An Unsupervised Model for Joint Phrase Alignment and Extraction

Graham Neubig[1,2] **Taro Watanabe**[2], **Eiichiro Sumita**[2], **Shinsuke Mori**[1], **Tatsuya Kawahara**[1]
[1]Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto, Japan
[2]National Institute of Information and Communication Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, Japan

Abstract

We present an unsupervised model for joint phrase alignment and extraction using nonparametric Bayesian methods and inversion transduction grammars (ITGs). The key contribution is that phrases of many granularities are included directly in the model through the use of a novel formulation that memorizes phrases generated not only by terminal, but also non-terminal symbols. This allows for a completely probabilistic model that is able to create a phrase table that achieves competitive accuracy on phrase-based machine translation tasks directly from unaligned sentence pairs. Experiments on several language pairs demonstrate that the proposed model matches the accuracy of traditional two-step word alignment/phrase extraction approach while reducing the phrase table to a fraction of the original size.

1 Introduction

The training of translation models for phrase-based statistical machine translation (SMT) systems (Koehn et al., 2003) takes unaligned bilingual training data as input, and outputs a scored table of phrase pairs. This phrase table is traditionally generated by going through a pipeline of two steps, first generating word (or minimal phrase) alignments, then extracting a phrase table that is consistent with these alignments.

However, as DeNero and Klein (2010) note, this two step approach results in word alignments that are not optimal for the final task of generating phrase tables that are used in translation. As a solution to this, they proposed a supervised discriminative model that performs joint word alignment and phrase extraction, and found that joint estimation of word alignments and extraction sets improves both word alignment accuracy and translation results.

In this paper, we propose the first *unsupervised* approach to joint alignment and extraction of phrases at multiple granularities. This is achieved by constructing a generative model that includes phrases at many levels of granularity, from minimal phrases all the way up to full sentences. The model is similar to previously proposed phrase alignment models based on inversion transduction grammars (ITGs) (Cherry and Lin, 2007; Zhang et al., 2008; Blunsom et al., 2009), with one important change: ITG symbols and phrase pairs are generated in the opposite order. In traditional ITG models, the branches of a biparse tree are generated from a nonterminal distribution, and each leaf is generated by a word or phrase pair distribution. As a result, only minimal phrases are directly included in the model, while larger phrases must be generated by heuristic extraction methods. In the proposed model, at each branch in the tree, we first attempt to generate a phrase pair from the phrase pair distribution, falling back to ITG-based divide and conquer strategy to generate phrase pairs that do not exist (or are given low probability) in the phrase distribution.

We combine this model with the Bayesian nonparametric Pitman-Yor process (Pitman and Yor, 1997; Teh, 2006), realizing ITG-based divide and conquer through a novel formulation where the Pitman-Yor process uses two copies of itself as a base measure. As a result of this modeling strategy, phrases of multiple granularities are generated, and thus memorized, by the Pitman-Yor process. This makes it possible to directly use probabilities of the phrase model as a replacement for the phrase table generated by heuristic extraction techniques.

Using this model, we perform machine translation experiments over four language pairs. We observe that the proposed joint phrase alignment and extraction approach is able to meet or exceed results attained by a combination of GIZA++ and heuristic phrase extraction with significantly smaller phrase table size. We also find that it achieves superior BLEU scores over previously proposed ITG-based phrase alignment approaches.

2 A Probabilistic Model for Phrase Table Extraction

The problem of SMT can be defined as finding the most probable target sentence $\mathbf{e}$ for the source sentence $\mathbf{f}$ given a parallel training corpus $\langle \mathcal{E}, \mathcal{F} \rangle$

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\text{argmax}} P(\mathbf{e}|\mathbf{f}, \langle \mathcal{E}, \mathcal{F} \rangle).$$

We assume that there is a hidden set of parameters θ learned from the training data, and that $\mathbf{e}$ is conditionally independent from the training corpus given θ. We take a Bayesian approach, integrating over all possible values of the hidden parameters:

$$P(\mathbf{e}|\mathbf{f}, \langle \mathcal{E}, \mathcal{F} \rangle) = \int_{\theta} P(\mathbf{e}|\mathbf{f}, \theta) P(\theta|\langle \mathcal{E}, \mathcal{F} \rangle). \quad (1)$$

If θ takes the form of a scored phrase table, we can use traditional methods for phrase-based SMT to find $P(\mathbf{e}|\mathbf{f}, \theta)$ and concentrate on creating a model for $P(\theta|\langle \mathcal{E}, \mathcal{F} \rangle)$. We decompose this posterior probability using Bayes law into the corpus likelihood and parameter prior probabilities

$$P(\theta|\langle \mathcal{E}, \mathcal{F} \rangle) \propto P(\langle \mathcal{E}, \mathcal{F} \rangle|\theta) P(\theta).$$

In Section 3 we describe an existing method, and in Section 4 we describe our proposed method for modeling these two probabilities.

3 Flat ITG Model

There has been a significant amount of work in many-to-many alignment techniques (Marcu and Wong (2002), DeNero et al. (2008), *inter alia*), and in particular a number of recent works (Cherry and Lin, 2007; Zhang et al., 2008; Blunsom et al., 2009) have used the formalism of inversion transduction grammars (ITGs) (Wu, 1997) to learn phrase alignments. By slightly limit reordering of words, ITGs make it possible to exactly calculate probabilities of phrasal alignments in polynomial time, which is a computationally hard problem when arbitrary reordering is allowed (DeNero and Klein, 2008).

The traditional flat ITG generative probability for a particular phrase (or sentence) pair $P_{flat}(\langle e, f \rangle; \theta_x, \theta_t)$ is parameterized by a phrase table θ_t and a symbol distribution θ_x. We use the following generative story as a representative of the flat ITG model.

1. Generate symbol x from the multinomial distribution $P_x(x; \theta_x)$. x can take the values TERM, REG, or INV.

2. According to the x take the following actions.

   (a) If $x =$ TERM, generate a phrase pair from the phrase table $P_t(\langle e, f \rangle; \theta_t)$.

   (b) If $x =$ REG, a regular ITG rule, generate phrase pairs $\langle e_1, f_1 \rangle$ and $\langle e_2, f_2 \rangle$ from P_{flat}, and concatenate them into a single phrase pair $\langle e_1 e_2, f_1 f_2 \rangle$.

   (c) If $x =$ INV, an inverted ITG rule, follows the same process as (b), but concatenate f_1 and f_2 in reverse order $\langle e_1 e_2, f_2 f_1 \rangle$.

By taking the product of P_{flat} over every sentence in the corpus, we are able to calculate the likelihood

$$P(\langle \mathcal{E}, \mathcal{F} \rangle|\theta) = \prod_{\langle e,f \rangle \in \langle \mathcal{E}, \mathcal{F} \rangle} P_{flat}(\langle e, f \rangle; \theta).$$

We will refer to this model as FLAT.

3.1 Bayesian Modeling

While the previous formulation can be used as-is in maximum likelihood training, this leads to a degenerate solution where every sentence is memorized as a single phrase pair. Zhang et al. (2008) and others propose dealing with this problem by putting a prior probability $P(\theta_x, \theta_t)$ on the parameters.

We assign θ_x a Dirichlet prior[1], and assign the phrase table parameters θ_t a prior using the Pitman-Yor process (Pitman and Yor, 1997; Teh, 2006), which is a generalization of the Dirichlet process prior used in previous research. It is expressed as

$$\theta_t \sim PY(d, s, P_{base}) \tag{2}$$

where d is the discount parameter, s is the strength parameter, and P_{base} is the base measure. The discount d is subtracted from observed counts, and when it is given a large value (close to one), less frequent phrase pairs will be given lower relative probability than more common phrase pairs. The strength s controls the overall sparseness of the distribution, and when it is given a small value the distribution will be sparse. P_{base} is the prior probability of generating a particular phrase pair, which we describe in more detail in the following section.

Non-parametric priors are well suited for modeling the phrase distribution because every time a phrase is generated by the model, it is "memorized" and given higher probability. Because of this, common phrase pairs are more likely to be re-used (the *rich-get-richer* effect), which results in the induction of phrase tables with fewer, but more helpful phrases. It is important to note that only phrases generated by P_t are actually memorized and given higher probability by the model. In FLAT, only minimal phrases generated after P_x outputs the terminal symbol TERM are generated from P_t, and thus only minimal phrases are memorized by the model.

While the Dirichlet process is simply the Pitman-Yor process with $d = 0$, it has been shown that the discount parameter allows for more effective modeling of the long-tailed distributions that are often found in natural language (Teh, 2006). We confirmed in preliminary experiments (using the data described in Section 7) that the Pitman-Yor process with automatically adjusted parameters results in superior alignment results, outperforming the sparse Dirichlet process priors used in previous research[2]. The average gain across all data sets was approximately 0.8 BLEU points.

3.2 Base Measure

P_{base} in Equation (2) indicates the prior probability of phrase pairs according to the model. By choosing this probability appropriately, we can incorporate prior knowledge of what phrases tend to be aligned to each other. We calculate P_{base} by first choosing whether to generate an unaligned phrase pair (where $|e| = 0$ or $|f| = 0$) according to a fixed probability p_u[3], then generating from P_{ba} for aligned phrase pairs, or P_{bu} for unaligned phrase pairs.

For P_{ba}, we adopt a base measure similar to that used by DeNero et al. (2008):

$$\begin{aligned} P_{ba}(\langle e,f \rangle) =& M_0(\langle e,f \rangle) P_{pois}(|e|;\lambda) P_{pois}(|f|;\lambda) \\ M_0(\langle e,f \rangle) =& (P_{m1}(f|e) P_{uni}(e) P_{m1}(e|f) P_{uni}(f))^{\frac{1}{2}}. \end{aligned}$$

P_{pois} is the Poisson distribution with the average length parameter λ. As long phrases lead to sparsity, we set λ to a relatively small value to allow us to bias against overly long phrases[4]. P_{m1} is the word-based Model 1 (Brown et al., 1993) probability of one phrase given the other, which incorporates word-based alignment information as prior knowledge in the phrase translation probability. We take the geometric mean[5]of the Model 1 probabilities in both directions to encourage alignments that are supported by both models (Liang et al., 2006). It should be noted that while Model 1 probabilities are used, they are only soft constraints, compared with the hard constraint of choosing a single word alignment used in most previous phrase extraction approaches.

For P_{bu}, if g is the non-null phrase in e and f, we calculate the probability as follows:

$$P_{bu}(\langle e,f \rangle) = P_{uni}(g) P_{pois}(|g|;\lambda)/2.$$

Note that P_{bu} is divided by 2 as the probability is considering null alignments in both directions.

4 Hierarchical ITG Model

While in FLAT only minimal phrases were memorized by the model, as DeNero et al. (2008) note

[1] The value of α had little effect on the results, so we arbitrarily set $\alpha = 1$.

[2] We put weak priors on s ($Gamma(\alpha = 2, \beta = 1)$) and d ($Beta(\alpha = 2, \beta = 2)$) for the Pitman-Yor process, and set $\alpha = 1^{-10}$ for the Dirichlet process.

[3] We choose 10^{-2}, 10^{-3}, or 10^{-10} based on which value gave the best accuracy on the development set.

[4] We tune λ to 1, 0.1, or 0.01 based on which value gives the best performance on the development set.

[5] The probabilities of the geometric mean do not add to one, but we found empirically that even when left unnormalized, this provided much better results than the using the arithmetic mean, which is more theoretically correct.

and we confirm in the experiments in Section 7, using only minimal phrases leads to inferior translation results for phrase-based SMT. Because of this, previous research has combined FLAT with heuristic phrase extraction, which exhaustively combines all adjacent phrases permitted by the word alignments (Och et al., 1999). We propose an alternative, fully statistical approach that directly models phrases at multiple granularities, which we will refer to as HIER. By doing so, we are able to do away with heuristic phrase extraction, creating a fully probabilistic model for phrase probabilities that still yields competitive results.

Similarly to FLAT, HIER assigns a probability $P_{hier}(\langle e, f\rangle; \theta_x, \theta_t)$ to phrase pairs, and is parameterized by a phrase table θ_t and a symbol distribution θ_x. The main difference from the generative story of the traditional ITG model is that symbols and phrase pairs are generated in the opposite order. While FLAT first generates branches of the derivation tree using P_x, then generates leaves using the phrase distribution P_t, HIER first attempts to generate the full sentence as a single phrase from P_t, then falls back to ITG-style derivations to cope with sparsity. We allow for this within the Bayesian ITG context by defining a new base measure P_{dac} ("divide-and-conquer") to replace P_{base} in Equation (2), resulting in the following distribution for θ_t.

$$\theta_t \sim PY(d, s, P_{dac}) \qquad (3)$$

P_{dac} essentially breaks the generation of a single longer phrase into two generations of shorter phrases, allowing even phrase pairs for which $c(\langle e, f\rangle) = 0$ to be given some probability. The generative process of P_{dac}, similar to that of P_{flat} from the previous section, is as follows:

1. Generate symbol x from $P_x(x; \theta_x)$. x can take the values BASE, REG, or INV.

2. According to x take the following actions.

   (a) If x = BASE, generate a new phrase pair directly from P_{base} of Section 3.2.

   (b) If x = REG, generate $\langle e_1, f_1\rangle$ and $\langle e_2, f_2\rangle$ from P_{hier}, and concatenate them into a single phrase pair $\langle e_1 e_2, f_1 f_2\rangle$.

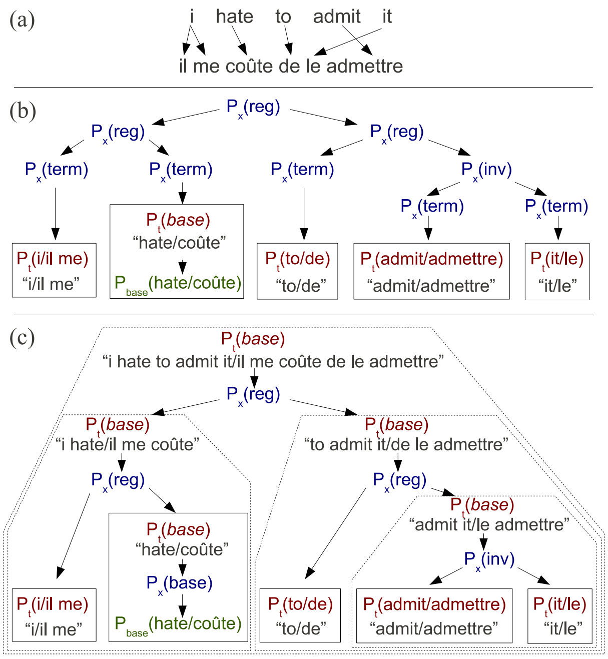

Figure 1: A word alignment (a), and its derivations according to FLAT (b), and HIER (c). Solid and dotted lines indicate minimal and non-minimal pairs respectively, and phrases are written under their corresponding instance of P_t. The pair hate/coûte is generated from P_{base}.

   (c) If x = INV, follow the same process as (b), but concatenate f_1 and f_2 in reverse order $\langle e_1 e_2, f_2 f_1\rangle$.

A comparison of derivation trees for FLAT and HIER is shown in Figure 1. As previously described, FLAT first generates from the symbol distribution P_x, then from the phrase distribution P_t, while HIER generates directly from P_t, which falls back to divide-and-conquer based on P_x when necessary. It can be seen that while P_t in FLAT only generates minimal phrases, P_t in HIER generates (and thus memorizes) phrases at all levels of granularity.

4.1 Length-based Parameter Tuning

There are still two problems with HIER, one theoretical, and one practical. Theoretically, HIER contains itself as its base measure, and stochastic process models that include themselves as base measures are deficient, as noted in Cohen et al. (2010). Practically, while the Pitman-Yor process in HIER shares the parameters s and d over all phrase pairs in the model, long phrase pairs are much more sparse

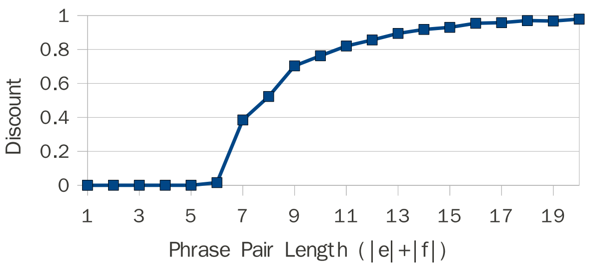

Figure 2: Learned discount values by phrase pair length.

than short phrase pairs, and thus it is desirable to appropriately adjust the parameters of Equation (2) according to phrase pair length.

In order to solve these problems, we reformulate the model so that each phrase length $l = |f|+|e|$ has its own phrase parameters $\theta_{t,l}$ and symbol parameters $\theta_{x,l}$, which are given separate priors:

$$\theta_{t,l} \sim PY(s, d, P_{dac,l})$$
$$\theta_{x,l} \sim Dirichlet(\alpha)$$

We will call this model HLEN.

The generative story is largely similar to HIER with a few minor changes. When we generate a sentence, we first choose its length l according to a uniform distribution over all possible sentence lengths

$$l \sim Uniform(1, L),$$

where L is the size $|e| + |f|$ of the longest sentence in the corpus. We then generate a phrase pair from the probability $P_{t,l}(\langle e, f\rangle)$ for length l. The base measure for HLEN is identical to that of HIER, with one minor change: when we fall back to two shorter phrases, we choose the length of the left phrase from $l_l \sim Uniform(1, l-1)$, set the length of the right phrase to $l_r = l - l_l$, and generate the smaller phrases from P_{t,l_l} and P_{t,l_r} respectively.

It can be seen that phrases at each length are generated from different distributions, and thus the parameters for the Pitman-Yor process will be different for each distribution. Further, as l_l and l_r must be smaller than l, $P_{t,l}$ no longer contains itself as a base measure, and is thus not deficient.

An example of the actual discount values learned in one of the experiments described in Section 7 is shown in Figure 2. It can be seen that, as expected, the discounts for short phrases are lower than those of long phrases. In particular, phrase pairs of length up to six (for example, $|e| = 3$, $|f| = 3$) are given discounts of nearly zero while larger phrases are more heavily discounted. We conjecture that this is related to the observation by Koehn et al. (2003) that using phrases where $\max(|e|, |f|) \leq 3$ cause significant improvements in BLEU score, while using larger phrases results in diminishing returns.

4.2 Implementation

Previous research has used a variety of sampling methods to learn Bayesian phrase based alignment models (DeNero et al., 2008; Blunsom et al., 2009; Blunsom and Cohn, 2010). All of these techniques are applicable to the proposed model, but we choose to apply the sentence-based blocked sampling of Blunsom and Cohn (2010), which has desirable convergence properties compared to sampling single alignments. As exhaustive sampling is too slow for practical purpose, we adopt the beam search algorithm of Saers et al. (2009), and use a probability beam, trimming spans where the probability is at least 10^{10} times smaller than that of the best hypothesis in the bucket.

One important implementation detail that is different from previous models is the management of phrase counts. As a phrase pair t_a may have been generated from two smaller component phrases t_b and t_c, when a sample containing t_a is removed from the distribution, it may also be necessary to decrement the counts of t_b and t_c as well. The Chinese Restaurant Process representation of P_t (Teh, 2006) lends itself to a natural and easily implementable solution to this problem. For each table representing a phrase pair t_a, we maintain not only the number of customers sitting at the table, but also the identities of phrases t_b and t_c that were originally used when generating the table. When the count of the table t_a is reduced to zero and the table is removed, the counts of t_b and t_c are also decremented.

5 Phrase Extraction

In this section, we describe both traditional heuristic phrase extraction, and the proposed model-based extraction method.

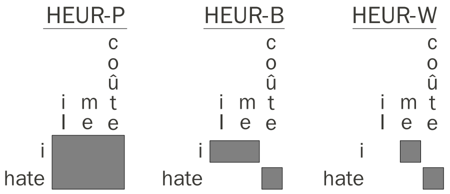

Figure 3: The phrase, block, and word alignments used in heuristic phrase extraction.

5.1 Heuristic Phrase Extraction

The traditional method for heuristic phrase extraction from word alignments exhaustively enumerates all phrases up to a certain length consistent with the alignment (Och et al., 1999). Five features are used in the phrase table: the conditional phrase probabilities in both directions estimated using maximum likelihood $P_{ml}(f|e)$ and $P_{ml}(e|f)$, lexical weighting probabilities (Koehn et al., 2003), and a fixed penalty for each phrase. We will call this heuristic extraction from word alignments HEUR-W. These word alignments can be acquired through the standard GIZA++ training regimen.

We use the combination of our ITG-based alignment with traditional heuristic phrase extraction as a second baseline. An example of these alignments is shown in Figure 3. In model HEUR-P, minimal phrases generated from P_t are treated as aligned, and we perform phrase extraction on these alignments. However, as the proposed models tend to align relatively large phrases, we also use two other techniques to create smaller alignment chunks that prevent sparsity. We perform regular sampling of the trees, but if we reach a minimal phrase generated from P_t, we continue traveling down the tree until we reach either a one-to-many alignment, which we will call HEUR-B as it creates alignments similar to the block ITG, or an at-most-one alignment, which we will call HEUR-W as it generates word alignments. It should be noted that forcing alignments smaller than the model suggests is only used for generating alignments for use in heuristic extraction, and does not affect the training process.

5.2 Model-Based Phrase Extraction

We also propose a method for phrase table extraction that directly utilizes the phrase probabilities $P_t(\langle e, f\rangle)$. Similarly to the heuristic phrase tables, we use conditional probabilities $P_t(f|e)$ and $P_t(e|f)$, lexical weighting probabilities, and a phrase penalty. Here, instead of using maximum likelihood, we calculate conditional probabilities directly from P_t probabilities:

$$P_t(f|e) = P_t(\langle e, f\rangle) / \sum_{\{\tilde{f}:c(\langle e,\tilde{f}\rangle)\geq 1\}} P_t(\langle e, \tilde{f}\rangle)$$

$$P_t(e|f) = P_t(\langle e, f\rangle) / \sum_{\{\tilde{e}:c(\langle \tilde{e},f\rangle)\geq 1\}} P_t(\langle \tilde{e}, f\rangle).$$

To limit phrase table size, we include only phrase pairs that are aligned at least once in the sample.

We also include two more features: the phrase pair joint probability $P_t(\langle e, f\rangle)$, and the average posterior probability of each span that generated $\langle e, f\rangle$ as computed by the inside-outside algorithm during training. We use the span probability as it gives a hint about the reliability of the phrase pair. It will be high for common phrase pairs that are generated directly from the model, and also for phrases that, while not directly included in the model, are composed of two high probability child phrases.

It should be noted that while for FLAT and HIER P_t can be used directly, as HLEN learns separate models for each length, we must combine these probabilities into a single value. We do this by setting

$$P_t(\langle e, f\rangle) = P_{t,l}(\langle e, f\rangle) c(l) / \sum_{\tilde{l}=1}^{L} c(\tilde{l})$$

for every phrase pair, where $l = |e| + |f|$ and $c(l)$ is the number of phrases of length l in the sample.

We call this model-based extraction method MOD.

5.3 Sample Combination

As has been noted in previous works, (Koehn et al., 2003; DeNero et al., 2006) exhaustive phrase extraction tends to out-perform approaches that use syntax or generative models to limit phrase boundaries. DeNero et al. (2006) state that this is because generative models choose only a single phrase segmentation, and thus throw away many good phrase pairs that are in conflict with this segmentation.

Luckily, in the Bayesian framework it is simple to overcome this problem by combining phrase tables from multiple samples. This is equivalent to approximating the integral over various parameter configurations in Equation (1). In MOD, we do this by taking the average of the joint probability and span probability features, and re-calculating the conditional probabilities from the averaged joint probabilities.

6 Related Work

In addition to the previously mentioned phrase alignment techniques, there has also been a significant body of work on phrase extraction (Moore and Quirk (2007), Johnson et al. (2007a), *inter alia*). DeNero and Klein (2010) presented the first work on joint phrase alignment and extraction at multiple levels. While they take a supervised approach based on discriminative methods, we present a fully unsupervised generative model.

A generative probabilistic model where longer units are built through the binary combination of shorter units was proposed by de Marcken (1996) for monolingual word segmentation using the minimum description length (MDL) framework. Our work differs in that it uses Bayesian techniques instead of MDL, and works on two languages, not one.

Adaptor grammars, models in which non-terminals memorize subtrees that lie below them, have been used for word segmentation or other monolingual tasks (Johnson et al., 2007b). The proposed method could be thought of as synchronous adaptor grammars over two languages. However, adaptor grammars have generally been used to specify only two or a few levels as in the FLAT model in this paper, as opposed to recursive models such as HIER or many-leveled models such as HLEN. One exception is the variational inference method for adaptor grammars presented by Cohen et al. (2010) that is applicable to recursive grammars such as HIER. We plan to examine variational inference for the proposed models in future work.

7 Experimental Evaluation

We evaluate the proposed method on translation tasks from four languages, French, German, Spanish, and Japanese, into English.

	de-en	es-en	fr-en	ja-en
TM (en)	1.80M	1.62M	1.35M	2.38M
TM (other)	1.85M	1.82M	1.56M	2.78M
LM (en)	52.7M	52.7M	52.7M	44.7M
Tune (en)	49.8k	49.8k	49.8k	68.9k
Tune (other)	47.2k	52.6k	55.4k	80.4k
Test (en)	65.6k	65.6k	65.6k	40.4k
Test (other)	62.7k	68.1k	72.6k	48.7k

Table 1: The number of words in each corpus for TM and LM training, tuning, and testing.

7.1 Experimental Setup

The data for French, German, and Spanish are from the 2010 Workshop on Statistical Machine Translation (Callison-Burch et al., 2010). We use the news commentary corpus for training the TM, and the news commentary and Europarl corpora for training the LM. For Japanese, we use data from the NTCIR patent translation task (Fujii et al., 2008). We use the first 100k sentences of the parallel corpus for the TM, and the whole parallel corpus for the LM. Details of both corpora can be found in Table 1. Corpora are tokenized, lower-cased, and sentences of over 40 words on either side are removed for TM training. For both tasks, we perform weight tuning and testing on specified development and test sets.

We compare the accuracy of our proposed method of joint phrase alignment and extraction using the FLAT, HIER and HLEN models, with a baseline of using word alignments from GIZA++ and heuristic phrase extraction. Decoding is performed using Moses (Koehn and others, 2007) using the phrase tables learned by each method under consideration, as well as standard bidirectional lexical reordering probabilities (Koehn et al., 2005). Maximum phrase length is limited to 7 in all models, and for the LM we use an interpolated Kneser-Ney 5-gram model.

For GIZA++, we use the standard training regimen up to Model 4, and combine alignments with `grow-diag-final-and`. For the proposed models, we train for 100 iterations, and use the final sample acquired at the end of the training process for our experiments using a single sample[6]. In addition,

[6] For most models, while likelihood continued to increase gradually for all 100 iterations, BLEU score gains plateaued after 5-10 iterations, likely due to the strong prior information

			de-en		es-en		fr-en		ja-en	
Align	Extract	# Samp.	BLEU	Size	BLEU	Size	BLEU	Size	BLEU	Size
GIZA++	HEUR-W	1	**16.62**	4.91M	**22.00**	4.30M	21.35	4.01M	**23.20**	4.22M
FLAT	MOD	1	13.48	136k	19.15	125k	17.97	117k	16.10	89.7k
HIER	MOD	1	**16.58**	1.02M	**21.79**	859k	**21.50**	751k	**23.23**	723k
HLEN	MOD	1	**16.49**	1.17M	21.57	930k	21.31	860k	**23.19**	820k
HIER	MOD	10	**16.53**	3.44M	**21.84**	2.56M	**21.57**	2.63M	**23.12**	2.21M
HLEN	MOD	10	**16.51**	3.74M	**21.69**	3.00M	**21.53**	3.09M	**23.20**	2.70M

Table 2: BLEU score and phrase table size by alignment method, extraction method, and samples combined. Bold numbers are not significantly different from the best result according to the sign test ($p < 0.05$) (Collins et al., 2005).

we also try averaging the phrase tables from the last ten samples as described in Section 5.3.

7.2 Experimental Results

The results for these experiments can be found in Table 2. From these results we can see that when using a single sample, the combination of using HIER and model probabilities achieves results approximately equal to GIZA++ and heuristic phrase extraction. This is the first reported result in which an unsupervised phrase alignment model has built a phrase table directly from model probabilities and achieved results that compare to heuristic phrase extraction. It can also be seen that the phrase table created by the proposed method is approximately 5 times smaller than that obtained by the traditional pipeline.

In addition, HIER significantly outperforms FLAT when using the model probabilities. This confirms that phrase tables containing only minimal phrases are not able to achieve results that compete with phrase tables that use multiple granularities.

Somewhat surprisingly, HLEN consistently slightly underperforms HIER. This indicates potential gains to be provided by length-based parameter tuning were outweighed by losses due to the increased complexity of the model. In particular, we believe the necessity to combine probabilities from multiple $P_{t,l}$ models into a single phrase table may have resulted in a distortion of the phrase probabilities. In addition, the assumption that phrase lengths are generated from a uniform distribution is likely too strong, and further gains could likely be achieved by more accurate modeling of phrase lengths. We leave further adjustments to the HLEN model to future work.

	FLAT		HIER	
MOD	17.97	117k	21.50	751k
HEUR-W	21.52	5.65M	21.68	5.39M
HEUR-B	21.45	4.93M	21.41	2.61M
HEUR-P	21.56	4.88M	21.47	1.62M

Table 3: Translation results and phrase table size for various phrase extraction techniques (French-English).

It can also be seen that combining phrase tables from multiple samples improved the BLEU score for HLEN, but not for HIER. This suggests that for HIER, most of the useful phrase pairs discovered by the model are included in every iteration, and the increased recall obtained by combining multiple samples does not consistently outweigh the increased confusion caused by the larger phrase table.

We also evaluated the effectiveness of model-based phrase extraction compared to heuristic phrase extraction. Using the alignments from HIER, we created phrase tables using model probabilities (MOD), and heuristic extraction on words (HEUR-W), blocks (HEUR-B), and minimal phrases (HEUR-P) as described in Section 5. The results of these experiments are shown in Table 3. It can be seen that model-based phrase extraction using HIER outperforms or insignificantly underperforms heuristic phrase extraction over all experimental settings, while keeping the phrase table to a fraction of the size of most heuristic extraction methods.

Finally, we varied the size of the parallel corpus for the Japanese-English task from 50k to 400k sen-

provided by P_{base}. As iterations took 1.3 hours on a single processor, good translation results can be achieved in approximately 13 hours, which could further reduced using distributed sampling (Newman et al., 2009; Blunsom et al., 2009).

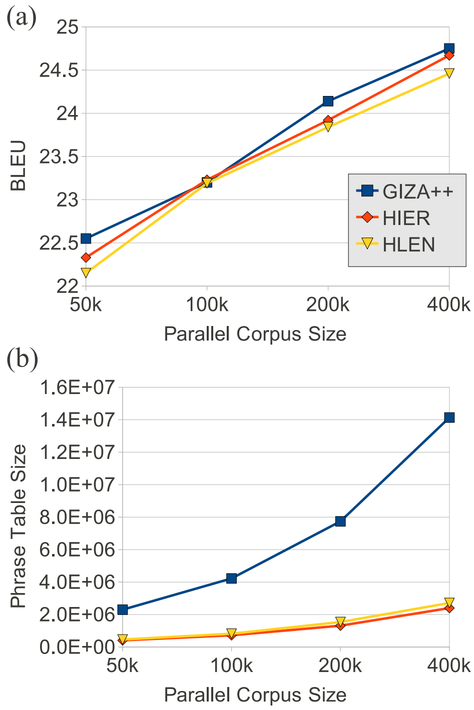

Figure 4: The effect of corpus size on the accuracy (a) and phrase table size (b) for each method (Japanese-English).

tences and measured the effect of corpus size on translation accuracy. From the results in Figure 4 (a), it can be seen that at all corpus sizes, the results from all three methods are comparable, with insignificant differences between GIZA++ and HIER at all levels, and HLEN lagging slightly behind HIER. Figure 4 (b) shows the size of the phrase table induced by each method over the various corpus sizes. It can be seen that the tables created by GIZA++ are significantly larger at all corpus sizes, with the difference being particularly pronounced at larger corpus sizes.

8 Conclusion

In this paper, we presented a novel approach to joint phrase alignment and extraction through a hierarchical model using non-parametric Bayesian methods and inversion transduction grammars. Machine translation systems using phrase tables learned directly by the proposed model were able to achieve accuracy competitive with the traditional pipeline of word alignment and heuristic phrase extraction, the first such result for an unsupervised model.

For future work, we plan to refine HLEN to use a more appropriate model of phrase length than the uniform distribution, particularly by attempting to bias against phrase pairs where one of the two phrases is much longer than the other. In addition, we will test probabilities learned using the proposed model with an ITG-based decoder. We will also examine the applicability of the proposed model in the context of hierarchical phrases (Chiang, 2007), or in alignment using syntactic structure (Galley et al., 2006). It is also worth examining the plausibility of variational inference as proposed by Cohen et al. (2010) in the alignment context.

Acknowledgments

This work was performed while the first author was supported by the JSPS Research Fellowship for Young Scientists.

References

Phil Blunsom and Trevor Cohn. 2010. Inducing synchronous grammars with slice sampling. In *Proceedings of the Human Language Technology: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A Gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 782–790.

Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53.

Colin Cherry and Dekang Lin. 2007. Inversion transduction grammar for joint phrasal translation modeling. In *Proceedings of the NAACL Workshop on Syntax and Structure in Machine Translation*.

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

Shay B. Cohen, David M. Blei, and Noah A. Smith. 2010. Variational inference for adaptor grammars. In *Proceedings of the Human Language Technology: The*

11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 564–572.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 531–540.

Carl de Marcken. 1996. *Unsupervised Language Acquisition*. Ph.D. thesis, Massachusetts Institute of Technology.

John DeNero and Dan Klein. 2008. The complexity of phrase alignment problems. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 25–28.

John DeNero and Dan Klein. 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.

John DeNero, Dan Gillick, James Zhang, and Dan Klein. 2006. Why generative phrase models underperform surface heuristics. In *Proceedings of the 1st Workshop on Statistical Machine Translation*, pages 31–38.

John DeNero, Alex Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 314–323.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2008. Overview of the patent translation task at the NTCIR-7 workshop. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pages 389–400.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, pages 961–968.

J. Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007a. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007b. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in Neural Information Processing Systems*, 19:641.

Philipp Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Phillip Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, pages 48–54.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 104–111.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. pages 133–139.

Robert C. Moore and Chris Quirk. 2007. An iteratively-trained segmentation-free phrase translation model for statistical machine translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 112–119.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10:1801–1828.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the 4th Conference on Empirical Methods in Natural Language Processing*, pages 20–28.

Jim Pitman and Marc Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Markus Saers, Joakim Nivre, and Dekai Wu. 2009. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In *Proceedings of the The 11th International Workshop on Parsing Technologies*.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Hao Zhang, Chris Quirk, Robert C. Moore, and Daniel Gildea. 2008. Bayesian learning of non-compositional phrases with synchronous parsing. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 97–105.