A Simple and Effective Hierarchical Phrase Reordering Model

Michel Galley
Computer Science Department
Stanford University
Stanford, CA 94305-9020
`galley@cs.stanford.edu`

Christopher D. Manning
Computer Science Department
Stanford University
Stanford, CA 94305-9010
`manning@cs.stanford.edu`

Abstract

While phrase-based statistical machine translation systems currently deliver state-of-the-art performance, they remain weak on word order changes. Current phrase reordering models can properly handle swaps between adjacent phrases, but they typically lack the ability to perform the kind of long-distance reorderings possible with syntax-based systems. In this paper, we present a novel hierarchical phrase reordering model aimed at improving non-local reorderings, which seamlessly integrates with a standard phrase-based system with little loss of computational efficiency. We show that this model can successfully handle the key examples often used to motivate syntax-based systems, such as the rotation of a prepositional phrase around a noun phrase. We contrast our model with reordering models commonly used in phrase-based systems, and show that our approach provides statistically significant BLEU point gains for two language pairs: Chinese-English (+0.53 on MT05 and +0.71 on MT08) and Arabic-English (+0.55 on MT05).

1 Introduction

Statistical phrase-based systems (Och and Ney, 2004; Koehn et al., 2003) have consistently delivered state-of-the-art performance in recent machine translation evaluations, yet these systems remain weak at handling word order changes. The reordering models used in the original phrase-based systems penalize phrase displacements proportionally to the amount of nonmonotonicity, with no consideration of the fact that some words are far more
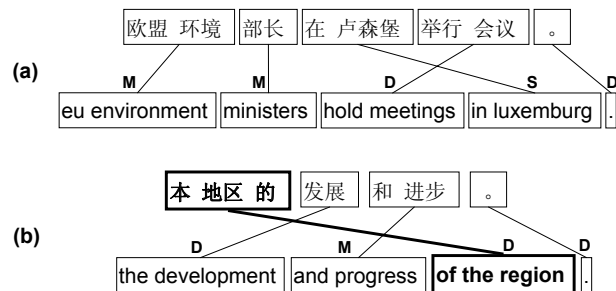

Figure 1: Phase orientations (monotone, swap, discontinuous) for Chinese-to-English translation. While previous work reasonably models phrase reordering in simple examples (a), it fails to capture more complex reorderings, such as the swapping of "of the region" (b).

likely to be displaced than others (e.g., in English-to-Japanese translation, a verb should typically move to the end of the clause).

Recent efforts (Tillman, 2004; Och et al., 2004; Koehn et al., 2007) have directly addressed this issue by introducing lexicalized reordering models into phrase-based systems, which condition reordering probabilities on the words of each phrase pair. These models distinguish three orientations with respect to the previous phrase—monotone (*M*), swap (*S*), and discontinuous (*D*)—and as such are primarily designed to handle *local* re-orderings of neighboring phrases. Fig. 1(a) is an example where such a model effectively swaps the prepositional phrase *in Luxembourg* with a verb phrase, and where the noun *ministers* remains in monotone order with respect to the previous phrase *EU environment*.

While these lexicalized re-ordering models have shown substantial improvements over unlexicalized phrase-based systems, these models only have a

limited ability to capture sensible long distance re-orderings, as can be seen in Fig. 1(b). The phrase *of the region* should swap with the rest of the noun phrase, yet these previous approaches are unable to model this movement, and assume the orientation of this phrase is discontinuous (*D*). Observe that, in a shortened version of the same sentence (without *and progress*), the phrase orientation would be different (*S*), even though the shortened version has essentially the same sentence structure. Coming from the other direction, such observations about phrase reordering between different languages are precisely the kinds of facts that parsing approaches to machine translation are designed to handle and do successfully handle (Wu, 1997; Melamed, 2003; Chiang, 2005).

In this paper, we introduce a novel orientation model for phrase-based systems that aims to better capture long distance dependencies, and that presents a solution to the problem illustrated in Fig. 1(b). In this example, our reordering model effectively treats the adjacent phrases *the development* and *and progress* as one single phrase, and the displacement of *of the region* with respect to this phrase can be treated as a swap. To be able identify that adjacent blocks (e.g., *the development* and *and progress*) can be merged into larger blocks, our model infers binary (non-linguistic) trees reminiscent of (Wu, 1997; Chiang, 2005). Crucially, our work distinguishes itself from previous hierarchical models in that it does not rely on any cubic-time parsing algorithms such as CKY (used in, e.g., (Chiang, 2005)) or the Earley algorithm (used in (Watanabe et al., 2006)). Since our reordering model does not attempt to resolve natural language ambiguities, we can effectively rely on (linear-time) shift-reduce parsing, which is done jointly with left-to-right phrase-based beam decoding and thus introduces no asymptotic change in running time. As such, the hierarchical model presented in this paper maintains all the effectiveness and speed advantages of statistical phrase-based systems, while being able to capture some key linguistic phenomena (presented later in this paper) which have motivated the development of parsing-based approaches. We also illustrate this with results that are significantly better than previous approaches, in particular the lexical reordering models of Moses, a widely used phrase-based SMT system (Koehn et al., 2007).

This paper is organized as follows: the training of lexicalized re-ordering models is described in Section 3. In Section 4, we describe how to combine shift-reduce parsing with left-to-right beam search phrase-based decoding with the same asymptotic running time as the original phrase-based decoder. We finally show in Section 6 that our approach yields results that are significantly better than previous approaches for two language pairs and different test sets.

2 Lexicalized Reordering Models

We compare our re-ordering model with related work (Tillman, 2004; Koehn et al., 2007) using a log-linear approach common to many state-of-the-art statistical machine translation systems (Och and Ney, 2004). Given an input sentence $\mathbf{f}$, which is to be translated into a target sentence $\mathbf{e}$, the decoder searches for the most probable translation $\hat{\mathbf{e}}$ according to the following decision rule:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \left\{ p(\mathbf{e}|\mathbf{f}) \right\} \tag{1}$$

$$= \arg\max_{\mathbf{e}} \left\{ \sum_{j=1}^{J} \lambda_j h_j(\mathbf{f},\mathbf{e}) \right\} \tag{2}$$

$h_j(\mathbf{f},\mathbf{e})$ are J arbitrary feature functions over sentence pairs. These features include lexicalized re-ordering models, which are parameterized as follows: given an input sentence $\mathbf{f}$, a sequence of target-language phrases $\mathbf{e} = (\bar{e}_1,\ldots,\bar{e}_n)$ currently hypothesized by the decoder, and a phrase alignment $\mathbf{a} = (a_1,\ldots,a_n)$ that defines a source $\bar{f}_{a_i}$ for each translated phrase $\bar{e}_i$, these models estimate the probability of a sequence of orientations $\mathbf{o} = (o_1,\ldots,o_n)$

$$p(\mathbf{o}|\mathbf{e},\mathbf{f}) = \prod_{i=1}^{n} p(o_i|\bar{e}_i,\bar{f}_{a_i},a_{i-1},a_i), \tag{3}$$

where each o_i takes values over the set of possible orientations $\mathscr{O} = \{M,S,D\}$.[1] The probability is conditioned on both a_{i-1} and a_i to make sure that the label o_i is consistent with the phrase alignment. Specifically, probabilities in these models can be

---

[1] We note here that the parameterization and terminology in (Tillman, 2004) is slightly different. We purposely ignore these differences in order to enable a direct comparison between Tillman's, Moses', and our approach.
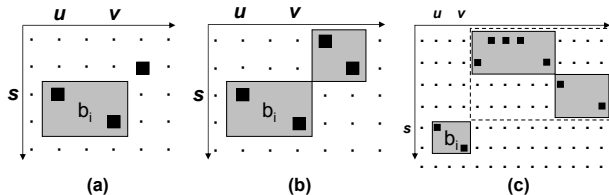
Figure 2: Occurrence of a swap according to the three orientation models: word-based, phrase-based, and hierarchical. Black squares represent word alignments, and gray squares represent blocks identified by phrase-extract. In (a), block $b_i = (e_i, f_{a_i})$ is recognized as a swap according to all three models. In (b), b_i is not recognized as a swap by the word-based model. In (c), b_i is recognized as a swap only by the hierarchical model.

greater than zero only if one of the following conditions is true:

- $o_i = M$ and $a_i - a_{i-1} = 1$

- $o_i = S$ and $a_i - a_{i-1} = -1$

- $o_i = D$ and $|a_i - a_{i-1}| \neq 1$

At decoding time, rather than using the log-probability of Eq. 3 as single feature function, we follow the approach of Moses, which is to assign three distinct parameters $(\lambda_m, \lambda_s, \lambda_d)$ for the three feature functions:

- $f_m = \sum_{i=1}^{n} \log p(o_i = M | \ldots)$

- $f_s = \sum_{i=1}^{n} \log p(o_i = S | \ldots)$

- $f_d = \sum_{i=1}^{n} \log p(o_i = D | \ldots)$.

There are two key differences between this work and previous orientation models (Tillman, 2004; Koehn et al., 2007): (1) the estimation of factors in Eq. 3 from data; (2) the segmentation of **e** and **f** into phrases, which is static in the case of (Tillman, 2004; Koehn et al., 2007), while it is dynamically updated with hierarchical phrases in our case. These differences are described in the two next sections.

3 Training

We present here three approaches for computing $p(o_i | \bar{e}_i, \bar{f}_{a_i}, a_{i-1}, a_i)$ on word-aligned data using relative frequency estimates. We assume here that phrase $\bar{e}_i$ spans the word range $s, \ldots, t$ in the target sentence **e** and that the phrase $\bar{f}_{a_i}$ spans the range

Orientation Model	$o_i = M$	$o_i = S$	$o_i = D$
word-based (Moses)	0.1750	0.0159	0.8092
phrase-based	0.3192	0.0704	0.6104
hierarchical	0.4878	0.1004	0.4116

Table 1: Class distributions of the three orientation models, estimated from 12M words of Chinese-English data using the grow-diag alignment symmetrization heuristic implemented in Moses, which is similar to the 'refined' heuristic of (Och and Ney, 2004).

$u, \ldots, v$ in the source sentence **f**. All phrase pairs in this paper are extracted with the phrase-extract algorithm (Och and Ney, 2004), with maximum length set to 7.

Word-based orientation model: This model analyzes word alignments at positions $(s-1, u-1)$ and $(s-1, v+1)$ in the alignment grid shown in Fig. 2(a). Specifically, orientation is set to $o_i = M$ if $(s-1, u-1)$ contains a word alignment and $(s-1, v+1)$ contains no word alignment. It is set to $o_i = S$ if $(s-1, u-1)$ contains no word alignment and $(s-1, v+1)$ contains a word alignment. In all other cases, it is set to $o_i = D$. This procedure is exactly the same as the one implemented in Moses.[2]

Phrase-based orientation model: The model presented in (Tillman, 2004) is similar to the word-based orientation model presented above, except that it analyzes adjacent phrases rather than specific word alignments to determine orientations. Specifically, orientation is set to $o_i = M$ if an adjacent phrase pair lies at $(s-1, u-1)$ in the alignment grid. It is set to S if an adjacent phrase pair covers $(s-1, v+1)$ (as shown in Fig. 2(b)), and is set to D otherwise.

Hierarchical orientation model: This model analyzes alignments beyond adjacent phrases. Specifically, orientation is set to $o_i = M$ if the phrase-extract algorithm is able to extract a phrase pair at $(s-1, u-1)$ given no constraint on maximum phrase length. Orientation is S if the same is true at $(s-1, v+1)$, and orientation is D otherwise.

Table 1 displays overall class distributions according to the three models. It appears clearly that occurrences of M and S are too sparsely seen in the word-based model, which assigns more than 80% of its

---

[2]http://www.statmt.org/moses/?n=Moses.AdvancedFeatures

			word	phrase	hier.
Monotone with previous			$p(o_i = M \mid \overline{e}_i, \overline{f}_{a_i}, a_{i-1}, a_i)$		
1	，是	and is	0.223	0.672	0.942
2	，也	and also	0.201	0.560	0.948
Swap with previous			$p(o_i = S \mid \overline{e}_i, \overline{f}_{a_i}, a_{i-1}, a_i)$		
3	中国 的	of china	0.303	0.617	0.651
4	他 说	, he said	0.003	0.030	0.395
Monotone with next			$p(o_i = M \mid \overline{e}_i, \overline{f}_{a_i}, a_{i+1}, a_i)$		
5	他 指出 ,	he pointed out that	0.601	0.770	0.991
6	然而 ,	however ,	0.517	0.728	0.968
Swap with next			$p(o_i = S \mid \overline{e}_i, \overline{f}_{a_i}, a_{i+1}, a_i)$		
7	的 发展	the development of	0.145	0.831	0.900
8	的 邀请	at the invitation of	0.272	0.834	0.925

Table 2: Monotone and swap probabilities for specific phrases according to the three models (word, phrase, and hierarchical). To ensure probabilities are representative, we only selected phrase pairs that occur at least 100 times in the training data.

probability mass to *D*. Conversely, the hierarchical model counts considerably less discontinuous cases, and is the only model that accounts for the fact that real data is predominantly monotone.

Since *D* is a rather uninformative default category that gives no clue how a particular phrase should be displaced, we will also provide MT evaluation scores (in Section 6) for a set of classes that distinguishes between left and right discontinuity $\{M, S, D_l, D_r\}$, a choice that is admittedly more linguistically motivated.

Table 2 displays orientation probabilities for concrete examples. Each example was put under one of the four categories that linguistically seems the best match, and we provide probabilities for that category according to each model. Note that, while we have so far only discussed left-to-right reordering models, it is also possible to build right-to-left models by substituting a_{i-1} with a_{i+1} in Eq. 3. Examples for right-to-left models appear in the second half of the table. The table strongly suggests that the hierarchical model more accurately determines the orientation of phrases with respect to large contextual blocks. In Examples 1 and 2, the hierarchical model captures the fact that coordinated clauses almost always remain in the same order, and that words should generally be forbidden to move from one side of "and" to the other side, a constraint that is difficult to enforce with the other two reordering models. In Example 4, the first two models completely ignore that "he said" sometimes rotates

around its neighbor clause.

4 Decoding

Computing reordering scores during decoding with word-based[3] and phrase-based models (Tillman, 2004) is trivial, since they only make use of local information to determine the orientation of a new incoming block b_i. For a left-to-right ordering model, b_i is scored based on its orientation with respect to b_{i-1}. For instance, if b_i has a swap orientation with respect to the previous phrase in the current translation hypothesis, feature $p(o_i = S \mid \ldots)$ becomes active.

Computing lexicalized reordering scores with the hierarchical model is more complex, since the model must identify contiguous blocks—monotone or swapping—that can be merged into hierarchical blocks. The employed method is an instance of the well-known shift-reduce parsing algorithm, and relies on a stack (*S*) of foreign substrings that have already been translated. Each time the decoder adds a new block to the current translation hypothesis, it shifts the source-language indices of the block onto *S*, then repeatedly tries reducing the top two elements of *S* if they are contiguous.[4] This parsing algorithm was first applied in computational geometry to identify convex hulls (Graham, 1972), and its running time was shown to be linear in the length of the sequence (a proof is presented in (Huang et al., 2008), which applies the same algorithm to the binarization of SCFG rules).

Figure 3 provides an example of the execution of this algorithm for the translation output shown in Figure 4, which was produced by a decoder incorporating our hierarchical reordering model. The decoder successively pushes source-language spans [1], [2], [3], which are successively merged into [1-3], and all correspond to monotone orientations.

---

[3]We would like to point out an inconsistency in Moses between training and testing. Despite the fact that Moses estimates a word-based orientation model during training (i.e., it analyzes the orientation of a given phrase with respect to adjacent word alignments), this model is then treated as a phrase-based orientation model during testing (i.e., as a model that orients phrases with respect to other phrases).

[4]It is not needed to store target-language indices onto the stack, since the decoder proceeds left to right, and thus successive blocks are always contiguous with respect to the target language.

Target phrase	Source	Op.	o_i	Stack
the russian side	[1]	S	M	
hopes	[2]	R	M	[1]
to	[3]	R	M	[1-2]
hold	[11]	S	D	[1-3]
consultations	[12]	R	M	[11], [1-3]
with iran	[9-10]	R	S	[11-12], [1-3]
on this	[6-7]	S	D	[9-12], [1-3]
issue	[8]	R,R	M	[6-7], [9-12], [1-3]
in the near future	[4-5]	R,R	S	[6-12], [1-3]
.	[13]	R,A	M	[1-12]

Figure 3: The application of the shift-reduce parsing algorithm for identifying hierarchical blocks. This execution corresponds to the decoding example of Figure 4. Operations (Op.) include shift (S), reduce (R), and accept (A). The source and stack columns contain source-language spans, which is the only information needed to determine whether two given blocks are contiguous. o_i is the label predicted by the hierarchical model by comparing the current block to the hierarchical phrase that is at the top of the stack.
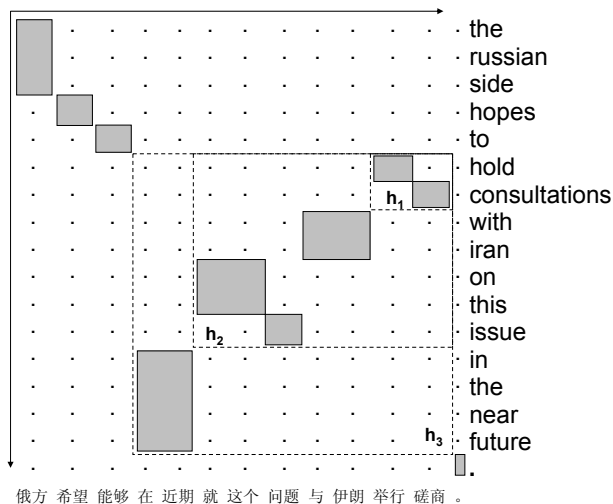

俄方 希望 能够 在 近期 就 这个 问题 与 伊朗 举行 磋商 。

Figure 4: Output of our phrase-based decoder using the hierarchical model on a sentence of MT06. Hierarchical phrases h_1 and h_2 indicate that *with Iran* and *in the near future* have a swap orientation. h_3 indicates that "*to*" and "." are monotone. In this particular example, distortion limit was set to 10.

It then encounters a discontinuity that prevents the next block [11] from being merged with [1-3]. As the decoder reaches the last words of the sentence (*in the near future*), [4-5] is successively merged with [6-12], then [1-3], yielding a stack that contains only [1-12].

A nice property of this parsing algorithm is that it does not worsen the asymptotic running time of beam-search decoders such as Moses (Koehn, 2004a). Such decoders run in time $O(n^2)$, where n is the length of the input sentence. Indeed, each time a partial translation hypothesis is expanded into a longer one, the decoder must perform an $O(n)$ operation in order to copy the coverage set (indicating which foreign words have already been translated) into the new hypothesis. Since this copy operation must be executed $O(n)$ times, the overall time complexity is quadratic. The incorporation of the shift-reduce parser into such a decoder does not worsen overall time complexity: whenever the decoder expands a given partial translation into a longer hypothesis, it simply copies its stack into the newly created hypothesis (similarly to copying the coverage vector, this is an $O(n)$ operation). Hence, the incorporation of the hierarchical models described in the paper into a phrase-based decoder preserves the $O(n^2)$ running time. In practice, we observe based on a set of experiments for Chinese-English and Arabic-English translation that our phrase-based decoder is on average only 1.35 times slower when it is running using hierarchical reordering features and the shift-reduce parser.

We finally note that the decoding algorithm presented in this section can only be applied left-to-right if the decoder itself is operating left-to-right. In order to predict orientations relative to the right-to-left hierarchical reordering model, we must resort to approximations at decoding time. We experimented with different approximations, and the one that worked best (in the experiments discussed in Section 6) is described as follows. First, we note that an analysis of the alignment grid often reveals that certain orientations are impossible. For instance, the block *issue* in Figure 4 can only have discontinuous orientation with respect to what comes next in English, since words surrounding the Chinese phrase have already been translated. When several hierarchical orientations are possible according to the alignment grid, we choose according to the following order of preference: (1) monotone, (2) swap, (3) discontinuous. For instance, in the case of *with iran* in Figure 4, only swap and discontinuous orientations are possible (monotone orientation is impossible because of the block *hold consultations*), hence we give preference to swap. This prediction turns out to be the correct one according to the decoding

steps that complete the alignment grid.

5 Discussion

We now analyze the system output of Figure 4 to further motivate the hierarchical model, this time from the perspective of the decoder. We first observe that the prepositional phrase *in the future* should rotate around a relatively large noun phrase headed by *consultations*. Unfortunately, localized reordering models such as (Tillman, 2004) have no means of identifying that such a displacement is a swap (S). According to these models, the orientation of *in the future* with respect to what comes previously is discontinuous (D), which is an uninformative fall-back category. By identifying h_2 (*hold ... issue*) as a hierarchical block, the hierarchical model can properly determine that the block *in the near future* should have a swap orientation.[5] Similar observations can be made regarding blocks h_1 and h_3, which leads our model to predict either monotone orientation (between h_3 and "*to*" and between h_3 and ".") or swap orientation (between h_1 and *with Iran*) while local models would predict discontinuous in all cases.

Another benefit of the hierarchical model is that its representation of phrases remains the same during both training and decoding, which is not the case for word-based and phrase-based reordering models. The deficiency of these local models lies in the fact that blocks handled by phrase-based SMT systems tend to be long at training time and short at test time, which has adverse consequences on nonhierarchical reordering models. For instance, in Figure 4, the phrase-based reordering model categorizes the block *in the near future* as discontinuous, though if the sentence pair had been a training example, this block would count as a swap because of the extracted phrase *on this issue*.

6 Results

In our experiments, we use a re-implementation of the Moses decoder (Koehn et al., 2007). Except for lexical reordering models, all other features are standard features implemented almost

exactly as in Moses: four translation features (phrase-based translation probabilities and lexically-weighted probabilities), word penalty, phrase penalty, linear distortion, and language model score. We experiment with two language pairs: Chinese-to-English (C-E) and Arabic-to-English (A-E). For C-E, we trained translation models using a subset of the Chinese-English parallel data released by LDC (mostly news, in particular FBIS and Xinhua News). This subset comprises 12.2M English words, and 11M Chinese words. Chinese words are segmented with a conditional random field (CRF) classifier that conforms to the Chinese Treebank (CTB) standard. The training set for our A-E systems also includes mostly news parallel data released by LDC, and contains 19.5M English words, and 18.7M Arabic tokens that have been segmented using the Arabic Treebank (ATB) (Maamouri et al., 2004) standard.[6]

For our language model, we trained a 5-gram model using the Xinhua and AFP sections of the Gigaword corpus (LDC2007T40), in addition to the target side of the parallel data. For both C-E and A-E, we manually removed documents of Gigaword that were released during periods that overlap with those of our development and test sets. The language model was smoothed with the modified Kneser-Ney algorithm, and we kept only trigrams, 4-grams, and 5-grams that respectively occurred two, three, and three times in the training data.

Parameters were tuned with minimum error-rate training (Och, 2003) on the NIST evaluation set of 2006 (MT06) for both C-E and A-E. Since MERT is prone to search errors, especially with large numbers of parameters, we ran each tuning experiment four times with different initial conditions. This precaution turned out to be particularly important in the case of the combined lexicalized reordering models (the combination of phrase-based and hierarchical discussed later), since MERT must optimize up to 26 parameters at once in these cases.[7] For testing,

---

[5]Note that the hierarchical phrase *hold ... issue* is not a well-formed syntactic phrase – i.e., it neither matches the bracketing of the verb phrase *hold ... future* nor matches the noun phrase *consultations ... issue* – yet it enables sensible reordering.

[6]Catalog numbers for C-E: LDC2002E18, LDC2003E07, LDC2003E14, LDC2005E83, LDC2005T06, LDC2006E26, and LDC2006E8. For A-E: LDC2007E103, LDC2005E83, LDC2006E24, LDC2006E34, LDC2006E85, LDC2006E92, LDC2007E06, LDC2007E101, LDC2007E46, LDC2007E86, and LDC2008E40.

[7]We combine lexicalized reordering models by simply treating them as distinct features, which incidentally increases the number of model parameters that must be tuned with MERT.
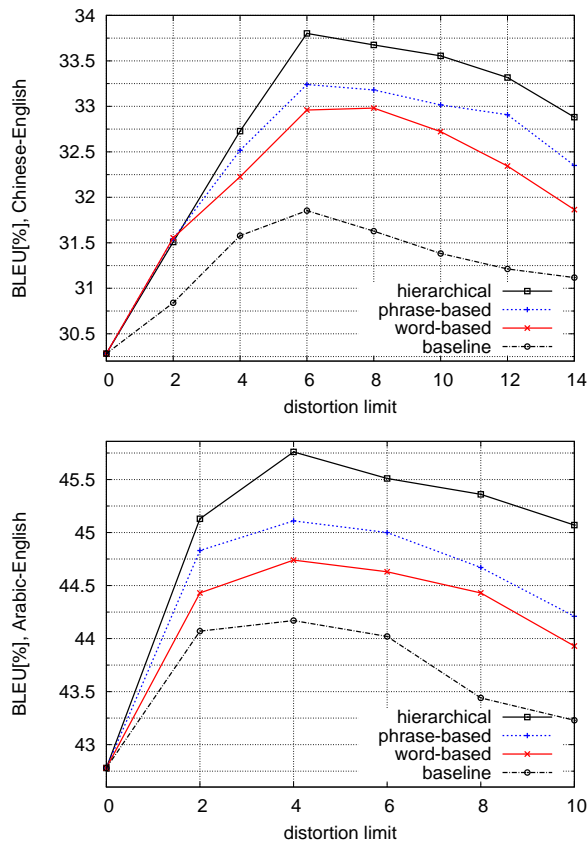
Figure 5: Performance on the Chinese-English and Arabic-English development sets (MT06) with increasing distortion limits for all lexicalized reordering models discussed in the paper. Our novel hierarchical model systematically outperforms all other models for distortion limit equal to or greater than 4. The baseline is Moses with no lexicalized reordering model.

we used the NIST evaluation sets of 2005 and 2008 (MT05 and MT08) for Chinese-English, and the test set of 2005 (MT05) for Arabic-English.

Statistical significance is computed using the approximate randomization test (Noreen, 1989), whose application to MT evaluation (Riezler and Maxwell, 2005) was shown to be less sensitive to type-I errors (i.e., incorrectly concluding that improvement is significant) than the perhaps more widely used bootstrap resampling method (Koehn, 2004b).

Tuning set performance is shown in Figure 5. Since this paper studies various ordering models, it is interesting to first investigate how the distor-

Lexicalized Reordering	MT06	MT05	MT08
none	31.85	29.75	25.22
word-based	32.96	31.45	25.86
phrase-based	33.24	31.23	26.01
hierarchical	33.80**	32.20**	26.38
phrase-based + hierarchical	33.86**	32.85**	26.53*

Table 3: BLEU[%] scores (uncased) for Chinese-English and the orientation categories $\{M, S, D\}$. Maximum distortion is set to 6 words, which is the default in Moses. The stars at the bottom of the tables indicate when a given hierarchical model is significantly better than all local models for a given development or test set (*: significance at the .05 level; **: significance at the .01 level).

Lexicalized Reordering	MT06	MT05	MT08
phrase-based	33.79	32.32	26.32
hierarchical	34.01	32.35	26.58
phrase-based + hierarchical	34.36**	32.33	27.03**

Table 4: BLEU[%] scores (uncased) for Chinese-English and the orientation categories $\{M, S, D_l, D_r\}$. Since the distinction between these four categories is not available in Moses, hence we have no baseline results for this case. Maximum distortion is set to 6 words.

tion limit affects performance.[8] As has been shown in previous work in Chinese-English and Arabic-English translation, limiting phrase displacements to six source-language words is a reasonable choice. For both C-E and A-E, the hierarchical model is significantly better ($p \leq .05$) than either other models for distortion limits equal to or greater than 6 (except for distortion limit 12 in the case of C-E). Since a distortion limit of 6 works reasonably well for both language pairs and is the default in Moses, we used this distortion limit value for all test-set experiments presented in this paper.

Our main results for Chinese-English are shown in Table 3. It appears that hierarchical models provide significant gains over all non-hierarchical models. Improvements on MT06 and MT05 are very significant ($p \leq .01$). In the case of MT08, significant improvement is reached through the combination of both phrase-based and hierarchical models. We often observe substantial gains when we combine such models, presumably because we get the benefit of identifying both local and long-distance swaps.

Since most orientations in the phrase-based model are discontinuous, it is reasonable to ask whether

---

[8]Note that we ran MERT separately for each distinct distortion limit.

LEXICALIZED REORDERING	MT06	MT05
none	44.03	54.87
word-based	44.64	54.96
phrase-based	45.01	55.09
hierarchical	45.51*	55.50*
phrase-based + hierarchical	45.64**	56.01**

Table 5: BLEU[%] scores (uncased) for Arabic-English and the reordering categories $\{M,S,D\}$.

LEXICALIZED REORDERING	MT06	MT05
phrase-based	44.74	55.52
hierarchical	45.53**	56.02**
phrase-based + hierarchical	45.63**	56.07**

Table 6: BLEU[%] scores (uncased) for Arabic-English and the reordering categories $\{M,S,D_l,D_r\}$.

the relatively poor performance of the phrase-based model is the consequence of an inadequate set of orientation labels. To try to answer this question, we use the set of orientation labels $\{M,S,D_l,D_r\}$ described in Section 3. Results for this different set of orientations are shown in Table 4. While the phrase-based model appears to benefit more from the distinction between left- and right-discontinuous, systems that incorporate hierarchical models remain the most competitive overall: their best performance on MT06, MT05, and MT08 are respectively 34.36, 32.85, and 27.03. The best non-hierarchical models achieve only 33.79, 32.32, and 26.32, respectively. All these differences (i.e., .57, .53, and .71) are statistically significant at the .05 level.

Our results for Arabic-English are shown in Tables 5 and 6. Similarly to C-E, we provide results for two orientation sets: $\{M,S,D\}$ and $\{M,S,D_l,D_r\}$. We note that the four-class orientation set is overall less effective for A-E than for C-E. This is probably due to the fact that there is less probability mass in A-E assigned to the D category, and thus it is less helpful to split the discontinuous category into two.

For both orientation sets, we observe in A-E that the hierarchical model significantly outperforms the local ordering models. Gains provided by the hierarchical model are no less significant than for Chinese-to-English. This positive finding is perhaps a bit surprising, since Arabic-to-English translation generally does not require many word order changes compared to Chinese-to-English translation, and this translation task so far has seldom benefited from hi-

erarchical approaches to MT. In our case, one possible explanation is that Arabic-English translation is benefiting from the fact that orientation predictions of the hierarchical model are consistent across training and testing, which is not the case for the other ordering models discussed in this paper (see Section 4). Overall, hierarchical models are the most effective on the two sets: their best performances on MT06 and MT05 are respectively 45.64 and 56.07. The best non-hierarchical models obtain only 45.01 and 55.52 respectively for the same sets. All these differences (i.e., .63 and .55) are statistically significant at the .05 level.

7 Conclusions and Future Work

In this paper, we presented a lexicalized orientation model that enables phrase movements that are more complex than swaps between adjacent phrases. This model relies on a hierarchical structure that is built as a by-product of left-to-right phrase-based decoding without increase of asymptotic running time. We show that this model provides statistically significant improvements for five NIST evaluation sets and for two language pairs. In future work, we plan to extend the parameterization of our models to not only predict phrase orientation, but also the length of each displacement as in (Al-Onaizan and Papineni, 2006). We believe such an extension would improve translation quality in the case of larger distortion limits. We also plan to experiment with discriminative approaches to estimating reordering probabilities (Zens and Ney, 2006; Xiong et al., 2006), which could also be applied to our work. We think the ability to condition reorderings on any arbitrary feature functions is also very effective in the case of our hierarchical model, since information encoded in the trees would seem beneficial to the orientation prediction task.

8 Acknowledgements

References

Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL (COLING/ACL)*, pages 529–536, Morristown, NJ, USA.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263–270, June.

Ronald L. Graham. 1972. An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters*, 1(4):132–133.

Liang Huang, Hao Zhang, Daniel Gildea, and Kevin Knight. 2008. Binarization of synchronous context-free grammars. Technical report, University of Pennsylvania.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*.

Philipp Koehn. 2004a. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Sixth Conference of the Association for Machine Translation in the Americas*, pages 115–124.

Philipp Koehn. 2004b. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. 2004. The Penn Arabic treebank: Building a large-scale annotated Arabic corpus.

I. Dan Melamed. 2003. Multitext grammars and synchronous parsers. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 79–86.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses. An Introduction*. Wiley, New York.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

F. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of HLT-NAACL*.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, June.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 101–104.

Taro Watanabe, Hajime Tsukada, and Hideki Isozaki. 2006. Left-to-right target generation for hierarchical phrase-based translation. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 777–784.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 521–528.

Richard Zens and Herman Ney. 2006. Discriminative reordering models for statistical machine translation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proceedings of the Workshop on Statistical Machine Translation*, pages 55–63, New York City, NY, June.