mainly due to the limitation of the Gibbs sampler employed by the aligner).

We use the Rochester corpus for Korean-English and remove the sentences longer than 60 words, resulting in about 52K sentence pairs for training. We retain 1118 sentence pairs each for tuning and testing. For Arabic-English, we randomly sample the ISI parallel-corpus to select 120K sentence pairs that satisfy the sentence length criterion. For tuning and testing, we use 1982 and 987 sentence pairs from the same corpus. The statistics of the two corpora are shown in Table 1.

Lang.	Corpus	Train (# of words)	Dev/ test (# of sents)
Korean-English	University of Rochester	1.5M/ 1.4M	1118/ 1118
Arabic-English	ISI web-crawled	3.1M/ 3.3M	1982/ 987

Table 1: Hiero-style binary grammar extraction: Corpus statistics for iterative-cascade experiments. The sentences are restricted to have at most 60 words due to the limitation of the aligner.

We use our implementation of conventional Hiero system for training the baseline models and our CKY-style hierarchical phrase-based decoder for decoding in all our experiments. We use the respective systems for the two steps of our cascade-framework, i.e. pialign (Neubig et al., 2011) and Variational inference models. For experiments involving pialign, we ran the aligner for 10 iterations with 9 burn-in iterations. The samples were read off from the last iteration. For extracting SCFG grammar we use the initial phrase-pairs obtained by pialign and pass them through either the heuristic (Chiang, 2007) or the Variational-Bayesian (Section 2.2) extractor. We tuned the feature weights using MERT and decoded the test set with the optimal weights. For language model, we use a 5-gram model trained on the gigaword corpus with Kneser-Ney smoothing using the SRILM toolkit. For the iterative-cascade setting, we iterate the two steps of the framework for three runs and do a sample combination to get the final grammar.

Aligner	**Extractor**	**BLEU**
Giza++	Heuristic	7.97
Giza++	Var. Bayes	8.03
Pialign	Heuristic	7.70
Pialign	Var. Bayes	*7.54*
Iterative-cascade (3 iters)		**8.19**

Table 2: Iterative-cascade framework: Korean-English BLEU scores. For the iterative-cascade framework we ran Pialign and VB inferences for three iterations and did a sample combination. The BLEU scores that are less than the baseline Moses (Giza++, Heuristic) BLEU of 8.23 by a statistically significant margin are *italicized*. The best BLEU score is in **boldface**.

The results of the iterative-cascade inference are summarized in Tables 2 and 3 for Korean-English and Arabic-English settings respectively. We use four baseline hierarchical translation systems that arise from different combinations of the aligner and extractor as listed in the tables.

The first two baselines use Giza++ aligner and then use the two different (heuristic and VB) methods for extracting translation rules, which are then tuned/ tested with our CKY-style decoder. Baselines 3 and 4 differ from the earlier ones in that, these baselines use pialign to generate many-to-many alignments. The last row corresponds to the iterative-cascade grammar setting, where we run the iterative inference three times and then aggregate the grammars.

In both language pairs, baselines employing pialign perform marginally worse and the first iteration of iterative-cascade model in fact results in statistically significant BLEU reduction compared to phrase-based baseline of 8.23. However when we run our cascade framework for three iterations, we see consistent BLEU score improvements ranging between 0.2 and 0.65 as compared to other baselines in the table.

One can also compare these scores to the phrase-based model for the sake of completeness. We consider two phrase-based models one using the regular heuristic training pipeline as Koehn et al. (2003) and the other using pialign. For pialign, we use the phrase table extracted by pialign and directly used it with Moses for tuning and decoding. Note that this baseline uses two additional features including span probability (see Neubig et al. (2011)) that are not used in the standard baseline or in the later models in the tables. The two phrase-based models obtained BLEU scores of 8.23 and 8.30 respectively and these are comparable to the performance of our iterative-cascade model.

Now turning our attention to the Arabic-English language pair we again notice a very similar behaviour as we saw for Korean-English. The only difference is that the scale of improvement is marginally less and our iterative-cascade framework improves the BLEU scores in the range of 0.25 and 0.5 over the other baselines. A phrase-based model using Moses achieves 25.34 BLEU score, while the pialign achieves 24.90.

Aligner	Extractor	BLEU
Giza++	Heuristic	25.13
Giza++	Var. Bayes	25.20
Pialign	Heuristic	24.97
Pialign	Var. Bayes	25.09
Iterative-cascade (3 iters)		**25.45**

Table 3: Iterative-cascade framework: Arabic-English BLEU scores. For the iterative-cascade framework we ran Pialign and VB inferences independently for three runs and did a sample combination. The best BLEU score is in **boldface**.

4 Related work

On models learning Alignments or Hiero-style grammar: The potential incompatibility between the word alignments and the translation rules for the syntactic translation models have been noticed earlier (DeNero and Klein, 2007). Apart from showing the incompatibility, they also propose an unsupervised HMM alignment model that soft constrains the alignments conditioned on the target sentences and the corresponding (automatically generated) parse trees. The main difference is that our approach seek to improve alignments of different granularities and not just the word-level alignments.

Several other works have focussed on learning phrase alignments from synchronous derivations using non-lexicalized (Blunsom et al., 2008) or lexicalized (Hiero-style) ITG (Blunsom et al., 2009; Levenberg et al., 2012) rules and apply them for hierarchical phrase-based models. While these models extract ITG-style rules, they use them only for obtaining the alignment information. In other words, the extracted ITG-style rules are not directly used by a hierarchical translation decoder, which are in fact obtained from the alignments suggested by these rules. Thus the biggest drawback is that these models, strictly speaking, are alignment models and they use the heuristic rule extractor (Chiang, 2007) for learning the Hiero-style translation grammar.

In contrast to these, Sankaran et al. (2012, 2013) proposed a set of Bayesian models that directly learns the SCFG grammar. However these models only focus on the rule extraction part and rely on the heuristically extracted phrasal alignments. Instead our iterative-cascade framework simplifies the entire hierarchical translation training pipeline.

On Joint models for PBMT: Several works have exploited word alignments to improving the performance of parsing (Burkett and Klein, 2008; Snyder et al., 2009) outside the machine translation setting. In the reverse direction syntactic parsing has been used to get better alignments (May and Knight, 2007; DeNero and Klein, 2007; Fossum et al., 2008) in the context of machine translations.

Joint models for learning alignments and translation rules have been a fairly recent direction. A joint model using two syntactic parsers and combined with an ITG derivation to model alignments, enables the trees to diverge if required and otherwise encouraging the derivation to synchronize with the trees (Burkett et al., 2010). However it requires a parallel treebank and gold alignments to train on in addition to parsers for the source and target languages, thus severely limiting its applicability. DeNero and Klein (2010) proposed a supervised model for extracting all overlapping bispans called *extraction sets* under a discriminative model by using phrase-level features in addition to the one-to-one alignments.

In contrast, Neubig et al. (2011) proposed an unsupervised hierarchical ITG model for jointly learning the alignments and translations as we explained in section 2.1. The extracted translation rules are then directly used in a phrase-based decoder. While, their alignments are based on the ITG, it uses with a flat (phrase-based) model for translation. We extend their approach through our iterative-cascade framework and extract SCFG rules in a separate phase.

Iterative approach has been used for directly training the phrase translation models for a phrase-based system (Wuebker et al., 2010). This method employs force decoding the training set (based on the IBM word alignments) to obtain phrase segmentations. Phrase probabilities are then estimated using leaving-one-out technique, in order to avoid overfitting. Our present work on learning hierarchical translation model differs from this in obvious way; additionally we just use model aggregation as opposed to their iterative force decoding.

Heger et al. (2010) extend the Wuebker's work by combining the iteratively learned phase alignments with the heuristically learned hierarchical translation model for Hiero-style system. This approach is similar in spirit to our goal of learning a hierarchical translation model that is consistent throughout. However their approach does not learn the hierarchical SCFG rules through force alignments, but only combines the iteratively learned phrase table with hierarchical translation grammar extracted traditionally. Secondly our framework allow both alignments and SCFG rules to be improved iteratively unlike theirs,[5] where only the phrase alignments are improved. Another major difference is that we use Variational-Bayesian inference to aggregate rule counts globally as opposed to the leave-one-out counting.

On Joint models for (Hiero-style) ITG: Recently, Saers et al. (2013b,a,c) proposed a Bayesian maximum *a posteriori* (MAP) driven model for extracting bracketing inversion transduction grammar. This approach aims to improve coherence and model consistency between the training and test. Unlike Neubig et al. (2011), this line of work is motivated to employ same (lexicalized-ITG) model in both training and testing. While this approach is elegant, the emphasis on a single coherent model is too restrictive and is unable to integrate well with other feature functions such as better reordering or language models, while our method is able to do so.

Similar to Saers et al. (2013b,a,c), our iterative-cascade framework simplifies the training pipeline of the hierarchical phrase-based system in order to obtain novel alignments and translation rules to be used in the SMT decoder. However, their approach stands separate from conventional phrase-based and hierarchical phrase-based SMT models. In contrast, while our approach does compute new alignments and translation rules, it can also be combined with some of the recent advancements in SMT, for example in reordering model and language model.

Finally many of the previous work on Bayesian grammar induction are trained and tested on datasets that have simple and short sentences (Blunsom et al., 2008; Saers et al., 2013b,a,c). Typically they use the IWSLT Chinese-English corpus consisting of sentences in the travel domain, where the average sentence length on English side is around 7 words. On the other hand,

[5] The present approach however only does model aggregation as opposed to full iterative learning, which we leave for future research.

we use realistic datasets with fairly long (average sentence length > 27 words) and complex sentences.

5 Conclusion and Future Directions

Our *iterative-cascade* framework reduces the serial, multi-component and heuristic-ridden Hiero training pipeline with a simple two-step iterative pipeline. The simplicity of the framework further enables any appropriate model to be plugged in for the alignment and rule extraction steps. Validation experiments with existing models demonstrate small but consistent gains over the traditional Hiero training baseline involving heuristic steps for two language pairs. Further the resulting synchronous context-free grammar has a sparse distribution, where the probability mass is concentrated on few rules unlike the flat distribution of rules generated by the conventional pipeline. Unlike the earlier research on Hiero-style Bayesian grammar induction (Blunsom et al., 2008, 2009; Levenberg et al., 2012), the grammar induced by our iterative-cascade framework are directly used by the CKY decoder.

A minor shortcoming of our work relates to the smaller training corpus we use for the experiments in this paper. However, as we noted earlier our experimental results are based on *realistic* SMT datasets that contain longer and complex sentences. This is unlike some earlier approaches that rely on some corpus consisting of shorter sentences with much simpler structure, where a large number of the sentences might share the same structure due to the nature of the domain. Further, we intend to address this shortcoming in near future (see below).

As a future work, we are currently working on adding the feedback loop to improve the alignments by using the information from the hierarchical translation grammar extracted in the second step. One could use the simple approach of parsing the training corpus using the SCFG extracted in the second step of iterative-cascade and use the resulting alignments to initialize the aligner in the next iteration. We are also exploring other approaches for doing this. Secondly, we also intend to replace the current ITG aligner to avoid the sampling issues due to the approximation employed by its beam search sampler. This would enable us to run experiments on large parallel corpora for better validation of our iterative-cascade framework.

Acknowledgments

We would like to thank Gholamreza Haffari, Adam Lopez and Taro Watanabe for discussions and useful suggestions.

References

Blunsom, P., Cohn, T., Dyer, C., and Osborne, M. (2009). A gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the Annual meeting of Association of Computational Linguistics*.

Blunsom, P., Cohn, T., and Osborne, M. (2008). Bayesian synchronous grammar induction. In *Proceedings of the Neural Information Processing Systems*.

Burkett, D., Blitzer, J., and Klein, D. (2010). Joint parsing and alignment with weakly synchronized grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter* *of the Association for Computational Linguistics*, pages 127–135. Association for Computational Linguistics.

Burkett, D. and Klein, D. (2008). Two languages are better than one (for syntactic parsing). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 877–886, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33.

DeNero, J., Bouchard-Cote, A., and Dan, K. (2008). Sampling alignment structure under a bayesian translation model. In *Proceedings of Empirical Methods in Natural Language Processing-08*, pages 314–323. Association for Computational Linguistics.

DeNero, J. and Klein, D. (2007). Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24, Prague, Czech Republic. Association for Computational Linguistics.

DeNero, J. and Klein, D. (2010). Discriminative modeling of extraction sets for machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1453–1463. Association for Computational Linguistics.

Fossum, V., Knight, K., and Abney, S. (2008). Using syntax to improve word alignment precision for syntax-based machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 44–52. Association for Computational Linguistics.

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., and Thayer, I. (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Heger, C., Wuebker, J., Vilar, D., and Ney, H. (2010). A combination of hierarchical systems with forced alignments from phrase-based systems. In *International Workshop on Spoken Language Translation, IWSLT*, pages 291–297.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Levenberg, A., Dyer, C., and Blunsom, P. (2012). A bayesian model for learning scfgs with discontiguous rules. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 223–232, Jeju Island, Korea. Association for Computational Linguistics.

May, J. and Knight, K. (2007). Syntactic re-alignment models for machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 360–368, Prague, Czech Republic. Association for Computational Linguistics.

Neubig, G., Watanabe, T., Sumita, E., Mori, S., and Kawahara, T. (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. Association for Computational Linguistics.

Saers, M., Addanki, K., and Wu, D. (2013a). Combining top-down and bottom-up search for unsupervised induction of transduction grammars. In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 48–57, Atlanta, Georgia. Association for Computational Linguistics.

Saers, M., Addanki, K., and Wu, D. (2013b). Iterative rule segmentation under minimum description length for unsupervised transduction grammar induction. In Dediu, A.-H., Martín-Vide, C., Mitkov, R., and Truthe, B., editors, *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 224–235. Springer Berlin Heidelberg.

Saers, M., Addanki, K., and Wu, D. (2013c). Unsupervised transduction grammar induction via minimum description length. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 67–73, Sofia, Bulgaria. Association for Computational Linguistics.

Sankaran, B., Haffari, G., and Sarkar, A. (2012). Compact rule extraction for hierarchical phrase-based translation. In *The 10th biennial conference of the Association for Machine Translation in the Americas (AMTA)*, San Diego, CA. Association for Computational Linguistics.

Sankaran, B., Haffari, G., and Sarkar, A. (2013). Scalable variational inference for extracting hierarchical phrase-based translation rules. In *Submitted to the 6th International Joint Conference on Natural Language Processing*, Nagoya, Japan. Association for Computational Linguistics.

Snyder, B., Naseem, T., and Barzilay, R. (2009). Unsupervised multilingual 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: Volume 1 - Volume 1*, ACL '09, pages 73–81, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wuebker, J., Mauser, A., and Ney, H. (2010). Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 475–484, Stroudsburg, PA, USA. Association for Computational Linguistics.

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