Local Phrase Reordering Models for Statistical Machine Translation

Shankar Kumar, William Byrne*
Center for Language and Speech Processing, Johns Hopkins University,
3400 North Charles Street, Baltimore, MD 21218, U.S.A.
Machine Intelligence Lab, Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, U.K.
skumar@jhu.edu , wjb31@cam.ac.uk

Abstract

We describe stochastic models of local phrase movement that can be incorporated into a Statistical Machine Translation (SMT) system. These models provide properly formulated, non-deficient, probability distributions over reordered phrase sequences. They are implemented by Weighted Finite State Transducers. We describe EM-style parameter re-estimation procedures based on phrase alignment under the complete translation model incorporating reordering. Our experiments show that the reordering model yields substantial improvements in translation performance on Arabic-to-English and Chinese-to-English MT tasks. We also show that the procedure scales as the bitext size is increased.

1 Introduction

Word and Phrase Reordering is a crucial component of Statistical Machine Translation (SMT) systems. However allowing reordering in translation is computationally expensive and in some cases even provably NP-complete (Knight, 1999). Therefore any translation scheme that incorporates reordering must necessarily balance model complexity against the ability to realize the model without approximation. In this paper our goal is to formulate models of local phrase reordering in such a way that they can be embedded inside a generative phrase-based model of translation (Kumar et al., 2005). Although this model of reordering is somewhat limited and cannot capture all possible phrase movement, it forms a proper parameterized probability distribution over reorderings of phrase sequences. We show that with this model it is possible to perform Maximum A Posteriori (MAP) decoding (with pruning) and Expectation Maximization (EM) style re-estimation of model parameters over large bitext collections.

We now discuss prior work on word and phrase reordering in translation. We focus on SMT systems that do not require phrases to form syntactic constituents.

The IBM translation models (Brown et al., 1993) describe word reordering via a distortion model defined over word positions within sentence pairs. The Alignment Template Model (Och et al., 1999) uses phrases rather than words as the basis for translation, and defines movement at the level of phrases. Phrase reordering is modeled as a first order Markov process with a single parameter that controls the degree of movement.

Our current work is inspired by the block (phrase-pair) orientation model introduced by Tillmann (2004) in which reordering allows neighboring blocks to swap. This is described as a sequence of orientations (left, right, neutral) relative to the monotone block order. Model parameters are block-specific and estimated over word aligned trained bitext using simple heuristics.

Other researchers (Vogel, 2003; Zens and Ney, 2003; Zens et al., 2004) have reported performance gains in translation by allowing deviations from monotone word and phrase order. In these cases,

* This work was supported by an ONR MURI Grant N00014-01-1-0685.

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 161–168, Vancouver, October 2005. ©2005 Association for Computational Linguistics

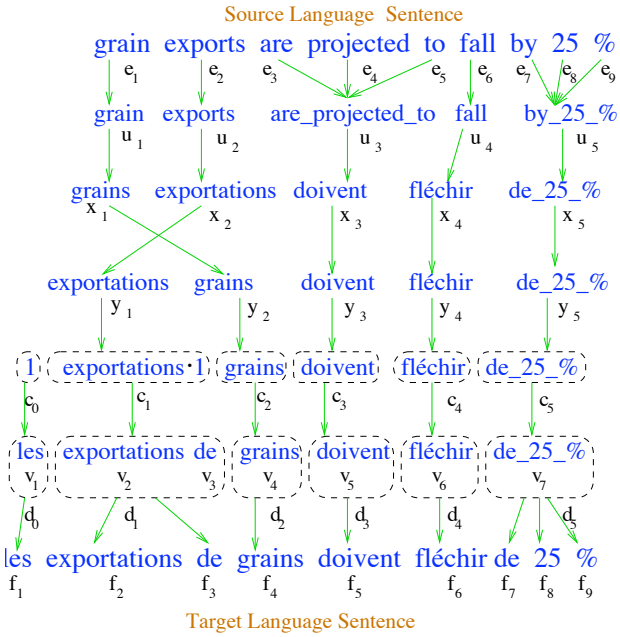

Figure 1: TTM generative translation process; here, $I = 9, K = 5, R = 7, J = 9$.

reordering is not governed by an explicit probabilistic model over reordered phrases; a language model is employed to select the translation hypothesis. We also note the prior work of Wu (1996), closely related to Tillmann's model.

2 The WFST Reordering Model

The Translation Template Model (TTM) is a generative model of phrase-based translation (Brown et al., 1993). Bitext is described via a stochastic process that generates source (English) sentences and transforms them into target (French) sentences (Fig 1 and Eqn 1).

$$\begin{aligned} &P(f_1^J, v_1^R, d_0^K, c_0^K, y_1^K, x_1^K, u_1^K, K, e_1^I) \;=\\ &P(e_1^I)\cdot\\ &\quad \textit{Source Language Model} \quad G\\ &P(u_1^K, K|e_1^I)\cdot\\ &\quad \textit{Source Phrase Segmentation} \quad W\\ &P(x_1^K|u_1^K, K, e_1^I)\cdot\\ &\quad \textit{Phrase Translation and Reordering} \quad R\\ &P(v_1^R, d_0^K, c_0^K, y_1^K|x_1^K, u_1^K, K, e_1^I)\cdot\\ &\quad \textit{Target Phrase Insertion} \quad \Phi\\ &P(f_1^J|v_1^R, d_0^K, c_0^K, y_1^K, x_1^K, u_1^K, K, e_1^I)\\ &\quad \textit{Target Phrase Segmentation} \quad \Omega \end{aligned} \tag{1}$$

The TTM relies on a Phrase-Pair Inventory (PPI) consisting of target language phrases and their source language translations. Translation is modeled via component distributions realized as WFSTs (Fig 1 and Eqn 1) : Source Language Model (G), Source Phrase Segmentation (W), Phrase Translation and Reordering (R), Target Phrase Insertion (Φ), and Target Phrase Segmentation (Ω) (Kumar et al., 2005).

TTM Reordering Previously, the TTM was formulated with reordering prior to translation; here, we perform reordering of phrase sequences following translation. Reordering prior to translation was found to be memory intensive and unwieldy (Kumar et al., 2005). In contrast, we will show that the current model can be used for both phrase alignment and translation.

2.1 The Phrase Reordering Model

We now describe two WFSTs that allow local reordering within phrase sequences. The simplest allows swapping of adjacent phrases. The second allows phrase movement within a three phrase window. Our formulation ensures that the overall model provides a proper parameterized probability distribution over reordered phrase sequences; we emphasize that the resulting distribution is not degenerate.

Phrase reordering (Fig 2) takes as its input a French phrase sequence in English phrase order $x_1, x_2, ..., x_K$. This is then reordered into French phrase order $y_1, y_2, ..., y_K$. Note that words within phrases are not affected.

We make the following conditional independence assumption:

$$P(y_1^K|x_1^K, u_1^K, K, e_1^I) = P(y_1^K|x_1^K, u_1^K). \tag{2}$$

Given an input phrase sequence x_1^K we now associate a unique *jump sequence* b_1^K with each permissible output phrase sequence y_1^K. The jump b_k measures the displacement of the k^{th} phrase x_k, i.e.

$$x_k \rightarrow y_{k+b_k}, k \in \{1, 2, ..., K\}. \tag{3}$$

The jump sequence b_1^K is constructed such that y_1^K is a permutation of x_1^K. This is enforced by constructing all models so that $\sum_{k=1}^K b_k = 0$.

We now redefine the model in terms of the jump sequence

$$\begin{aligned} &P(y_1^K|x_1^K, u_1^K) \\ &= \begin{cases} P(b_1^K|x_1^K, u_1^K) & y_{k+b_k} = x_k \;\forall k \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{4}$$

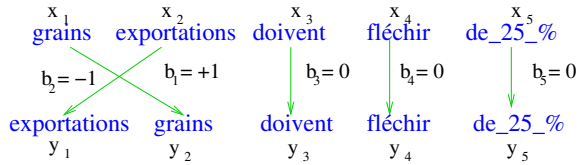

Figure 2: Phrase reordering and jump sequence.

-

where y_1^K is determined by x_1^K and b_1^K.

Each jump b_k depends on the phrase-pair (x_k, u_k) and preceding jumps b_1^{k-1}

$$P(b_1^K|x_1^K, u_1^K) = \prod_{k=1}^{K} P(b_k|x_k, u_k, \phi_{k-1}), \quad (5)$$

where ϕ_{k-1} is an equivalence classification (state) of the jump sequence b_1^{k-1}.

The jump sequence b_1^K can be described by a deterministic finite state machine. $\phi(b_1^{k-1})$ is the state arrived at by b_1^{k-1}; we will use ϕ_{k-1} to denote $\phi(b_1^{k-1})$.

We will investigate phrase reordering by restricting the maximum allowable jump to 1 phrase and to 2 phrases; we will refer to these reordering models as MJ-1 and MJ-2. In the first case, $b_k \in \{0, +1, -1\}$ while in the second case, $b_k \in \{0, +1, -1, +2, -2\}$.

2.2 Reordering WFST for MJ-1

We first present the Finite State Machine of the phrase reordering process (Fig 3) which has two equivalence classes (FSM states) for any given history b_1^{k-1}; $\phi(b_1^{k-1}) \in \{1, 2\}$. A jump of $+1$ has to be followed by a jump of -1, and 1 is the start and end state; this ensures $\sum_{k=1}^{K} b_k = 0$.

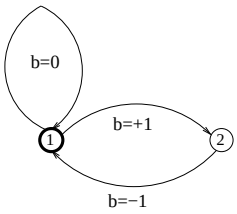

Figure 3: Phrase reordering process for MJ-1.

Under this restriction, the probability of the jump b_k (Eqn 5) can be simplified as

$$P(b_k|x_k, u_k, \phi(b_1^{k-1})) = \quad (6)$$
$$\begin{cases} \beta_1(x_k, u_k) & b_k = +1, \phi_{k-1} = 1 \\ 1 - \beta_1(x_k, u_k) & b_k = 0, \phi_{k-1} = 1 \\ 1 & b_k = -1, \phi_{k-1} = 2. \end{cases}$$

There is a single parameter jump probability $\beta_1(x, u) = P(b = +1|x, u)$ associated with each phrase-pair (x, u) in the phrase-pair inventory. This is the probability that the phrase-pair (x, u) appears out of order in the transformed phrase sequence.

We now describe the MJ-1 WFST. In the presentation, we use upper-case letters to denote the English phrases (u_k) and lower-case letters to denote the French phrases (x_k and y_k).

The PPI for this example is given in Table 1.

English	French	Parameters	
u	x	$P(x\|u)$	$\beta_1(x, u)$
A	a	0.5	0.2
A	d	0.5	0.2
B	b	1.0	0.4
C	c	1.0	0.3
D	d	1.0	0.8

Table 1: Example phrase-pair inventory with translation and reordering probabilities.

The input to the WFST (Fig 4) is a lattice of French phrase sequences derived from the French sentence to be translated. The outputs are the corresponding English phrase sequences. Note that the reordering is performed on the English side.

The WFST is constructed by adding a self-loop for each French phrase in the input lattice, and a 2-arc path for every pair of adjacent French phrases in the lattice. The WFST incorporates the translation model $P(x|u)$ and the reordering model $P(b|x, u)$. The score on a self-loop with labels (u, x) is $P(x|u) \times (1 - \beta_1(x, u))$; on a 2-arc path with labels (u_1, x_1) and (u_2, x_2), the score on the 1st arc is $P(x_2|u_1) \times \beta_1(x_2, u_1)$ and on the 2nd arc is $P(x_1|u_2)$.

In this example, the input to this transducer is a single French phrase sequence $V : a, b, c$. We perform the WFST composition $R \circ V$, project the result on the input labels, and remove the epsilons to form the acceptor $(R \circ V)_1$ which contains the six English phrase sequences (Fig 4).

Translation Given a French sentence, a lattice of translations is obtained using the weighted finite state composition: $\mathcal{T} = G \circ W \circ R \circ \Phi \circ \Omega \circ T$. The most-likely translation is obtained as the path with the highest probability in $\mathcal{T}$.

Alignment Given a sentence-pair (E, F), a lattice of phrase alignments is obtained by the finite state composition: $\mathcal{B} = S \circ W \circ R \circ \Phi \circ \Omega \circ T$, where

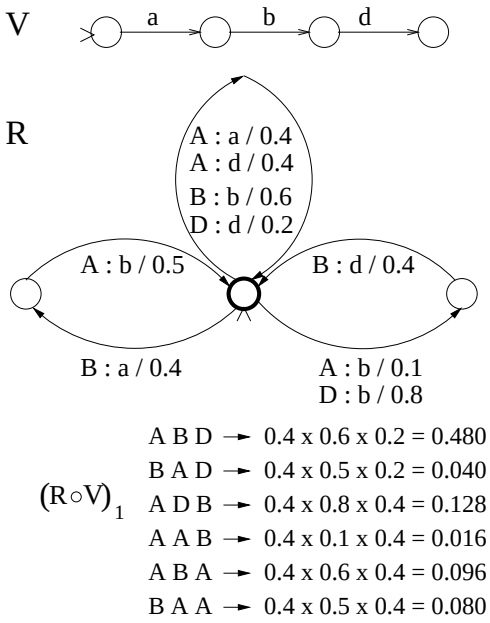

Figure 4: WFST for the MJ-1 model.

S is an acceptor for the English sentence E, and T is an acceptor for the French sentence F. The Viterbi alignment is found as the path with the highest probability in $\mathcal{B}$. The WFST composition gives the word-to-word alignments between the sentences. However, to obtain the phrase alignments, we need to construct additional FSTs not described here.

2.3 Reordering WFST for MJ-2

MJ-2 reordering restricts the maximum allowable jump to 2 phrases and also insists that the reordering take place within a window of 3 phrases. This latter condition implies that for an input sequence $\{a, b, c, d\}$, we disallow the three output sequences: $\{b, d, a, c;\ c, a, d, b;\ c, d, a, b;\}$. In the MJ-2 finite state machine, a given history b_1^{k-1} can lead to one of the six states in Fig 5.

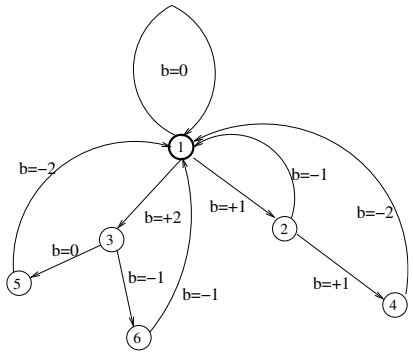

Figure 5: Phrase reordering process for MJ-2.

The jump probability of Eqn 5 becomes

$$P(b_k|x_k, u_k, \phi_{k-1}) =$$

$$\begin{cases} \beta_1(x_k, u_k) & b_k = 1, \phi_{k-1} = 1 \\ \beta_2(x_k, u_k) & b_k = 2, \phi_{k-1} = 1 \\ \begin{cases} 1 - \beta_1(x_k, u_k) \\ -\beta_2(x_k, u_k) \end{cases} & b_k = 0, \phi_{k-1} = 1 \end{cases} \quad (7)$$

$$\begin{cases} \beta_1(x_k, u_k) & b_k = 1, \phi_{k-1} = 2 \\ 1 - \beta_1(x_k, u_k) & b_k = -1, \phi_{k-1} = 2 \end{cases} \quad (8)$$

$$\begin{cases} 0.5 & b_k = 0, \phi_{k-1} = 3 \\ 0.5 & b_k = -1, \phi_{k-1} = 3. \end{cases} \quad (9)$$

$$\begin{cases} 1 & b_k = -2, \phi_{k-1} = 4 \end{cases} \quad (10)$$

$$\begin{cases} 1 & b_k = -2, \phi_{k-1} = 5 \end{cases} \quad (11)$$

$$\begin{cases} 1 & b_k = -1, \phi_{k-1} = 6 \end{cases} \quad (12)$$

We note that the distributions (Eqns 7 and 8) are based on two parameters $\beta_1(x, u)$ and $\beta_2(x, u)$ for each phrase-pair (x, u).

Suppose the input is a phrase sequence a, b, c, the MJ-2 model (Fig 5) allows 6 possible reorderings: $a, b, c; a, c, b; b, a, c; b, c, a; c, a, b; c, b, a$. The distribution Eqn 9 ensures that the sequences b, c, a and c, b, a are assigned equal probability. The distributions in Eqns 10-12 ensure that the maximum jump is 2 phrases and the reordering happens within a window of 3 phrases. By insisting that the process start and end at state 1 (Fig 5), we ensure that the model is not deficient. A WFST implementing the MJ-2 model can be easily constructed for both phrase alignment and translation, following the construction described for the MJ-1 model.

3 Estimation of the Reordering Models

The Translation Template Model relies on an inventory of target language phrases and their source language translations. Our goal is to estimate the reordering model parameters $P(b|x, u)$ for each phrase-pair (x, u) in this inventory. However, when translating a given test set, only a subset of the phrase-pairs is needed. Although there may be an advantage in estimating the model parameters under an inventory that covers all the training bitext, we fix the phrase-pair inventory to cover only the phrases on the test set. Estimation of the reordering model parameters over the training bitext is then performed under this test-set specific inventory.

We employ the EM algorithm to obtain Maximum Likelihood (ML) estimates of the reordering model parameters. Applying EM to the MJ-1 reordering model gives the following ML parameter estimates for each phrase-pair (u, x).

$$\hat{\beta}_1(x,u) = \frac{C_{x,u}(0,+1)}{C_{x,u}(0,+1) + C_{x,u}(0,0)}. \quad (13)$$

$C_{x,u}(\phi, b)$ is defined for $\phi = 1, 2$ and $b = -1, 0, +1$. Any permissible phrase alignment of a sentence pair corresponds to a b_1^K sequence, which in turn specifies a ϕ_1^K sequence. $C_{x,u}(\phi, b)$ is the expected number of times the phrase-pair x, u is aligned with a jump of b phrases when the jump history is ϕ. We do not use full EM but a Viterbi training procedure that obtains the counts for the best (Viterbi) alignments. If a phrase-pair (x, u) is never seen in the Viterbi alignments, we back-off to a flat parameter $\beta_1(x, u) = 0.05$.

The ML parameter estimates for the MJ-2 model are given in Table 2, with $C_{x,u}(\phi, b)$ defined similarly. In our training scenario, we use WFST operations to obtain Viterbi phrase alignments of the training bitext where the initial reordering model parameters ($\beta_0(x, u)$) are set to a uniform value of 0.05. The counts $C_{x,u}(s, b)$ are then obtained over the phrase alignments. Finally the ML estimates of the parameters are computed using Eqn 13 (MJ-1) or Eqn 14 (MJ-2). We will refer to the Viterbi trained models as MJ-1 VT and MJ-2 VT. Table 3 shows the MJ-1 VT parameters for some example phrase-pairs in the Arabic-English (A-E) task.

u	x	$\beta_1(x,u)$
which_is_the_closest	Aqrb	1.0
international_trade	tjArp_EAlmyp	0.8
the_foreign_ministry	wzArp_xArjyp	0.6
arab_league	jAmEp_dwl_Erbyp	0.4

Table 3: MJ-1 parameters for A-E phrase-pairs.

To validate alignment under a PPI, we measure performance of the TTM word alignments on French-English (500 sent-pairs) and Chinese-English (124 sent-pairs) (Table 4). As desired, the Alignment Recall (AR) and Alignment Error Rate (AER) improve modestly while Alignment Precision (AP) remains constant. This suggests that the models allow more words to be aligned and thus improve the recall; MJ-2 gives a further improvement in AR and AER relative to MJ-1. Alignment precision depends on the quality of the word alignments within the phrase-pairs and does not change much by allowing phrase reordering. This experiment validates the estimation procedure based on the phrase alignments; however, we do not advocate the use of TTM as an alternate word alignment technique.

Reordering	Metrics (%)					
	Frn-Eng			Chn-Eng		
	AP	AR	AER	AP	AR	AER
None	94.2	84.8	10.0	85.1	47.1	39.3
MJ-1 VT	94.1	86.8	9.1	85.3	49.4	37.5
MJ-2 VT	93.9	87.4	8.9	85.3	50.9	36.3

Table 4: Alignment Performance with Reordering.

4 Translation Experiments

We perform our translation experiments on the large data track of the NIST Arabic-to-English (A-E) and Chinese-to-English (C-E) MT tasks; we report results on the NIST 2002, 2003, and 2004 evaluation test sets [1].

4.1 Exploratory Experiments

In these experiments the training data is restricted to FBIS bitext in C-E and the news bitexts in A-E. The bitext consists of chunk pairs aligned at sentence and sub-sentence level (Deng et al., 2004). In A-E, the training bitext consists of 3.8M English words, 3.2M Arabic words and 137K chunk pairs. In C-E, the training bitext consists of 11.7M English words, 8.9M Chinese words and 674K chunk pairs.

Our Chinese text processing consists of word segmentation (using the LDC segmenter) followed by grouping of numbers. For Arabic our text processing consisted of a modified Buckwalter analysis (LDC2002L49) followed by post processing to separate conjunctions, prepostions and pronouns, and Al-/w- deletion. The English text is processed using a simple tokenizer based on the text processing utility available in the the NIST MT-eval toolkit.

The Language Model (LM) training data consists of approximately 400M words of English text derived from Xinhua and AFP (English Gigaword), the English side of FBIS, the UN and A-E News texts, and the online archives of The People's Daily.

Table 5 gives the performance of the MJ-1 and MJ-2 reordering models when translation is performed using a 4-gram LM. We report performance on the 02, 03, 04 test sets and the combined test set

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

$$\hat{\beta}_1(x,u) = \frac{C_{x,u}(1,+1) + C_{x,u}(2,+1)}{C_{x,u}(1,+1) + C_{x,u}(1,0) + C_{x,u}(1,+2) + C_{x,u}(2,+1) + C_{x,u}(2,-1)}$$

$$\hat{\beta}_2(x,u) = \frac{(C_{x,u}(1,0) + C_{x,u}(2,-1) + C_{x,u}(1,+2))C_{x,u}(1,+2)}{(C_{x,u}(1,+1) + C_{x,u}(1,0) + C_{x,u}(1,+2) + C_{x,u}(2,+1) + C_{x,u}(2,-1))(C_{x,u}(1,+2) + C_{x,u}(1,0))}$$

Table 2: ML parameter estimates for MJ-2 model.

Reordering	BLEU (%)							
	Arabic-English				Chinese-English			
	02	03	04	ALL	02	03	04	ALL
None	37.5	40.3	36.8	37.8 ± 0.6	24.2	23.7	26.0	25.0 ± 0.5
MJ-1 flat	40.4	43.9	39.4	40.7 ± 0.6	25.7	24.5	27.4	26.2 ± 0.5
MJ-1 VT	41.3	44.8	40.3	41.6 ± 0.6	25.8	24.5	27.8	26.5 ± 0.5
MJ-2 flat	41.0	44.4	39.7	41.1 ± 0.6	26.4	24.9	27.7	26.7 ± 0.5
MJ-2 VT	41.7	45.3	40.6	42.0 ± 0.6	26.5	24.9	27.9	26.8 ± 0.5

Table 5: Performance of MJ-1 and MJ-2 reordering models with a 4-gram LM.

(ALL=02+03+04). For the combined set (ALL), we also show the 95% BLEU confidence interval computed using bootstrap resampling (Och, 2003).

Row 1 gives the performance when no reordering model is used. The next two rows show the influence of the MJ-1 reordering model; in row 2, a flat probability of $\beta_1(x,u) = 0.05$ is used for all phrase-pairs; in row 3, a reordering probability is estimated for each phrase-pair using Viterbi Training (Eqn 13). The last two rows show the effect of the MJ-2 reordering model; row 4 uses flat probabilities ($\beta_1(x,u) = 0.05, \beta_2(x,u) = 0.01$) for all phrase-pairs; row 5 applies reordering probabilities estimating with Viterbi Training for each phrase-pair (Table 2).

On both language-pairs, we observe that reordering yields significant improvements. The gains from phrase reordering are much higher on A-E relative to C-E; this could be related to the fact that the word order differences between English and Arabic are much higher than the differences between English and Chinese. MJ-1 VT outperforms flat MJ-1 showing that there is value in estimating the reordering parameters from bitext. Finally, the MJ-2 VT model performs better than the flat MJ-2 model, but only marginally better than the MJ-1 VT model. Therefore estimation does improve the MJ-2 model but allowing reordering beyond a window of 1 phrase is not useful when translating either Arabic or Chinese into English in this framework.

The flat MJ-1 model outperforms the no-reordering case and the flat MJ-2 model is better than the flat MJ-1 model; we hypothesize that phrase reordering increases search space of translations that allows the language model to select a higher quality hypothesis. This suggests that these models of phrase reordering actually require strong language models to be effective. We now investigate the interaction between language models and reordering.

Our goal here is to measure translation performance of reordering models over variable span n-gram LMs (Table 6). We observe that both MJ-1 and MJ-2 models yield higher improvements under higher order LMs: e.g. on A-E, gains under 3g (3.6 BLEU points on MJ-1, 0.2 points on MJ-2) are higher than the gains with 2g (2.4 BLEU points on MJ-1, 0.1 points on MJ-2).

Reordering	BLEU (%)					
	A-E			C-E		
	2g	3g	4g	2g	3g	4g
None	21.0	36.8	37.8	16.1	24.8	25.0
MJ-1 VT	23.4	40.4	41.6	16.2	25.9	26.5
MJ-2 VT	23.5	40.6	42.0	16.0	26.1	26.8

Table 6: Reordering with variable span n-gram LMs on Eval02+03+04 set.

We now measure performance of the reordering models across the three test set genres used in the NIST 2004 evaluation: news, editorials, and speeches. On A-E, MJ-1 and MJ-2 yield larger improvements on News relative to the other genres; on C-E, the gains are larger on Speeches and Editorials relative to News. We hypothesize that the Phrase-Pair Inventory, reordering models and language models could all have been biased away from the test set due to the training data. There may also be less movement across these other genres.

Reordering	BLEU (%)					
	A-E			C-E		
	News	Eds	Sphs	News	Eds	Sphs
None	41.1	30.8	33.3	23.6	25.9	30.8
MJ-1 VT	45.6	32.6	35.7	24.8	27.8	33.3
MJ-2 VT	46.2	32.7	35.5	24.8	27.8	33.7

Table 7: Performance across Eval 04 test genres.

	BLEU (%)					
	Arabic-English			Chinese-English		
Reordering	02	03	04n	02	03	04n
None	40.2	42.3	43.3	28.9	27.4	27.3
MJ-1 VT	43.1	45.0	45.6	30.2	28.2	28.9
MET-Basic	44.8	47.2	48.2	31.3	30.3	30.3
MET-IBM1	45.2	48.2	49.7	31.8	30.7	31.0

Table 8: Translation Performance on Large Bitexts.

4.2 Scaling to Large Bitext Training Sets

We here describe the integration of the phrase reordering model in an MT system trained on large bitexts. The text processing and language models have been described in § 4.1. Alignment Models are trained on all available bitext (7.6M chunk pairs/207.4M English words/175.7M Chinese words on C-E and 5.1M chunk pairs/132.6M English words/123.0M Arabic words on A-E), and word alignments are obtained over the bitext. Phrase-pairs are then extracted from the word alignments (Koehn et al., 2003). MJ-1 model parameters are estimated over all bitext on A-E and over the non-UN bitext on C-E. Finally we use Minimum Error Training (MET) (Och, 2003) to train log-linear scaling factors that are applied to the WFSTs in Equation 1. 04news (04n) is used as the MET training set.

Table 8 reports the performance of the system. Row 1 gives the performance without phrase reordering and Row 2 shows the effect of the MJ-1 VT model. The MJ-1 VT model is used in an initial decoding pass with the four-gram LM to generate translation lattices. These lattices are then rescored under parameters obtained using MET (MET-basic), and 1000-best lists are generated. The 1000-best lists are augmented with IBM Model-1 (Brown et al., 1993) scores and then rescored with a second set of MET parameters. Rows 3 and 4 show the performance of the MET-basic and MET-IBM1 models.

We observe that the maximum likelihood phrase reordering model (MJ-1 VT) yields significantly improved translation performance relative to the monotone phrase order translation baseline. This confirms the translation performance improvements found over smaller training bitexts.

We also find additional gains by applying MET to optimize the scaling parameters that are applied to the WFST component distributions within the TTM (Equation 1). In this procedure, the scale factor applied to the MJ-1 VT Phrase Translation and Reordering component is estimated along with scale factors applied to the other model components; in other words, the ML-estimated phrase reordering model itself is not affected by MET, but the likelihood that it assigns to a phrase sequence is scaled by a single, discriminatively optimized weight. The improvements from MET (see rows MET-Basic and MET- IBM1) demonstrate that the MJ-1 VT reordering models can be incorporated within a discriminative optimized translation system incorporating a variety of models and estimation procedures.

5 Discussion

In this paper we have described local phrase reordering models developed for use in statistical machine translation. The models are carefully formulated so that they can be implemented as WFSTs, and we show how the models can be incorporated into the Translation Template Model to perform phrase alignment and translation using standard WFST operations. Previous approaches to WFST-based reordering (Knight and Al-Onaizan, 1998; Kumar and Byrne, 2003; Tsukada and Nagata, 2004) constructed permutation acceptors whose state spaces grow exponentially with the length of the sentence to be translated. As a result, these acceptors have to be pruned heavily for use in translation. In contrast, our models of local phrase movement do not grow explosively and do not require any pruning or approximation in their construction. In other related work, Bangalore and Ricardi (2001) have trained WFSTs for modeling reordering within translation; their WFST parses word sequences into trees containing reordering information, which are then checked for well-formed brackets. Unlike this approach, our model formulation does not use a tree representation and also ensures that the output sequences are valid permutations of input phrase sequences; we emphasize again that the probability distribution induced over reordered phrase sequences is not degenerate.

Our reordering models do resemble those of (Tillmann, 2004; Tillmann and Zhang, 2005) in that we

treat the reordering as a sequence of jumps relative to the original phrase sequence, and that the likelihood of the reordering is assigned through phrase-pair specific parameterized models. We note that our implementation allows phrase reordering beyond simply a 1-phrase window, as was done by Tillmann. More importantly, our model implements a generative model of phrase reordering which can be incorporated directly into a generative model of the overall translation process. This allows us to perform 'embedded' EM-style parameter estimation, in which the parameters of the phrase reordering model are estimated using statistics gathered under the complete model that will actually be used in translation. We believe that this estimation of model parameters directly from phrase alignments obtained under the phrase translation model is a novel contribution; prior approaches derived the parameters of the reordering models from word aligned bitext, e.g. within the phrase pair extraction procedure.

We have shown that these models yield improvements in alignment and translation performance on Arabic-English and Chinese-English tasks, and that the reordering model can be integrated into large evaluation systems. Our experiments show that discriminative training procedures such Minimum Error Training also yield additive improvements by tuning TTM systems which incorporate ML-trained reordering models. This is essential for integrating our reordering model inside an evaluation system, where a variety of techniques are applied simultaneously.

The MJ-1 and MJ-2 models are extremely simple models of phrase reordering. Despite their simplicity, these models provide large improvements in BLEU score when incorporated into a monotone phrase order translation system. Moreover, they can be used to produced translation lattices for use by more sophisticated reordering models that allow longer phrase order movement. Future work will build on these simple structures to produce more powerful models of word and phrase movement in translation.

References

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Y. Deng, S. Kumar, and W. Byrne. 2004. Bitext chunk alignment for statistical machine translation. In *Research Note, Center for Language and Speech Processing, Johns Hopkins University*.

K. Knight and Y. Al-Onaizan. 1998. Translation with finite-state devices. In *AMTA*, pages 421–437, Langhorne, PA, USA.

K. Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics, Squibs & Discussion*, 25(4).

P. Koehn, F. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL*, pages 127–133, Edmonton, Canada.

S. Kumar and W. Byrne. 2003. A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In *HLT-NAACL*, pages 142–149, Edmonton, Canada.

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(4).

F. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *EMNLP-VLC*, pages 20–28, College Park, MD, USA.

F. Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*, Sapporo, Japan.

C. Tillmann and T. Zhang. 2005. A localized prediction model for statistical machine translation. In *ACL*, Ann Arbor, Michigan, USA.

C. Tillmann. 2004. A block orientation model for statistical machine translation. In *HLT-NAACL*, Boston, MA, USA.

H. Tsukada and M. Nagata. 2004. Efficient decoding for statistical machine translation with a fully expanded WFST model. In *EMNLP*, Barcelona, Spain.

S. Vogel. 2003. SMT Decoder Dissected: Word Reordering. In *NLPKE*, Beijing, China.

D. Wu. 1996. A polynomial-time algorithm for statistical machine translation. In *ACL*, pages 152–158, Santa Cruz, CA, USA.

R. Zens and H. Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *ACL*, pages 144–151, Sapporo, Japan.

R. Zens, H. Ney, T. Watanabe, and E. Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *COLING*, pages 205–211, Boston, MA, USA.