Electronic Preprint for Journal of Information Processing Vol.20 No.2

Regular Paper

Joint Phrase Alignment and Extraction for Statistical Machine Translation

Graham Neubig[1,2,a)] Taro Watanabe[2] Eiichiro Sumita[2] Shinsuke Mori[1]
Tatsuya Kawahara[1]

Received: May 24, 2011, Accepted: December 16, 2011

Abstract: The phrase table, a scored list of bilingual phrases, lies at the center of phrase-based machine translation systems. We present a method to directly learn this phrase table from a parallel corpus of sentences that are not aligned at the word level. The key contribution of this work is that while previous methods have generally only modeled phrases at one level of granularity, in the proposed method phrases of many granularities are included directly in the model. This allows for the direct learning of a phrase table that achieves competitive accuracy without the complicated multi-step process of word alignment and phrase extraction that is used in previous research. The model is achieved through the use of non-parametric Bayesian methods and inversion transduction grammars (ITGs), a variety of synchronous context-free grammars (SCFGs). Experiments on several language pairs demonstrate that the proposed model matches the accuracy of the more traditional two-step word alignment/phrase extraction approach while reducing its phrase table to a fraction of its original size.

Keywords: statistical machine translation, phrase alignment, non-parametric Bayesian statistics, inversion transduction grammars

1. Introduction

Statistical machine translation (SMT) has seen great improvements over the past decade thanks largely to the introduction of phrase-based translation, which helps resolve lexical ambiguity and short-distance reordering by translating multi-word phrases as single chunks. The most important element of phrase-based SMT systems is the "phrase table," a scored list of bilingual phrase pairs that are translations of each other. This phrase table is generated from a parallel corpus of translated sentences that are not aligned at the word or phrase level.

Traditional systems construct phrase tables by going through a two-step pipeline. The first step consists of finding alignments between words or minimal phrases in both sentences, while the second step extracts an expanded phrase table from these alignments through heuristic combination of words or minimal phrases into longer units. The ability to use both short single-word units and longer phrases is one of the major reasons why phrase-based translation achieves superior results to word-based methods. However, it has been shown in previous research [13] that 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. In addition, exhaustively extracted phrase tables are often unnecessarily large, which results in an increase in the amount of time and memory required to run machine translation systems.

In this paper, we propose an approach that is able to reduce the two steps of alignment and extraction into a single step by including phrases of multiple granularities in a probabilistic alignment model. The model is based on inversion transduction grammars (ITGs [34]), a variety of synchronous context free grammars (SCFGs). ITGs allow for efficient word or phrase alignment [2], [5], [35] through the use of bilingual chart parsing, similar to parsing algorithms used widely for the parsing of monolingual CFGs.

In contrast to previous approaches, which generally only attempt to model word (or minimal phrase) alignments, the proposed method models phrases at multiple levels of granularity through a novel recursive formulation, where larger phrase pairs are probabilistically constructed from two smaller phrase pairs. The model uses methods from non-parametric Bayesian statistics, which favor simpler models, preventing the over-fitting that occurs in some previous alignment approaches [22].

Using this model, we create phrase tables and perform machine translation experiments over four language pairs. We observe that the proposed hierarchical model is able to meet or exceed results attained by the traditional combination of word alignment and heuristic phrase extraction with significantly smaller phrase table size. We also find that in contrast, previously proposed ITG-based phrase alignment approaches are not able to achieve competitive accuracy without heuristic phrase extraction and the accompanying increase in phrase table size.

[1] Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[2] National Institute of Information and Communications Technology, Keihanna Science City, Kyoto 619–2089, Japan
[a)] neubig@ar.media.kyoto-u.ac.jp

2. Phrase-Based Statistical Machine Translation

Machine translation is the process of automatically translating a sentence F in a source (foreign) language, into an equivalent sentence E in the target (English) language. Many modern machine translation (MT) systems utilize phrase-based MT [20] techniques, which break F into phrases of one or more words, each of which is individually translated and reordered to form E. An example of a phrase-based translation is shown in **Fig. 1**. It should be noted that in addition to F and E, there is a string A of alignment spans that indicates which parts of F were translated into which parts of E. Each element of A takes the form $\{[a_{e1}, a_{e2}], [a_{f1}, a_{f2}]\}$ indicating a single pair of phrases in the source and target sentences. The variables a_{e1} and a_{e2} indicate the position of the first and last words of the target phrase, while a_{f1} and a_{f2} indicate the position of the first and last words of the source phrase respectively.

For any particular source sentence F there are many possible translations, some more natural or semantically correct than others. Statistical machine translation (SMT) attempts to resolve this ambiguity by creating a statistical model for the target sentence and alignment given the source sentence, and finding the target sentence that maximizes this probability:

$$\hat{E} = \underset{E}{\operatorname{argmax}} P(E, A|F). \tag{1}$$

The predominant paradigm for calculating this probability is the log-linear model of Ref. [27]. This model defines the logarithm of the translation probability as a linear combination of a set of feature functions $\phi_1, \ldots, \phi_I$ over E, F, and A, weighted with weights $\lambda_1, \ldots, \lambda_I$

$$\log P(E, A|F) = \sum_{i=1}^{I} \lambda_i \phi_i(E, F, A). \tag{2}$$

This formulation allows arbitrary features of E, F, and A to be used in determining the translation probabilities. Commonly used feature functions include log language model probabilities and reordering probabilities. The language model probabilities are defined over E, and attempt to measure the fluency of the generated sentence. The reordering probabilities are defined over A, and attempt to ensure that the word order is appropriate.

However, the features that most directly affect the translation quality are those that belong to the *phrase table*. As shown in the example in **Fig. 2**, the phrase table is a collection of phrase pairs, consisting of equivalent source and target language phrases ($\boldsymbol{f}$ and $\boldsymbol{e}$ respectively). Each phrase pair is additionally scored with several feature functions, which will be explained in more detail in Section 6. These feature functions are used to provide an indication of the reliability or frequency of each phrase pair, and can be learned from a corpus consisting of translated pairs of sentences in the source and target languages.

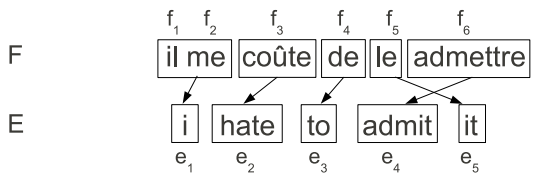

Fig. 1 The target sentence F, source sentence E, and alignment A.

3. Alignment Using Inversion Transduction Grammars

The first step in creating a phrase table from a sentence-aligned parallel corpus is *alignment*, the process of finding which words or phrases in the source and target sides of the training data correspond to each other. Following the definitions presented in the previous section, this means that we are given a parallel training corpus consisting of $\mathcal{F} = F_1, \ldots, F_n$ and $\mathcal{E} = E_1, \ldots, E_n$, and we must find the corresponding alignments $\mathcal{A} = A_1, \ldots, A_n$. One framework for learning these alignments that has been used in a number of recent works [2], [5], [35] is the inversion transduction grammar (ITG) [34].

ITGs are a form of context-free grammar (CFG) in Chomsky normal form [7], defined over two languages instead of one. Like normal CFGs, ITGs have non-terminal, pre-terminal, and terminal symbols, but each node generates bilingual phrase pairs $\langle \boldsymbol{e}, \boldsymbol{f} \rangle$ instead of the single monolingual phrases generated by normal CFGs. The most important characteristic of ITGs is the non-terminal symbols, which can be either "straight" or "inverted." If a non-terminal node's left and right child nodes have generated the phrase pairs $\langle \boldsymbol{e}_1, \boldsymbol{f}_1 \rangle$ and $\langle \boldsymbol{e}_2, \boldsymbol{f}_2 \rangle$ respectively, in the case of the straight non-terminals, these will be concatenated in order as $\langle \boldsymbol{e}_1\boldsymbol{e}_2, \boldsymbol{f}_1\boldsymbol{f}_2 \rangle$, while in the case of inverted nodes, the phrases of f are concatenated in inverted reverse order as $\langle \boldsymbol{e}_1\boldsymbol{e}_2, \boldsymbol{f}_2\boldsymbol{f}_1 \rangle$. An example of straight and inverted nodes is shown in **Fig. 3**.

Like probabilistic CFGs, which assign probabilities to each generative grammar rule over monolingual phrases, ITGs can also be assigned a generative probability distribution over bilingual phrase pairs. The traditional ITG generative probability for a particular phrase pair $P_{flat}(\langle \boldsymbol{e}, \boldsymbol{f} \rangle; \theta_x, \theta_t)$ is parameterized by a phrase table θ_t (which specifies a probability distribution over terminal symbols) and a symbol distribution θ_x (which specifies a probability distribution over non-terminal and pre-terminal symbols). A number of small variations of this traditional ITG model have been proposed in the literature, but we use the following generative story as a representative of previously proposed models.

$\boldsymbol{e}$	$\boldsymbol{f}$	$\phi_1(\boldsymbol{e}, \boldsymbol{f})$	$\phi_2(\boldsymbol{e}, \boldsymbol{f})$	
admit	admettre	0.5	0.3	...
admit	avouer	0.5	0.8	
admit it	le admettre	1.0	0.28	
...				

Fig. 2 An example of part of the phrase table with source phrases $\boldsymbol{e}$, target phrases $\boldsymbol{f}$, and feature functions ϕ_i.

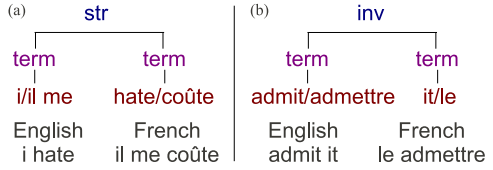

Fig. 3 A straight (a) and inverted (b) ITG production.

(1) Generate symbol x from the multinomial distribution $P_x(x;\theta_x)$. x can take the values TERM, STR, or INV.
(2) According to the value of x take the following actions.
  (a) If x = TERM, the pre-terminal, generate a phrase pair from the phrase table $P_t(\langle \boldsymbol{e}, \boldsymbol{f} \rangle; \theta_t)$.
  (b) If x = STR, a straight ITG non-terminal, generate phrase pairs $\langle \boldsymbol{e}_1, \boldsymbol{f}_1 \rangle$ and $\langle \boldsymbol{e}_2, \boldsymbol{f}_2 \rangle$ from P_{flat}, and concatenate them into a single phrase pair $\langle \boldsymbol{e}_1\boldsymbol{e}_2, \boldsymbol{f}_1\boldsymbol{f}_2 \rangle$.
  (c) If x = INV, an inverted ITG non-terminal, follows the same process as (b), but concatenate $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ in reverse order $\langle \boldsymbol{e}_1\boldsymbol{e}_2, \boldsymbol{f}_2\boldsymbol{f}_1 \rangle$.

The result of this generative process is a bilingual phrase pair, along with its corresponding generative probability. We will refer to this model as FLAT.

ITG-based models can be used to find alignments for words in parallel sentences through the process of biparsing [34]. Within the previously described ITG framework, a sentence pair $\langle E, F \rangle$ can be defined as the phrase pair that is generated by the node at the top of the derivation tree. Biparsing for ITGs finds the most likely derivation for this sentence pair given the ITG probabilities. Once we have this most likely derivation, we treat all phrase pairs that were generated from the same terminal symbols as aligned (for example, in Fig. 3 : "i/il me," "hate/coûte," "to/de," "admit/admettre," and "it/le").

4. Bayesian Modeling for Inversion Transduction Grammars

The probabilities of ITG models can be calculated in the same manner as traditional unsupervised PCFGs using the expectation-maximization algorithm and maximum likelihood estimation. However, as noted by Ref. [12], when many-to-many alignments are allowed, the solution that maximizes the likelihood is often to simply memorize every sentence as a single phrase pair, a degenerate solution that defeats the purpose of performing alignment. Reference [35] and others propose dealing with this problem by putting a prior probability $P(\theta_x, \theta_t)$ on the parameters, which allows us to bias towards compact models and prevent this degenerate solution.

Priors based on Bayesian statistics have proven useful for controlling model complexity in previous work, so we adapt a similar approach here. The symbol distribution parameters θ_x specify a multinomial distribution over 3 elements. Because of this it is natural to use a Dirichlet distribution as a prior for θ_x, as the Dirichlet distribution is the conjugate prior of the multinomial distribution.

$$\theta_x \sim Dirichlet(\alpha). \tag{3}$$

α is a hyper-parameter controlling the sparsity of the distribution, but this has little empirical effect on the results, so we arbitrarily set $\alpha = 1$.

The phrase table parameters θ_t specify a multinomial distribution over an undetermined number of elements (every possible phrase pair). Previous work on both word alignment [2], [35] and other natural language processing tasks has used Bayesian non-parametric techniques to specify priors over these sort of infinite multinomial distributions. In particular we use a prior based on the non-parametric Pitman-Yor process [31], [33]. The Pitman-Yor process is a generalization of the better-known Dirichlet process prior that has been used in previous work on word alignment. This prior is expressed as

$$\theta_t \sim PY(d, s, P_{base}). \tag{4}$$

In the Pitman-Yor process, 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. Within the framework of the ITG model, this indicates that phrase pairs that are generated by P_t many times are more likely to be reused (the *rich-get-richer* effect), which results in the induction of phrase tables with fewer, but more helpful phrases. In the FLAT model, non-terminal nodes are first generated from P_x, reducing the sentence to managable chunks, followed by the generation of the pre-terminal from P_x, then a generation of a minimal phrase pair from P_t. As P_t will only generate a phrase pair at the end of the generative process, only phrase pairs of the smallest level of granularity will be memorized and given higher probability 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 [33]. We confirmed in preliminary experiments (using the data described in Section 8) that the Pitman-Yor process with automatically adjusted parameters results in superior alignment results, outperforming the sparse Dirichlet process priors used in previous research [*1].

4.1 Base Measure

P_{base} in Eq. (4) is the base measure, 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. In particular, there are three pieces of prior knowledge that we would like to provide through the base measure. First, we would like to minimize the number of phrases that are not aligned to any phrase in the other language, as we can assume that most of the phrases will have some corresponding translation. Second, we would like to bias against overly long phrases, as these are likely to cause sparsity and hurt generalization performance when the model is tested on new data. Finally, when aligning multi-word phrases, it makes sense to align phrases that are composed of words that are good

[*1] Following Ref. [33], we put priors on s ($Gamma(\alpha = 2, \beta = 1)$) and d ($Beta(\alpha = 2, \beta = 2)$) for the Pitman-Yor process, and sample their values. These priors do not provide a strong bias towards any particular value of s or d, allowing the model freedom to choose values that maximize the likelihood of the training data. We set $\alpha = 1^{-10}$ for the Dirichlet process.

translations of each-other.

Here, we adopt a formulation similar to that of Ref. [11] that is able to satisfy all of these desiderata. P_{base} is first calculated by choosing whether to generate an unaligned phrase pair (where $|\boldsymbol{e}| = 0$ or $|\boldsymbol{f}| = 0$) according to a fixed probability p_u. p_u should generally be a small value to minimize the number of unaligned phrases [*2]. Based on this choice, we next generate an aligned phrase pair from P_{ba}, or an unaligned phrase pair from P_{bu}.

For P_{ba}, we use the following probability:

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

P_{pois} is the Poisson distribution with the average length parameter λ, where k represents the phrase length $|\boldsymbol{f}|$ or $|\boldsymbol{e}|$.

$$P_{pois}(k|\lambda) = \frac{(\lambda - 1)^{k-1}}{(k-1)!} e^{-(\lambda - 1)}. \tag{5}$$

We set λ to a relatively small value, which allows us to bias against overly long phrases [*3].

P_{uni} is the unigram probability of a particular phrase, and P_{m1} is the word-based Model 1 [3] probability of one phrase given the other. Model 1 probabilities are word-based translation probabilities that help to indicate whether the words in each phrase are good translations of each-other. The phrase-based Model 1 probability is calculated according to the following equation:

$$P_{m1}(\boldsymbol{e}|\boldsymbol{f}) = \prod_{i=1}^{|\boldsymbol{e}|} \frac{1}{|\boldsymbol{f}|} \sum_{j=1}^{|\boldsymbol{f}|} P_{m1}(e_i|f_j) \tag{6}$$

where e_i and f_j are the ith and jth words in phrases $\boldsymbol{e}$ and $\boldsymbol{f}$ respectively. The word-based probabilities $P_{m1}(e_i|f_j)$ and $P_{m1}(f_j|e_i)$ are parameters of the model, and can be calculated efficiently using the expectation maximization algorithm [3] before starting phrase alignment. Following Ref. [21], we combine the Model 1 probabilities in both directions using the geometric mean [*4], which allows us to encourage alignments that are supported by both models.

For P_{bu}, in the case of $|\boldsymbol{f}| = 0$, we calculate the probability as follows:

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

The probability can be calculated similarly when $|\boldsymbol{e}| = 0$. Note that P_{bu} is divided by 2 as the probability is considering null alignments in both directions.

5. Hierarchical ITG Model

While in FLAT only minimal phrases were memorized by the model, as Ref. [11] notes and we confirm in the experiments in Section 8, using only minimal phrases leads to inferior translation results for phrase-based translation. Because of this, previous research has combined FLAT with heuristic phrase extraction, which exhaustively combines all adjacent phrases permitted by the word alignments [29]. 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 phrase table that is able to achieve competitive accuracy in a single step through a fully probabilistic process.

Similarly to FLAT, HIER assigns a probability $P_{hier}(\langle \boldsymbol{e}, \boldsymbol{f} \rangle; \theta_x, \theta_t)$ to phrase pairs, and is parameterized by a phrase table θ_t and a symbol distribution θ_x. The main difference between the two models is that non/pre-terminal symbols and phrase pairs are generated in reverse 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 Eq. (4), resulting in the following distribution for θ_t.

$$\theta_t \sim PY(d, s, P_{dac}) \tag{7}$$

P_{dac} essentially generates a single longer phrase through two generations and a combination of shorter phrases, allowing even long phrase pairs to be given significant amounts of probability when justified. 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, STR, 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 4.1.
  (b) If x = STR, generate $\langle \boldsymbol{e}_1, \boldsymbol{f}_1 \rangle$ and $\langle \boldsymbol{e}_2, \boldsymbol{f}_2 \rangle$ from P_{hier}, and concatenate them into a single phrase pair $\langle \boldsymbol{e}_1\boldsymbol{e}_2, \boldsymbol{f}_1\boldsymbol{f}_2 \rangle$.
  (c) If x = INV, follow the same process as (b), but concatenate $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ in reverse order $\langle \boldsymbol{e}_1\boldsymbol{e}_2, \boldsymbol{f}_2\boldsymbol{f}_1 \rangle$.

A comparison of derivation trees for FLAT and HIER is shown in **Fig. 4**. As previously described, FLAT first generates from the symbol distribution P_x, then from the phrase distribution P_t. On the other hand, HIER generates directly from P_t, which falls back to divide-and-conquer based on P_x when necessary. The minimal and non-minimal phrase pairs that are generated by P_t are surrounded by solid and dotted lines respectively. 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.

5.1 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 technically deficient, as noted in Ref. [8]. 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 than short phrase pairs, and thus it is desir-

[*2] We choose 10^{-2}, 10^{-3}, or 10^{-10} based on which value gave the best translation accuracy on the development set.
[*3] We tune λ to 1, 0.1, or 0.01 based on which value gives the best translation accuracy on the development set.
[*4] The probabilities of the geometric mean do not add to one, and are thus not, strictly speaking, proper probabilities. However, we found empirically that even when left unnormalized, they provided much better results than the model using the arithmetic mean, which is mathematically correct.

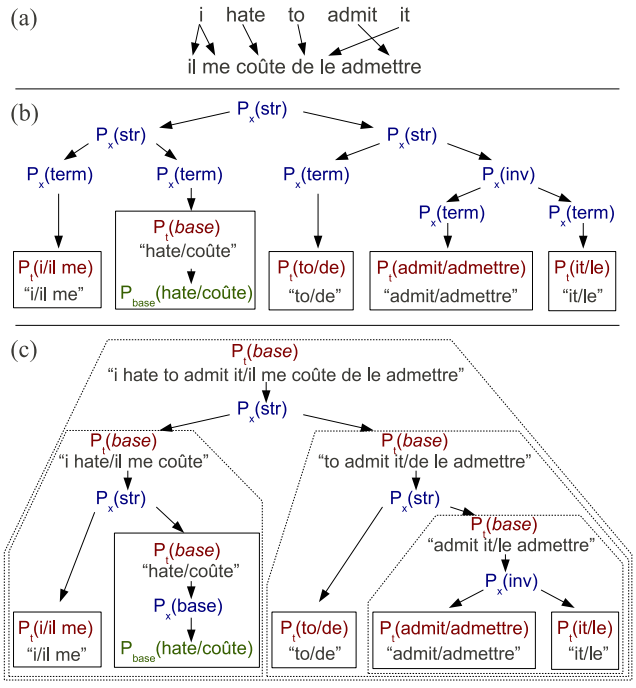

Fig. 4 A word alignment (a), and its derivation according to FLAT (b), and HIER (c). Solid and dotted lines indicate minimal and non-minimal pairs respectively, and phrases memorized by the model are written in quotes under their corresponding instance of P_t. The pair hate/coûte is generated from P_{base}.

able to appropriately adjust the parameters of Eq. (4) according to phrase pair length.

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

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

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 $P_{t,l}(\langle \boldsymbol{e}, \boldsymbol{f} \rangle)$ 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 8 is shown in **Fig. 5**. 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, $|\boldsymbol{e}| = 3$, $|\boldsymbol{f}| = 3$) are given discounts of nearly zero while larger phrases are more heavily discounted. We conjecture that this is related to the observation by Ref. [20] that using phrases where $\max(|\boldsymbol{e}|, |\boldsymbol{f}|) \leq 3$ cause significant improvements in translation accuracy, while using larger phrases results in diminishing returns.

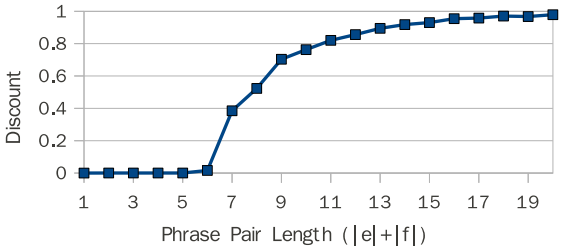

Fig. 5 Learned discount values by phrase pair length.

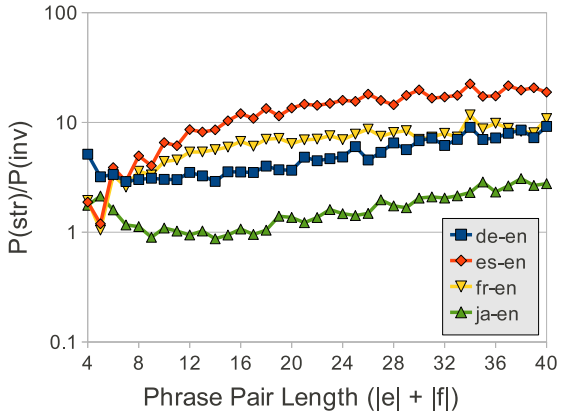

Fig. 6 The ratio of $P_x(str)$ to $P_x(inv)$ by length. Higher values indicate more monotonic alignments.

In addition, the HLEN model has the potential to learn different ITG reordering probabilities for different lengths. An example of the ratio between $P_x(str)$ and $P_x(inv)$ learned for phrases of length 4 to 40 in German, Spanish, French, and Japanese is shown in **Fig. 6**. It can be seen that at the shortest phrase length of 4, which generally corresponds to the reordering of two single-word translations, German has a higher ratio than all other languages. This is intuitive, as both French and Spanish order adjective-noun pairs in the opposite order of English, so there should be more swaps of single words than in German, which places adjective-noun pairs in the same order as English. On the other hand, as sentence length grows longer, French and Spanish surpass German in monotonicity, a result of German having greater divergence in syntax from English. One typical example of this is that sentence-final verbs must be reordered over long distances to their natural position in the middle of the sentence for English. Finally, Japanese has significantly lower monotonicity than all of the European languages at almost all phrase pair lengths, a result of the vast differences in sentence structure between Japanese and English. In contrast, HIER can only learn a single value for $P_x(str)$ and $P_x(inv)$. For German, Spanish, French, and Japanese, the values of $P_x(str)/P_x(inv)$ were 4.83, 5.81, 4.99, and 1.83 respectively, showing that the overall preference for monotonicity or non-monotonicity can be learned, although not in the fine-grained

manner allowed by HLEN.

5.2 Implementation

Previous research has used a variety of methods to learn Bayesian phrase based alignment models, all of which have used Gibbs sampling as their central learning algorithm [1], [2], [11]. All of these techniques are applicable to the proposed model, but we choose to apply the sentence-based sampling proposed by Ref. [1], which has desirable convergence properties compared to sampling single alignments. The majority of computation in the sampling process takes place in the parsing step where probabilities for each possibly aligned bilingual span are calculated to allow for proper sampling of an ITG parse tree for each sentence. Exhaustive parsing of ITGs can be performed in $O(n^6)$, but this is too slow in practical situations for all but the smallest of sentences. To solve this problem, we adopt the beam search algorithm of Ref. [32] as an approximation of full exhaustive parsing, 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 [33] 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.

6. Phrase Extraction

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

6.1 Heuristic Phrase Extraction

The traditional method for heuristic phrase extraction from word alignments exhaustively enumerates all phrases up to a certain length that are consistent with the alignment [29]. After counts for each phrase pair $\langle \boldsymbol{e}, \boldsymbol{f} \rangle$ have been enumerated, these counts are used to calculate five features used in the phrase table:

- **Phrase conditional probabilities:** These are calculated in both directions estimated using maximum likelihood estimation over phrase pair counts:

$$P_{ml}(\boldsymbol{e}|\boldsymbol{f}) = c(\boldsymbol{e}, \boldsymbol{f})/c(\boldsymbol{f})$$
$$P_{ml}(\boldsymbol{f}|\boldsymbol{e}) = c(\boldsymbol{e}, \boldsymbol{f})/c(\boldsymbol{e}).$$

- **Lexical weighting probabilities:** As many phrases have very low counts, simple phrase conditional probabilities are sparse and often do not provide reliable information about the correctness of the phrase pair. To solve this problem, Ref. [20] proposes a method of breaking each phrase down into its respective words, and using the conditional probabilities of the words in the phrase to calculate a more robust estimate of the phrase translation probabilities. The lexical weighting probabilities in both directions are used as two additional features in the model.
- **Phrase penalty:** The last feature is a fixed penalty or bonus for every phrase used. If it is a penalty, the model will prefer to use fewer but longer phrases, and if it is a bonus the model will prefer to use many shorter phrases.

These features are combined in a weighted manner to indicate the overall score of each phrase, with the weights being learned using a training regimen such as minimum error rate training (MERT [26]).

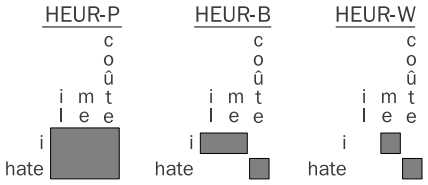

Fig. 7 The phrase, block, and word alignments used in heuristic phrase extraction.

As a baseline, we perform heuristic phrase extraction over the alignments acquired by the FLAT and HIER models. As the proposed method often aligns relatively long phrases, not words, a variety of alignment granularities can be used to create the phrase table (**Fig. 7**). In model HEUR-P, minimal phrases generated from P_t are treated as aligned, and we perform phrase extraction on these alignments. 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 of "blocks," 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.

6.2 Model-Based Phrase Extraction

For our proposed model, we also are able to perform phrase table extraction that directly utilizes the phrase probabilities $P_t(\langle \boldsymbol{e}, \boldsymbol{f} \rangle)$. Similarly to the heuristic phrase tables, we use conditional probabilities $P_t(\boldsymbol{f}|\boldsymbol{e})$ and $P_t(\boldsymbol{e}|\boldsymbol{f})$, lexical weighting probabilities, and a phrase penalty. Here, instead of using maximum likelihood, we calculate conditional probabilities directly from P_t probabilities:

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

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:

- **Model joint probability:** As the proposed method assigns a probability $P_t(\langle \boldsymbol{e}, \boldsymbol{f} \rangle)$ to all phrase pairs, we can use this as an additional feature.

- **Span generative probability:** We also use the average generative probability of each span that generated $\langle \boldsymbol{e}, \boldsymbol{f} \rangle$ as computed by the chart parser during training. This is similar to the joint probability, but is more reliable for low-frequency phrases, where the model probability tends to over-estimate the actual probability. The generative probability 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 and thus can be assumed to be more reliable.

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

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

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

We call this model-based extraction method MOD.

6.3 Sample Combination

As has been noted in previous works [12], [20], exhaustive phrase extraction tends to outperform approaches that use syntax or generative models to limit phrase boundaries. Reference [12] states 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. 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.

7. Related Work

While ITGs have been growing in popularity in recent years, they are by no means the only method for word or phrase alignment. In fact, the seminal IBM models presented in Ref. [3] and the implementation provided by the open-source software GIZA++ [28] are still widely used for word alignment in a large number of systems. The IBM models, while quite powerful, are fundamentally different from the models previously described in this paper in that they are not able to handle many-to-many alignments. As a result, it is necessary to find one-to-many word alignments in both directions, which allows for the capturing of multi-word units on both the source and target sides. These one-to-many alignments can then be combined using heuristics into a many-to-many alignment [20]. Finally, using this alignment, heuristic phrase extraction enumerates all possible phrases that do not conflict with the word alignments [29]. In the next section, we present experimental results comparing alignments acquired using the IBM models with those acquired using ITG-based alignment methods.

In addition to the previously mentioned alignment techniques, there has also been a significant body of work on improving phrase extraction methods (such as Refs. [23] and [16]). Reference [13] 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.

The generative probabilistic model where longer units are built through the binary combination of shorter units that we use in this model was inspired by the model proposed by Ref. [10] for monolingual word segmentation using the minimum description length (MDL) framework. Our work differs in that it uses Bayesian techniques instead of MDL, works on two languages instead of one, and uses words as its basic unit instead of phrases.

Adaptor grammars, models in which non-terminals memorize subtrees that lie below them, have been used for word segmentation or other monolingual tasks [17]. The proposed method could be thought of as a synchronous adaptor grammar 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 Ref. [8] that is applicable to recursive grammars such as HIER. We plan to examine variational inference for the proposed models in future work.

8. Experimental Evaluation

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

8.1 Experimental Setup

The data for French, German, and Spanish are from the 2010 Workshop on Statistical Machine Translation [4]. We use the news commentary corpus for training the phrase table, and the news commentary and Europarl corpora for training the LM. For Japanese, we use data from the NTCIR patent translation task [14]. We use the first 100 k sentences of the parallel corpus for the phrase table, 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 phrase table training. For both tasks, we perform weight tuning and testing on specified development and test sets. As an evaluation measure, we use case-insensitive BLEU score [30], a widely used evaluation metric for machine translation.

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

Table 1 The number of words in each corpus for phrase table (PT) and LM training, tuning, and testing.

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

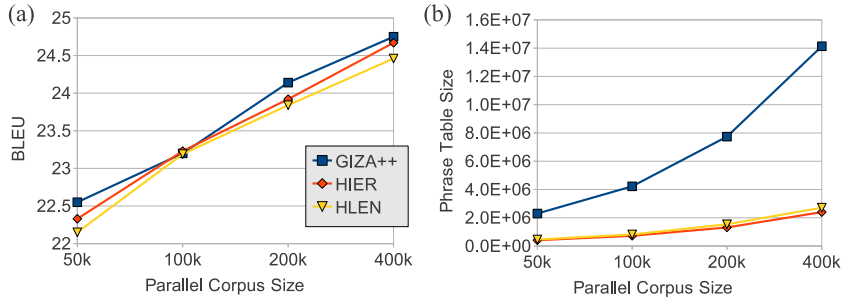

Fig. 8 The effect of corpus size on the accuracy (a) and phrase table size (b) for each method (Japanese-English).

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 ($p < 0.05$) [9]. GIZA++ uses HEUR-W for phrase extraction and all other models use MOD.

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

GIZA++ [28] and heuristic phrase extraction. Translation is performed using the Moses phrase-based machine translation decoder [19] using the phrase tables learned by each method under consideration. Phrase reordering probabilities are calculated using Moses's standard lexicalized reordering model [18] for all experimental settings. 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 the `grow-diag-final-and` combination heuristic [18]. 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 [*5]. In addition, we also try averaging the phrase tables from the last ten samples as described in Section 6.3.

8.2 Experimental Results

The results for these experiments can be found in **Table 2**. From these results it can be seen 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, as we examine further in Section 8.3, the assumption that phrase lengths are generated from a uniform distribution is likely too strong, and further gains could possibly be achieved by more accurate modeling of phrase lengths.

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.

8.2.1 Effect of Corpus Size

In order to ensure that the proposed method works well at all data sizes, we also performed an experiment varying the size of the training corpus. As there are not large amounts of in-domain data for the WMT news commentary task, we performed these experiments only on the patent task, varying the number of training sentences from 50 k to 400 k. The accuracy results are shown in **Fig. 8** (a). It can be seen that the results are largely consistent across all data sizes over, with statistically insignificant differences between HIER and GIZA++, and HLEN lagging slightly behind HIER. Figure 8 (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

[*5] 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 provided by P_{base}. As iterations took 1.3 hours on a single processor, good translation results were achieved in approximately 13 hours, which could further be reduced using distributed sampling [2], [24].

Table 5 Examples of phrases that exist only in GIZA or ITG-based models.

GIZA only		ITG only	
our réduire les	in reducing the	sensationnalisme	sensational
perceptuel implique une	implies a	tapageur	flashy
élections est	elections	évolué	moving
des attentes qui	the expectations that	dégénèrent	degenerate
vanterait	might	inscrire	enroll

Table 6 Phrase pairs that are used more often by GIZA than HIER. #GIZA and #HIER are the number of times the phrase was used by each system.

Source	Target	#GIZA	#HIER	HIER Phrase
les	the	529	475	*with noun*
qu’	that	74	38	*with verb*
: ”	: ”	33	0	*separate*
c’ est	it is	32	0	*separate*
opérateur	opérateur	32	0	operator
de la	the	33	2	of
2010 .	2010 .	2	0	*separate*
, ou alors	, or	2	0	*separate*
qui sont	who are	2	0	*separate*
il nous	we	2	0	*with comma*
travaillait	”	2	0	depravity (correct: “was working”)

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

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

Table 4 Overlap of phrase tables. The numbers indicate the percentage of the phrase table in the column that is also included in the phrase table in the row.

	GIZA	FLAT	HIER	HLEN
GIZA	-	40.46%	47.94%	41.54%
FLAT	1.68%	-	14.84%	12.51%
HIER	9.24%	68.72%	-	31.61%
HLEN	9.59%	69.40%	37.89%	-

corpus sizes.

8.2.2 Phrase Alignment/Heuristic Extraction

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 6. 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.

8.3 Acquired Phrases

In addition, we performed a quantitative and qualitative analysis of the phrase tables acquired using GIZA, FLAT, HIER, and HLEN for the French-English task. Phrase extraction was performed with HEUR-W for GIZA and MOD for all other alignment methods.

First, we performed an analysis of how much overlap there was between the extracted phrase tables, the results of which are shown in **Table 4**. Interestingly, the GIZA phrase table only covers approximately 40–50% of the acquired phrases in each of the ITG models, despite being much larger. To help understand the difference between GIZA and the ITG-based methods, we examined the phrases that occurred in only the GIZA phrase table, as well as the phrases that occurred in all of the other phrase tables, but not the GIZA phrase table (**Table 5**). For words found by all of the ITG models but not GIZA, the majority were rare single-word translations that were misaligned by GIZA due to the “garbage-collecting” phenomenon, where rare words are aligned to too many words, and thus not extracted properly. Among the shorter phrases that were found by none of the ITG models, but found by GIZA, most were the combination of one or several content words with a preposition.

In addition, we show examples of the phrases that are not just included in the phrase table, but actually used in translation by GIZA and HIER, focusing on phrase pairs that were used much more often by one system than the other, as well as less frequent phrase pairs that were used by one system twice, but the other system no times. Phrase pairs more commonly used in the GIZA and HIER systems are shown in **Table 6** and **Table 7** respectively. It can be seen that as an overall trend, the GIZA system tends to translate function words and punctuation in phrases together with the neighboring words, while the HIER system tends to translate these words separately, reflecting previous observations about the composition of the respective phrase tables. This combination of function words and punctuation into longer phrases does not change translation results, but increases the size of the phrase table, lending a convincing explanation for why HIER is able to achieve translation results that match GIZA with a smaller phrase table.

The next most common case were examples where one of the two systems dropped a frequent word in a multi-word translation (such as “de la” above). This was a problem for both systems, and there was not a clear trend favoring either system in these cases. In addition, both GIZA and HIER see words that are unknown for one of the two systems (“opérateur” and “communiqué” respectively) due to missed alignments preventing the creation of phrases for rarer words. Overall, GIZA was able to successfully generate phrases for fewer words, resulting in a total of 4,738 untranslatable words in the test set compared to 3,843 for HIER.

Table 7 Phrase pairs that are used more often by HIER than GIZA.

Source	Target	#GIZA	#HIER	GIZA Phrase
,	,	2,061	2,833	*with word*
de	of	685	1,366	*with noun*
.	.	1,443	2,002	*with word*
la	the	495	820	*with noun*
le	the	574	795	*with noun*
définitif	final	0	2	done, means
communiqué	communiqué	0	2	declaration, communicated
fréquente	frequent	0	2	frequently
connaissant	surplus	0	2	moreover (correct: "knowing")
moment où	moment when	0	2	*with "the"*

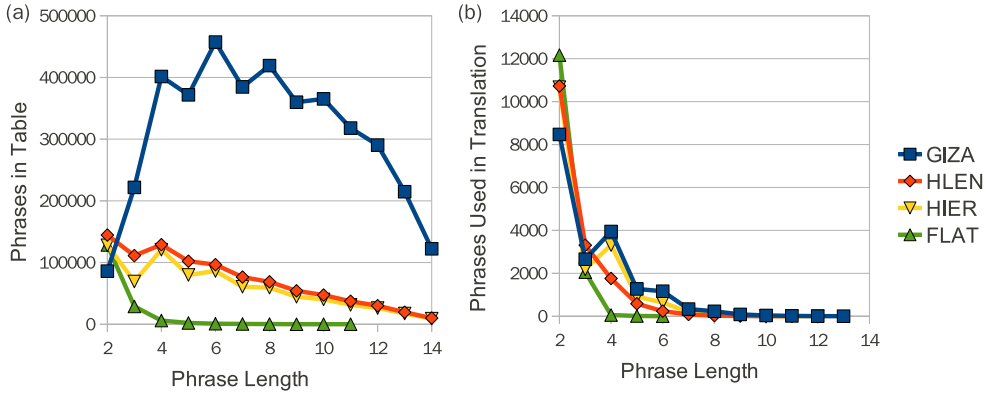

Fig. 9 The distribution of unique phrases by length (a) included in the phrase table and (b) used in translation.

Finally, there were a number of examples where equally valid translations with different lexical choice ("définitif") and syntactic form ("fréquente"), as well as examples where neither system was able to create a translation correctly ("travaillait" and "connaissant").

Finally, **Fig. 9** shows a break-down by length of the phrases in the acquired phrase table, as well as of the phrases that were actually used during translation. ¿From this graph it can be seen that GIZA creates large numbers of long phrases of length 6 or higher, despite the fact that the majority of used phrases are of length 2 or 4 (for 1-to-1 or 2-to-2 translations respectively). In general the distribution of phrases used by HIER in translation is similar to that of GIZA (with a slight tendency towards using shorter phrases), but the overall distribution of extracted phrases decreases gradually with length. FLAT, as expected, tends to both extract and use very short phrases.

Comparing HIER and HLEN, it can be seen that their patterns are largely similar, with the exception of phrases of length 3. Phrases of length 3 must be 1-to-2 or 2-to-1, and thus should be less common for language pairs such as English and French where one word tends to correspond to one word. HLEN creates more 3-word patterns than HIER because each length of phrase is given its own unique phrase distribution. Specifically, if $P_{t,3}$ has fewer phrases than $P_{t,2}$ and $P_{t,4}$, it also has more probability to "give away" to new 3-word phrases, creating an implicit bias towards creating more new phrases of less common phrase lengths. It is possible that this bias can be corrected by introducing priors that prefer phrase pairs where the number of words is roughly equal on both sides. However, this will require significant expansions to the current generative story, which does not explicitly keep track of the number of words on each side, so we leave this to future work.

9. Conclusion

In this paper, we presented a novel approach to joint phrase alignment and extraction through a hierarchical model using nonparametric 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.

One of the advantages of the proposed model is that it lends itself to relatively simple extension, allowing for further gains in accuracy through the introduction of more sophisticated alignment models. One promising future direction is the introduction of more intelligent prior knowledge through the base measure P_{base}. For example, P_{base} could be adjusted to take into account spelling similarities, parts of speech, phrase-based translation dictionaries, or bilingually acquired classes such as those proposed by Ref. [25]. We also 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 [6], or in alignment using syntactic structure [15]. It is also worth examining the plausibility of variational inference as proposed by Ref. [8] in the alignment context.

Reference

[1] Blunsom, P. and Cohn, T.: Inducing synchronous grammars with slice sampling, *Proc. Human Language Technology: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp.238–241 (2010).

[2] Blunsom, P., Cohn, T., Dyer, C. and Osborne, M.: A Gibbs sampler for phrasal synchronous grammar induction, *Proc. 47th Annual Meeting of the Association for Computational Linguistics*, pp.782–790 (2009).

[3] Brown, P.F., Pietra, V.J., Pietra, S.A.D. and Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol.19, pp.263–311 (1993).

[4] Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M. and Zaidan, O.F.: Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation, *Proc. Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pp.17–53 (2010).

[5] Cherry, C. and Lin, D.: Inversion Transduction Grammar for Joint Phrasal Translation Modeling, *Proc. NAACL Workshop on Syntax and Structure in Machine Translation*, pp.17–24 (2007).

[6] Chiang, D.: Hierarchical phrase-based translation, *Computational Linguistics*, Vol.33, No.2, pp.201–228 (2007).

[7] Chomsky, N.: Three models for the description of language, *IRE Trans. Inf. Theory*, Vol.2, No.3, pp.113–124 (1956).

[8] Cohen, S.B., Blei, D.M. and Smith, N.A.: Variational inference for adaptor grammars, *Proc. Human Language Technology: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp.564–572 (2010).

[9] Collins, M., Koehn, P. and Kučerová, I.: Clause restructuring for statistical machine translation, *Proc. 43rd Annual Meeting of the Association for Computational Linguistics*, pp.531–540 (2005).

[10] de Marcken, C.: Unsupervised Language Acquisition, PhD Thesis, Massachusetts Institute of Technology (1996).

[11] DeNero, J., Bouchard-Côté, A. and Klein, D.: Sampling alignment structure under a Bayesian translation model, *Proc. Conference on Empirical Methods in Natural Language Processing*, pp.314–323 (2008).

[12] DeNero, J., Gillick, D., Zhang, J. and Klein, D.: Why generative phrase models underperform surface heuristics, *Proc. 1st Workshop on Statistical Machine Translation*, pp.31–38 (2006).

[13] DeNero, J. and Klein, D.: Discriminative modeling of extraction sets for machine translation, *Proc. 48th Annual Meeting of the Association for Computational Linguistics*, pp.1453–1463 (2010).

[14] Fujii, A., Utiyama, M., Yamamoto, M. and Utsuro, T.: Overview of the patent translation task at the NTCIR-7 workshop, *Proc. 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp.389–400 (2008).

[15] Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W. and Thayer, I.: Scalable inference and training of context-rich syntactic translation models, *Proc. 44th Annual Meeting of the Association for Computational Linguistics*, pp.961–968 (2006).

[16] Johnson, J.H., Martin, J., Foster, G. and Kuhn, R.: Improving translation quality by discarding most of the phrasetable, *Proc. Conference on Empirical Methods in Natural Language Processing*, pp.967–975 (2007).

[17] Johnson, M., Griffiths, T.L. and Goldwater, S.: Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models, *Advances in Neural Information Processing Systems*, Vol.19, pp.641–648 (2007).

[18] Koehn, P., Axelrod, A., Mayne, A.B., Callison-Burch, C., Osborne, M. and Talbot, D.: Edinburgh system description for the 2005 IWSLT speech translation evaluation, *Proc. International Workshop on Spoken Language Translation* (2005).

[19] Koehn, P. et al.: Moses: Open Source Toolkit for Statistical Machine Translation, *Proc. 45th Annual Meeting of the Association for Computational Linguistics*, pp.177–180 (2007).

[20] Koehn, P., Och, F.J. and Marcu, D.: Statistical phrase-based translation, *Proc. Human Language Technology Conference* (*HLT-NAACL*), pp.48–54 (2003).

[21] Liang, P., Taskar, B. and Klein, D.: Alignment by agreement, *Proc. Human Language Technology Conference – North American Chapter of the Association for Computational Linguistics Annual Meeting* (*HLT-NAACL*), pp.104–111 (2006).

[22] Marcu, D. and Wong, W.: A phrase-based, joint probability model for statistical machine translation, *Proc. Conference on Empirical Methods in Natural Language Processing*, pp.133–139 (2002).

[23] Moore, R.C. and Quirk, C.: An iteratively-trained segmentation-free phrase translation model for statistical machine translation, *Proc. 2nd Workshop on Statistical Machine Translation*, pp.112–119 (2007).

[24] Newman, D., Asuncion, A., Smyth, P. and Welling, M.: Distributed algorithms for topic models, *Journal of Machine Learning Research*, Vol.10, pp.1801–1828 (2009).

[25] Och, F.J.: An efficient method for determining bilingual word classes, *Proc. 9th European Chapter of the Association for Computational Linguistics*, pp.71–76 (1999).

[26] Och, F.J.: Minimum Error Rate Training in Statistical Machine Translation, *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, pp.160–167 (2003).

[27] Och, F.J. and Ney, H.: Discriminative training and maximum entropy models for statistical machine translation, *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pp.295–302 (2002).

[28] Och, F.J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol.29, No.1, pp.19–51 (2003).

[29] Och, F.J., Tillmann, C. and Ney, H.: Improved alignment models for statistical machine translation, *Proc. 4th Conference on Empirical Methods in Natural Language Processing*, pp.20–28 (1999).

[30] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation, *Proc. 19th International Conference on Computational Linguistics*, pp.311–318 (2002).

[31] Pitman, J. and Yor, M.: The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator, *The Annals of Probability*, Vol. 25, No. 2, pp. 855–900 (1997).

[32] Saers, M., Nivre, J. and Wu, D.: Learning Stochastic Bracketing Inversion Transduction Grammars with a Cubic Time Biparsing Algorithm, *Proc. 11th International Workshop on Parsing Technologies*, pp.29–32 (2009).

[33] Teh, Y.W.: A hierarchical Bayesian language model based on Pitman-Yor processes, *Proc. 44th Annual Meeting of the Association for Computational Linguistics*, pp.985–992 (2006).

[34] Wu, D.: Stochastic inversion transduction grammars and bilingual parsing of parallel corpora, *Computational Linguistics*, Vol.23, No.3, pp.377–403 (1997).

[35] Zhang, H., Quirk, C., Moore, R.C. and Gildea, D.: Bayesian learning of non-compositional phrases with synchronous parsing, *Proc. 46th Annual Meeting of the Association for Computational Linguistics*, pp.97–105 (2008).

Graham Neubig received his B.S. from University of Illinois, Urbana-Champaign, U.S.A, in 2005, and his M.E. in informatics from Kyoto University, Kyoto, Japan in 2010, where he is currently pursuing his Ph.D. He is a recipient of the JSPS Research Fellowship for Young Scientists (DC1). His research interests include speech and natural language processing, with a focus on unsupervised learning for applications such as automatic speech recognition and machine translation.

Taro Watanabe received his B.E. and M.E. degrees in information science from Kyoto University, Kyoto, Japan in 1994 and 1997, respectively, and obtained M.S. degree in language and information technologies from the School of Computer Science, Carnegie Mellon University in 2000. In 2004, he received his Ph.D. in informatics from Kyoto University, Kyoto, Japan. After serving as a researcher at ATR and NTT, Dr. Watanabe is a senior researcher at National Institute of Information and Communications Technology. His research interests include natural language processing, machine learning and statistical machine translation.

Eiichiro Sumita received his M.S. degree in computer science from the University of Electro-Communications in 1982 and Ph.D. degree in engineering from Kyoto University in 1999. Dr. Sumita is the director of multilingual translation laboratory of NICT and a visiting professor of Kobe University Graduate School. His research interests include machine translation and e-Learning.

Shinsuke Mori received his B.S., M.S., and Ph.D. degrees in electrical engineering from Kyoto University, Kyoto, Japan in 1993, 1995, and 1998, respectively. After joining Tokyo Research Laboratory of International Business Machines (IBM) in 1998, he studied the language model and its application to speech recognition and language processing. He is currently an associate professor of Academic Center for Computing and Media Studies, Kyoto University.

Tatsuya Kawahara received his B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. In 1990, he became a research associate in the Department of Information Science, Kyoto University. From 1995 to 1996, he was a visiting researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a professor in the Academic Center for Computing and Media Studies and an affiliated professor in the School of Informatics, Kyoto University. He has also been an invited researcher at ATR and NICT. He has published more than 200 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been managing several speech-related projects in Japan including a free large vocabulary continuous speech recognition software project (http://julius.sourceforge.jp/). Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. From 2003 to 2006, he was a member of IEEE SPS Speech Technical Committee. From 2011, he is a secretary of IEEE SPS Japan Chapter. He was a general chair of IEEE Automatic Speech Recognition & Understanding workshop (ASRU 2007). He also served as a tutorial chair of INTERSPEECH 2010. He is a senior member of IEEE.