To appear in *HLT-EMNLP 2005, Vancouver, Canada*

HMM Word and Phrase Alignment for Statistical Machine Translation

Yonggang Deng[1] **, William Byrne**[1,2]
Center for Language and Speech Processing, Johns Hopkins University
Baltimore, MD 21210, USA [1]
Machine Intelligence Lab, Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, UK [2]
dengyg@jhu.edu , wjb31@cam.ac.uk

Abstract

HMM-based models are developed for the alignment of words and phrases in bitext. The models are formulated so that alignment and parameter estimation can be performed efficiently. We find that Chinese-English word alignment performance is comparable to that of IBM Model-4 even over large training bitexts. Phrase pairs extracted from word alignments generated under the model can also be used for phrase-based translation, and in Chinese to English and Arabic to English translation, performance is comparable to systems based on Model-4 alignments. Direct phrase pair induction under the model is described and shown to improve translation performance.

1 Introduction

Describing word alignment is one of the fundamental goals of Statistical Machine Translation (SMT). Alignment specifies how word order changes when a sentence is translated into another language, and given a sentence and its translation, alignment specifies translation at the word level. It is straightforward to extend word alignment to phrase alignment: two phrases align if their words align.

Deriving phrase pairs from word alignments is now widely used in phrase-based SMT. Parameters of a statistical word alignment model are estimated from bitext, and the model is used to generate word alignments over the same bitext. Phrase pairs are extracted from the aligned bitext and used in the SMT system. With this approach the quality of the underlying word alignments can have a strong influence on phrase-based SMT system performance. The common practice therefore is to extract phrase pairs from the best attainable word alignments. Currently, Model-4 alignments (Brown and others, 1993) as produced by GIZA++ (Och and Ney, 2000) are often the best that can be obtained, especially with large bitexts.

Despite its modeling power and widespread use, Model-4 has shortcomings. Its formulation is such that maximum likelihood parameter estimation and bitext alignment are implemented by approximate, hill-climbing, methods. Consequently parameter estimation can be slow, memory intensive, and difficult to parallelize. It is also difficult to compute statistics under Model-4. This limits its usefulness for modeling tasks other than the generation of word alignments.

We describe an HMM alignment model developed as an alternative to Model-4. In the word alignment and phrase-based translation experiments to be presented, its performance is comparable or improved relative to Model-4. Practically, we can train the model by the Forward-Backward algorithm, and by parallelizing estimation, we can control memory usage, reduce the time needed for training, and increase the bitext used for training. We can also compute statistics under the model in ways not practical with Model-4, and we show the value of this in the extraction of phrase pairs from bitext.

2 HMM Word and Phrase Alignment

Our goal is to develop a generative probabilistic model of Word-to-Phrase (WtoP) alignment. We start with an l-word source sentence $\mathbf{e} = e_1^l$, and an

m-word target sentence $\mathbf{f} = f_1^m$, which is realized as a sequence of K phrases: $\mathbf{f} = v_1^K$.

Each phrase is generated as a translation of one source word, which is determined by the alignment sequence a_1^K: $e_{a_k} \rightarrow v_k$. The length of each phrase is specified by the process ϕ_1^K, which is constrained so that $\sum_{k=1}^K \phi_k = m$.

We also allow target phrases to be inserted, i.e. to be generated by a NULL source word. For this, we define a binary hallucination sequence h_1^K : if $h_k = 0$, then NULL $\rightarrow v_k$; if $h_k = 1$ then $e_{a_k} \rightarrow v_k$.

With all these quantities gathered into an alignment $\mathbf{a} = (\phi_1^K, a_1^K, h_1^K, K)$, the modeling objective is to realize the conditional distribution $P(\mathbf{f}, \mathbf{a}|\mathbf{e})$. With the assumption that $P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = 0$ if $\mathbf{f} \neq v_1^K$, we write $P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = P(v_1^K, K, a_1^K, h_1^K, \phi_1^K|\mathbf{e})$ and

$$\begin{aligned} &P(v_1^K, K, a_1^K, h_1^K, \phi_1^K|\mathbf{e}) \\ &\quad = \epsilon(m|l) \times P(K|m, \mathbf{e}) \\ &\qquad \times P(a_1^K, \phi_1^K, h_1^K|K, m, \mathbf{e}) \\ &\qquad\quad \times P(v_1^K|a_1^K, h_1^K, \phi_1^K, K, m, \mathbf{e}) \end{aligned}$$

We now describe the component distributions.

Sentence Length $\epsilon(m|l)$ determines the target sentence length. It is not needed during alignment, where sentence lengths are known, and is ignored.

Phrase Count $P(K|m, \mathbf{e})$ specifies the number of target phrases. We use a simple, single parameter distribution, with $\eta = 8.0$ throughout

$$P(K|m, \mathbf{e}) = P(K|m, l) \propto \eta^K$$

Word-to-Phrase Alignment Alignment is a Markov process that specifies the lengths of phrases and their alignment with source words

$$\begin{aligned} &P(a_1^K, h_1^K, \phi_1^K|K, m, \mathbf{e}) \\ &= \prod_{k=1}^K P(a_k, h_k, \phi_k|a_{k-1}, \phi_{k-1}, \mathbf{e}) \\ &= \prod_{k=1}^K p(a_k|a_{k-1}, h_k; l)\, d(h_k)\, n(\phi_k; e_{a_k}) \end{aligned}$$

The actual word-to-phrase alignment (a_k) is a first-order Markov process, as in HMM-based word-to-word alignment (Vogel et al., 1996). It necessarily depends on the hallucination variable

$$\begin{aligned} &p(a_j|a_{j-1}, h_j; l) \\ &\quad = \begin{cases} 1 & a_j = a_{j-1},\ h_j = 0 \\ 0 & a_j \neq a_{j-1},\ h_j = 0 \\ a(a_j|a_{j-1}; l) & h_j = 1 \end{cases} \end{aligned}$$

This formulation allows target phrases to be inserted without disrupting the Markov dependencies of phrases aligned to actual source words.

The phrase length model $n(\phi; e)$ gives the probability that a word e produces a phrase with ϕ words in the target language; $n(\phi; e)$ is defined for $\phi = 1, \cdots, N$. The hallucination process is a simple i.i.d. process, where $d(0) = p_0$, and $d(1) = 1 - p_0$.

Word-to-Phrase Translation The translation of words to phrases is given as

$$P(v_1^K|a_1^K, h_1^K, \phi_1^K, K, m, \mathbf{e}) = \prod_{k=1}^K p(v_k|e_{a_k}, h_k, \phi_k)$$

We introduce the notation $v_k = v_k[1], \ldots, v_k[\phi_k]$ and a dummy variable x_k (for phrase insertion) :

$$x_k = \begin{cases} e_{a_k} & h_k = 1 \\ \text{NULL} & h_k = 0 \end{cases}$$

We define two models of word-to-phrase translation. This simplest is based on context-independent word-to-word translation

$$p(v_k|e_{a_k}, h_k, \phi_k) = \prod_{j=1}^{\phi_k} t(v_k[j] \,|\, x_k)$$

We also define a model that captures foreign word context with *bigram translation probabilities*

$$\begin{aligned} &p(v_k|e_{a_k}, h_k, \phi_k) \\ &\quad = t(v_k[1] \,|\, x_k) \prod_{j=2}^{\phi_k} t_2(v_k[j] \,|\, v_k[j-1], x_k) \end{aligned}$$

Here, $t(f|e)$ is the usual context independent word-to-word translation probability. The bigram translation probability $t_2(f|f', e)$ specifies the likelihood that target word f is to follow f' in a phrase generated by source word e.

2.1 Properties of the Model and Prior Work

The formulation of the WtoP alignment model was motivated by both the HMM word alignment model (Vogel et al., 1996) and IBM Model-4 with the goal of building on the strengths of each.

The relationship with the word-to-word HMM alignment model is straightforward. For example, constraining the phrase length component $n(\phi; e)$ to permit only phrases of one word would give a word-to-word HMM alignment model. The extensions introduced are the phrase count, and the phrase length models, and the bigram translation distribution. The hallucination process is motivated by the use of NULL alignments into Markov alignment models as done by (Och and Ney, 2003).

The phrase length model is motivated by Toutanova et al. (2002) who introduced 'stay' probabilities in HMM alignment as an alternative to word fertility. By comparison, Word-to-Phrase HMM alignment models contain detailed models of state occupancy, motivated by the IBM fertility model, which are more powerful than a single staying parameter. In fact, the WtoP model is a segmental Hidden Markov Model (Ostendorf et al., 1996), in which states emit observation sequences.

Comparison with Model-4 is less straightforward. The main features of Model-4 are NULL source words, source word fertility, and the distortion model. The WtoP alignment model includes the first two of these. However distortion, which allows hypothesized words to be distributed throughout the target sentence, is difficult to incorporate into a model that supports efficient DP-based search. We preserve efficiency in the WtoP model by insisting that target words form connected phrases; this is not as general as Model-4 distortion. This weakness is somewhat offset by a more powerful (Markov) alignment process as well as by the phrase count distribution. Despite these differences, the WtoP alignment model and Model-4 allow similar alignments. For example, in Fig. 1, Model-4 would allow

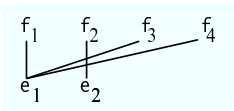

Figure 1: Word-to-Word and Word-to-Phrase Links

f_1, f_3, and f_4 to be generated by e_1 with a fertility of 3. Under the WtoP model, e_1 could generate f_1 and $f_3 f_4$ with phrase lengths 1 and 2, respectively: source words can generate more than one phrase.

This alignment could also be generated via four single word foreign phrases. The balance between word-to-word and word-to-phrase alignments is set by the phrase count distribution parameter η. As η increases, alignments with shorter phrases are favored, and for very large η the model allows only word-to-word alignments (see Fig. 2). Although the WtoP alignment model is more complex than the word-to-word HMM alignment model, the Baum-Welch and Viterbi algorithms can still be used. Word-to-word alignments are generated by the Viterbi algorithm: $\hat{\mathbf{a}} = \text{argmax}_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e})$; if $e_{a_k} \rightarrow v_k$, e_{a_k} is linked to all the words in v_k.

The bigram translation probability relies on word context, known to be helpful in translation (Berger et al., 1996), to improve the identification of target phrases. As an example, f is the Chinese word for "world trade center". Table 1 shows how the likelihood of the correct English phrase is improved with bigram translation probabilities; this example is from the C→E, N=4 system of Table 2.

Model	unigram	bigram
P(world$\|f$)	0.06	0.06
P(trade|world, f)	0.06	0.99
P(center|trade, f)	0.06	0.99
P(world trade center$\|f, 3$)	0.0002	0.0588

Table 1: Context in Bigram Phrase Translation.

There are of course much prior work in translation that incorporates phrases. Sumita et al. (2004) develop a model of phrase-to-phrase alignment, which while based on HMM alignment process, appears to be deficient. Marcu and Wong (2002) propose a model to learn lexical correspondences at the phrase level. To our knowledge, ours is the first non-syntactic model of bitext alignment (as opposed to translation) that links words and phrases.

3 Embedded Alignment Model Estimation

We now discuss estimation of the WtoP model parameters by the EM algorithm. Since the WtoP model can be treated as an HMM with a very complex state space, it is straightforward to apply Baum-

Welch parameter estimation. We show the forward recursion as an example.

Given a sentence pair (e_1^l, f_1^m), the forward probability $\alpha_j(i,\phi)$ is defined as the probability of generating the first j target words with the added condition that the target words $f_{j-\phi+1}^{j}$ form a phrase aligned to source word e_i. It can be calculated recursively (omitting the hallucination process, for simplicity) as

$$\alpha_j(i,\phi) = \Big\{ \sum_{i',\phi'} \alpha_{j-\phi}(i',\phi') a(i|i',l) \Big\} \cdot \eta$$

$$\cdot\, n(\phi; e_i) \cdot t(f_{j-\phi+1}|e_i) \cdot \prod_{j'=j-\phi+2}^{j} t_2(f_{j'}|e_i)$$

This recursion is over a trellis of $l(N+1)m$ nodes.

Models are trained from a flat-start. We begin with 10 iterations of EM to train Model-1, followed by 5 EM iterations to train Model-2 (Brown and others, 1993). We initialize the parameters of the word-to-word HMM alignment model by collecting word alignment counts from the Model-2 Viterbi alignments, and refine the word-to-word HMM alignment model by 5 iterations of the Baum-Welch algorithm. We increase the order of the WtoP model (N) from 2 to the final value in increments of 1, by performing 5 Baum Welch iterations at each step. At the final value of N, we introduce the bigram translation probability; we use Witten-Bell smoothing (1991) as a backoff strategy for t_2, and other strategies are possible.

4 Bitext Word Alignment

We now investigate bitext word alignment performance. We start with the FBIS Chinese/English parallel corpus which consists of approx. 10M English/7.5M Chinese words. The Chinese side of the corpus is segmented into words by the LDC segmenter[1]. The alignment test set consists of 124 sentences from the NIST 2001 dry-run MT-eval[2] set that are manually word aligned.

We first analyze the distribution of word links within these manual alignments. Of the Chinese words which are aligned to more than one English words, 82% of these words align with consecutive English words (phrases). In the other direction, among all English words which are aligned to multiple Chinese words, 88% of these align to Chinese phrases. In this collection, at least, word-to-phrase alignments are plentiful.

[1] http://www.ldc.upenn.edu/Projects/Chinese

[2] http://www.nist.gov/speech/tests/mt

Model	AER_{1-1}	AER_{1-N}	AER
	C⟶E		
Model-4	37.9	68.3	37.3
HMM, N=1	42.8	72.9	42.0
HMM, N=2	38.3	71.2	38.1
HMM, N=3	37.4	69.5	37.8
HMM, N=4	37.1	69.1	37.8
+ bigram t-table	37.5	65.8	37.1
	E⟶C		
Model-4	42.3	87.2	45.0
HMM, N=1	45.0	90.6	47.2
HMM, N=2	42.7	87.5	44.5
+ bigram t-table	44.2	85.5	45.1

Table 2: FBIS Bitext Alignment Error Rate.

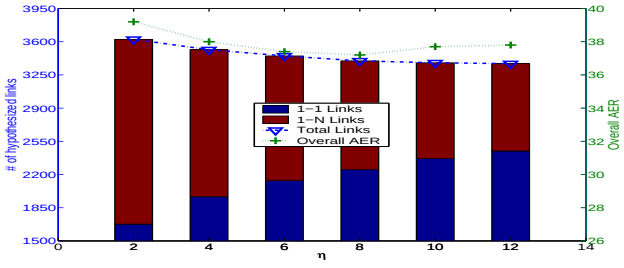

Figure 2: Balancing Word and Phrase Alignments

Alignment performance is measured by the Alignment Error Rate (AER) (Och and Ney, 2003)

$$AER(B;B') = 1 - 2 \times |B \cap B'| / (|B'| + |B|)$$

where B is a set reference word links, and B' are the word links generated automatically.

AER gives a general measure of word alignment quality. We are also interested in how the model performs over the word-to-word and word-to-phrase alignments it supports. We split the reference alignments into two subsets: B_{1-1} contains word-to-word reference links (e.g. 1⟶1 in Fig 1); and B_{1-N} contains word-to-phrase reference links (e.g. 1⟶3, 1⟶4 in Fig 1); The automatic alignment B' is partitioned similarly. We define additional AERs: $AER_{1-1} = AER(B_{1-1}, B'_{1-1})$, and $AER_{1-N} = AER(B_{1-N}, B'_{1-N})$, which measure word-to-word and word-to-phrase alignment, separately.

Table 2 presents the three AER measurements for

the WtoP alignment models trained as described in Section 3. GIZA++ Model 4 alignment performance is also presented for comparison. We note first that the word-to-word HMM (N=1) alignment model is worse than Model 4, as expected. For the WtoP models in the C→E direction, we see reduced AER for phrases lengths up to 4, although in the E→C direction, AER is reduced only for phrases of length 2; performance for $N > 2$ is not reported.

In introducing the bigram phrase translation (the bigram t-table), there is a tradeoff between word-to-word and word-to-phrase alignment quality. As mentioned, the bigram t-table increases the likelihood of word-to-phrase alignments. In both translation directions, this reduces the AER_{1-N}. However, it also causes increases in AER_{1-1}, primarily due to a drop in recall: fewer word-to-word alignments are produced. For C→E, this is not severe enough to cause an overall AER increase; however, in E→C, AER does increase.

Fig. 2 (C→E, N=4) shows how the 1-1 and 1-N alignment behavior is balanced by the phrase count parameter. As η increases, the model favors alignments with more word-to-word links and fewer word-to-phrase links; the overall Alignment Error Rate (AER) suggests a good balance at $\eta = 8.0$.

After observing that the WtoP model performs as well as Model-4 over the FBIS C-E bitext, we investigated performance over these large bitexts :
- "NEWS" containing non-UN parallel Chinese/English corpora from LDC (mainly FBIS, Xinhua, Hong Kong, Sinorama, and Chinese Treebank).
- "NEWS+UN01-02" also including UN parallel corpora from the years 2001 and 2002.
- "ALL C-E" refers to all the C-E bitext available from LDC as of his submission; this consists of the NEWS corpora with the UN bitext from all years.

Over all these collections, WtoP alignment performance (Table 3) is comparable to that of Model-4. We do note a small degradation in the E→C WtoP alignments. It is quite possible that this one-to-many model suffers slightly with English as the source and Chinese as the target, since English sentences tend to be longer. Notably, simply increasing the amount of bitext used in training need not improve AER. However, larger aligned bitexts can give improved phrase pair coverage of the test set.

One of the desirable features of HMMs is that the Forward-Backward steps can be run in parallel: bitext is partitioned; the Forward-Backward algorithm is run over the subsets on different CPUs; statistics are merged to reestimate model parameters. Partitioning the bitext also reduces the memory usage, since different cooccurrence tables can be kept for each partition. With the "ALL C-E" bitext collection, a single set of WtoP models (C→E, N=4, bigram t-table) can be trained over 200M words of Chinese-English bitext by splitting training over 40 CPUs; each Forward-Backward process takes less than 2GB of memory and the training run finishes in five days. By contrast, the 96M English word NEWS+UN01-02 is about the largest C-E bitext over which we can train Model-4 with our GIZA++ configuration and computing infrastructure.

Bitext	English Words	Model	C→E	E→C
NEWS	71M	M-4	37.1	45.3
		WtoP	36.1	44.8
NEWS+ UN01-02	96M	M-4	36.1	43.4
		WtoP	36.4	44.2
ALL C-E	200M	WtoP	36.8	44.7

Table 3: AER Over Large C-E Bitexts.

Based on these and other experiments, in this paper we set a maximum value of $N = 4$ for F→E; in E→F, we set N=2 and omit the bigram phrase translation probability; η is set to 8.0. We do not claim that this is optimal, however.

5 Phrase Pair Induction

A common approach to phrase-based translation is to extract an inventory of phrase pairs (PPI) from bitext (Koehn et al., 2003), For example, in the *phrase-extract* algorithm (Och, 2002), a word alignment $\hat{a}_1^m$ is generated over the bitext, and all word subsequences $e_{i_1}^{i_2}$ and $f_{j_1}^{j_2}$ are found that satisfy :

$$\hat{a}_1^m : \hat{a}_j \in [i_1, i_2] \text{ iff } j \in [j_1, j_2] \ . \qquad (1)$$

The PPI comprises all such phrase pairs $(e_{i_1}^{i_2}, f_{j_1}^{j_2})$.

The process can be stated slightly differently. First, we define a set of alignments :

$$A(i_1, i_2; j_1, j_2) = \{a_1^m : a_j \in [i_1, i_2] \text{ iff } j \in [j_1, j_2]\} \ .$$

If $\hat{a}_1^m \in A(i_1, i_2; j_1, j_2)$ then $(e_{i_1}^{i_2}, f_{j_1}^{j_2})$ form a phrase pair.

Viewed in this way, there are many possible alignments under which phrases might be paired, and

the selection of phrase pairs need not be based on a single alignment. Rather than simply accepting a phrase pair $(e_{i_1}^{i_2}, f_{j_1}^{j_2})$ if the unique MAP alignment satisfies Equation 1, we can assign a probability to phrases occurring as translation pairs :

$$P(\mathbf{f},\, A(i_1, i_2; j_1, j_2)\,|\,\mathbf{e}) = \sum_{\mathbf{a}\,:\,a_1^m \in A(i_1,i_2;j_1,j_2)} P(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

For a fixed set of indices i_1, i_2, j_1, j_2, the quantity $P(\mathbf{f},\, A(i_1, i_2; j_1, j_2)\,|\,\mathbf{e})$ can be computed efficiently using a modified Forward algorithm. Since $P(\mathbf{f}|\mathbf{e})$ can also be computed by the Forward algorithm, the *phrase-to-phrase posterior distribution* $P(\,A(i_1, i_2; j_1, j_2)\,|\,\mathbf{f}, \mathbf{e})$ is easily found.

PPI Induction Strategies In the *phrase-extract* algorithm (Och, 2002), the alignment $\hat{a}$ is generated as follows: Model-4 is trained in both directions (e.g. F→E and E→F); two sets of word alignments are generated by the Viterbi algorithm for each set of models; and the two alignments are merged. This forms a static aligned bitext. Next, all foreign word sequences up to a given length (here, 5 words) are extracted from the test set. For each of these, a phrase pair is added to the PPI if the foreign phrase can be found aligned to an English phrase under Eq 1. We refer to the result as the Model-4 Viterbi Phrase-Extract PPI.

Constructed in this way, the PPI is limited to phrase pairs which can be found in the Viterbi alignments. Some foreign phrases which do appear in the training bitext will not be included in the PPI because suitable English phrases cannot be found. To add these to the PPI we can use the phrase-to-phrase posterior distribution to find English phrases as candidate translations. This adds phrases to the Viterbi Phrase-Extract PPI and increase the test set coverage. A somewhat *ad hoc* PPI Augmentation algorithm is given to the right.

Condition (A) extracts phrase pairs based on the geometric mean of the E→F and F→E posteriors ($T_g = 0.01$ throughout). The threshold T_p selects additional phrase pairs under a more forgiving criterion: as T_p decreases, more phrase pairs are added and PPI coverage increases. Note that this algorithm is constructed specifically to improve a Viterbi PPI; it is certainly not the only way to extract phrase pairs under the phrase-to-phrase posterior distribution.

Once the PPI phrase pairs are set, the phrase translation probabilities are set based on the number of times each phrase pair is extracted from a sentence pair, i.e. from relative frequencies.

For each foreign phrase v not in the Viterbi PPI :
For all pairs (f_1^m, e_1^l) and j_1, j_2 s.t. $f_{j_1}^{j_2} = v$:
For $1 \le i_1 \le i_2 \le l$, find

$$f(i_1, i_2) = P_{F\to E}(\,A(i_1, i_2; j_1, j_2)\,|\,e_1^l, f_1^m)$$

$$b(i_1, i_2) = P_{E\to F}(\,A(i_1, i_2; j_1, j_2)\,|\,e_1^l, f_1^m)$$

$$g(i_1, i_2) = \sqrt{f(1_1, i_2)\, b(i_1, i_2)}$$

$$(\hat{i}_1, \hat{i}_2) = \underset{1\le i_1, i_2 \le l}{\operatorname{argmax}}\; g(i_1, i_2)\text{ , and set } u = e_{\hat{i}_1}^{\hat{i}_2}$$

Add (u, v) to the PPI if any of A, B, or C hold :

$b(\hat{i}_1, \hat{i}_2) \ge T_g$ and $f(\hat{i}_1, \hat{i}_2) \ge T_g$ (A)
$b(\hat{i}_1, \hat{i}_2) < T_g$ and $f(\hat{i}_1, \hat{i}_2) > T_p$ (B)
$f(\hat{i}_1, \hat{i}_2) < T_g$ and $b(\hat{i}_1, \hat{i}_2) > T_p$ (C)

PPI Augmentation via Phrase-Posterior Induction

HMM-based models are often used if posterior distributions are needed. Model-1 can also be used in this way (Venugopal et al., 2003), although it is a relatively weak alignment model. By comparison, finding posterior distributions under Model-4 is difficult. The Word-to-Phrase alignment model appears not to suffer this tradeoff: it is a good model of word alignment under which statistics such as the phrase-to-phrase posterior can be calculated.

6 Translation Experiments

We evaluate the quality of phrase pairs extracted from the bitext through the translation performance of the Translation Template Model (TTM) (Kumar et al., 2005), which is a phrase-based translation system implemented using weighted finite state transducers. Performance is measured by BLEU (Papineni and others, 2001).

Chinese→English Translation We report performance on the NIST Chinese/English 2002, 2003 and 2004 (News only) MT evaluation sets. These consist of 878, 919, and 901 sentences, respectively. Each Chinese sentence has 4 reference translations.

We evaluate two C→E translation systems. The smaller system is built on the FBIS C-E bitext collection. The language model used for this system is a trigram word language model estimated with 21M

	V-PE Model	WtoP T_p	eval02 cvg	eval02 BLEU	eval03 cvg	eval03 BLEU	eval04 cvg	eval04 BLEU	eval02 cvg	eval02 BLEU	eval03 cvg	eval03 BLEU	eval04 cvg	eval04 BLEU
			FBIS C→E System						News A→E System					
1	M-4	-	20.1	23.8	17.7	22.8	20.2	23.0	19.5	36.9	21.5	39.1	18.5	40.0
2		0.7	24.6	24.6	21.4	23.7	24.6	23.7	23.8	37.6	26.6	40.2	22.4	40.3
3	WtoP	-	19.7	23.9	17.4	23.3	19.8	23.3	18.4	36.2	20.6	38.6	17.4	39.2
4		1.0	23.1	24.0	20.0	23.7	23.2	23.5	21.8	36.7	24.3	39.3	20.4	39.7
5		0.9	24.0	24.8	20.9	23.9	24.0	23.8	23.2	37.2	25.8	39.7	21.8	40.1
6		0.7	24.6	24.9	21.3	24.0	24.7	23.9	23.7	37.2	26.5	39.7	22.4	39.9
7		0.5	24.9	24.9	21.6	24.1	24.8	23.9	24.0	37.2	26.9	39.7	22.7	39.8
			Large C→E System						Large A→E System					
8	M-4	-	32.5	27.7	29.3	27.1	32.5	26.6	26.4	38.1	28.1	40.1	28.2	39.9
9	WtoP	-	30.6	27.9	27.5	27.0	30.6	26.4	24.8	38.1	26.6	40.1	26.7	40.6
10		0.7	38.2	28.2	32.3	27.3	37.1	26.8	30.7	39.3	32.9	41.6	32.5	41.9

Table 4: Translation Analysis and Performance of PPI Extraction Procedures

words taken from the English side of the bitext; all language models are built with the SRILM toolkit using Kneser-Ney smoothing (Stolcke, 2002).

The larger system is based on alignments generated over all available C-E bitext (the "ALL C-E" collection of Section 4). The language model is an equal-weight interpolated trigram model trained over 373M English words taken from the English side of the bitext and the LDC Gigaword corpus.

Arabic→English Translation We also evaluate our WtoP alignment models in Arabic-English translation. We report results on a small and a large system. In each, Arabic text is tokenized by the Buckwalter analyzer provided by LDC. We test our models on NIST Arabic/English 2002, 2003 and 2004 (News only) MT evaluation sets that consists of 1043, 663 and 707 Arabic sentences, respectively. Each Arabic sentence has 4 reference translations.

In the small system, the training bitext is from A-E News parallel text, with ∼3.5M words on the English side. We follow the same training procedure and configurations as in Chinese/English system in both translation directions. The language model is an equal-weight interpolated trigram built over ∼400M words from the English side of the bitext, including UN text, and the LDC English Gigaword collection. The large Arabic/English system employs the same language model. Alignments are generated over all A-E bitext available from LDC as of this submission; this consists of approx. 130M words on the English side.

WtoP Model and Model-4 Comparison We first look at translation performance of the small A→E and C→E systems, where alignment models are trained over the smaller bitext collections. The baseline systems (Table 4, line 1) are based on Model-4 Viterbi Phrase-Extract PPIs.

We compare WtoP alignments directly to Model-4 alignments by extracting PPIs from the WtoP alignments using the Viterbi Phrase-Extract procedure (Table 4, line 3). In C→E translation, performance is comparable to that of Model-4; in A→E translation, performance lags slightly. As we add phrase pairs to the WtoP Viterbi Phrase-Extract PPI via the Phrase-Posterior Augmentation procedure (Table 4, lines 4-7), we obtain a ∼1% improvement in BLEU; the value of $T_p = 0.7$ gives improvements across all sets. In C→E translation, this yields good gains relative to Model-4, while in A→E we match or improve the Model-4 performance.

The performance gains through PPI augmentation are consistent with increased PPI coverage of the test set. We tabulate the percentage of test set phrases that appear in each of the PPIs (the 'cvg' values in Table 4). The augmentation scheme is designed specifically to increase coverage, and we find that BLEU score improvements track the phrase coverage of the test set. This is further confirmed by the experiment of Table 4, line 2 in which we take the PPI extracted from Model-4 Viterbi alignments, and add phrase pairs to it using the Phrase-Posterior augmentation scheme with $T_p = 0.7$. We find that the augmentation scheme under the WtoP models can be used to improve the Model-4 PPI itself.

We also investigate C→E and A→E translation performance with PPIs extracted from large bitexts.

Performance of systems based on Model-4 Viterbi Phrase-Extract PPIs is shown in Table 4, line 8. To train Model-4 using GIZA++, we split the bitexts into two (A-E) or three (C-E) partitions, and train models for each division separately; we find that memory usage is otherwise too great. These serve as a single set of alignments for the bitext, as if they had been generated under a single alignment model. When we translate with Viterbi Phrase-Extract PPIs taken from WtoP alignments created over all available bitext, we find comparable performance to the Model-4 baseline (Table 4, line 9). Using the Phrase-Posterior augmentation scheme with $T_p = 0.7$ yields further improvement (Table 4, line 10). Pooling the sets to form two large C→E and A→E test sets, the A→E system improvements are significant at a 95% level (Och, 2003); the C→E systems are only equivalent.

7 Conclusion

We have described word-to-phrase alignment models capable of good quality bitext word alignment. In Arabic-English and Chinese-English translation and alignment they compare well to Model-4, even with large bitexts. The model architecture was inspired by features of Model-4, such as fertility and distortion, but care was taken to ensure that dynamic programming procedures, such as EM and Viterbi alignment, could still be performed. There is practical value in this: training and alignment are easily parallelized. Working with HMMs also makes it straightforward to explore new modeling approaches. We show an augmentation scheme that adds to phrases extracted from Viterbi alignments; this improves translation with both the WtoP and the Model-4 phrase pairs, even though it would be infeasible to implement the scheme under Model-4 itself. We note that these models are still relatively simple, and we anticipate further alignment and translation improvement as the models are refined.

Acknowledgments The TTM translation system was provided by Shankar Kumar. This work was funded by ONR MURI Grant N00014-01-1-0685.

References

A. L. Berger, S. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

P. F. Brown et al. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19:263–312.

P. Koehn, F. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of HLT-NAACL*.

S. Kumar, Y. Deng, and W. Byrne. 2005. A weighted finite state transducer translation template model for statistical machine translation. *Journal of Natural Language Engineering*, 11(3).

D. Marcu and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. of EMNLP*.

F. Och and H. Ney. 2000. Improved statistical alignment models. In *Proc. of ACL*, Hong Kong, China.

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

F. Och. 2002. *Statistical Machine Translation: From Single Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen, Germany.

F. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*.

M. Ostendorf, V. Digalakis, and O. Kimball. 1996. From HMMs to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Trans. Acoustics, Speech and Signal Processing*, 4:360–378.

K. Papineni et al. 2001. BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. ICSLP*.

E. Sumita et al. 2004. EBMT, SMT, Hybrid and More: ATR spoken language translation system. In *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan.

K. Toutanova, H. T. Ilhan, and C. Manning. 2002. Extentions to HMM-based statistical word alignment models. In *Proc. of EMNLP*.

A. Venugopal, S. Vogel, and A. Waibel. 2003. Effective phrase translation extraction from alignment models. In *Proc. of ACL*.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM based word alignment in statistical translation. In *Proc. of the COLING*.

I. H. Witten and T. C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. In *IEEE Trans. Inform Theory*, volume 37, pages 1085–1094, July.