# Source-side Dependency Tree Reordering Models with Subtree Movements and Constraints

**Nguyen Bach, Qin Gao and Stephan Vogel**
InterACT, Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{nbach, qing, stephan.vogel}@cs.cmu.edu

## Abstract

We propose a novel source-side dependency tree reordering model for statistical machine translation, in which subtree movements and constraints are represented as reordering events associated with the widely used lexicalized reordering models. This model allows us to not only efficiently capture the statistical distribution of the subtree-to-subtree transitions in training data, but also utilize it directly at the decoding time to guide the search process. Using subtree movements and constraints as features in a log-linear model, we are able to help the reordering models make better selections. It also allows the subtle importance of monolingual syntactic movements to be learned alongside other reordering features. We show improvements in translation quality in English→Spanish and English→Iraqi translation tasks.

## 1 Introduction

Word movement is a defining characteristic of the machine translation problem. The fact that word order can change during translation makes the problem fundamentally different from related tasks such as tagging and automatic speech recognition. In fact, if one allows unrestricted changes in word order during translation, that alone is sufficient to show it to be NP complete, by analogy to the Traveling Salesman Problem (Knight, 1999). Despite the importance of word movement, the popular phrase-based translation paradigm (Koehn et al., 2003) devotes surprisingly little modeling capacity to the issue. A very simple reordering model is to base the cost for word movement only on the distance in the source sentence between the previous and the current word or phrase during the translation process. Later on, lexicalized reordering models, which condition the probability of phrase-to-phrase transitions on the words involved, have been proposed to address the word reordering issue (Tillman, 2004; Koehn et al., 2005; Al-Onaizan and Papineni, 2006; Kuhn et al., 2006). Alternatively, one can employ syntax in the modeling of movement. By viewing sentence in terms of its hierarchical structure, one can more easily expose regularities in the sorts of movement that occur during translation. A number of syntactic methods are driven by formal syntax alone (Wu, 1997; Chiang, 2005; Shen et al., 2008), while others employ linguistic syntax derived from a parse tree (Galley et al., 2004; Quirk et al., 2005; Liu et al., 2006). Each of these approaches requires a parser-like decoder, and represents a departure from phrase-based decoding. Galley and Manning (2008) demonstrated how to integrate hierarchical phrase structures to lexicalized reordering models.

The well-studied phrase-based architecture can also benefit from syntactic intuitions. Phrasal decoding can be augmented easily, either by syntactic pre-processing or through search-space constraints. Pre-processing approaches parse the source sentence and use the tree to apply rules which reorder the source into a more target-like structure before the translation begins. These rules can be learned (Xia and McCord, 2004; Rottmann and Vogel, 2007) or designed by hand (Collins et al., 2005; Wang et al., 2007; Xu et al., 2009). The pre-processing approach benefits from its simplicity and modularity, but it suffers from limitation of providing at most a first-best guess at syntactic movement. Search space constraints limit the phrasal decoder's translation search using syntactic intuitions. Zens et al.(2004) demonstrated how to incorporate formally syntactic binary-bracketing constraints into phrase-based decoding. Recently, it has been shown that syntactic cohesion, the notion that syntactic phrases in the source sentence tend to remain contiguous in the target (Fox, 2002), can be incorporated into phrasal decoding as well, by following the simple intuition that any source subtree that has begun translation, must be completed before translating another part of the tree (Cherry, 2008; Yamamoto et al., 2008).

In this paper, we introduce a novel reordering model for phrase-based systems which exploits dependency subtree movements and constraints. In order to do, we

must first consider several questions. Should subtree movements be conditioned on source dependency structures? How can we estimate reliable probability distributions from training data? How do we incorporate the reordering model with dependency structures and cohesive constraints into a phrase-based decoder? We investigate these questions by presenting the model, training and decoding procedure in Section 2. Furthermore, we present experimental results on English-Iraqi and English-Spanish systems in Section 3. Finally, we investigate the impact of the proposed models in Section 4 .

## 2 Source-tree Reordering Models

Nowadays most statistical machine translation systems are based on log-linear model which tries to provide a parameterized form of the probability of translating a sentence $f_1^J$ to $e_1^I$, subject to

$$\hat{e}_1^{\hat{I}} = \arg\max_{\{e_1^I\}} P(e_1^I|f_1^J) \quad (1)$$

$P(e_1^I|f_1^J)$ can be modeled as a log-linear model with components $h_m(.)$ and scaling factors $\lambda_m$:

$$\begin{aligned} \hat{e}_1^{\hat{I}} &= \arg\max_{\{e_1^I\}} P(e_1^I|f_1^J) \quad (2) \\ &= \arg\max_{\{e_1^I\}} \exp[\sum_1^M \lambda_m h_m(e_1^I, f_1^J)] \end{aligned}$$

A common feature set includes reordering models which provide the decoder the capability to determine the orientation sequence of phrases. The beam search strategy is used during decoding, in which the intermediate states correspond to partial translations. The decoding process advances by extending a state with the translation of a source phrase and the final state is reached when each source word has been translated exactly once. Reordering occurs when the source phrase to be translated does not immediately follow the previously translated phrase. The reordering is integrated into the target function by using discriminatively-trained distortion penalties, such as the widely used lexicalized reordering model (Koehn et al., 2005). It can be parameterized as follows:

$$p(O|e,f) = \prod_{i=1}^{n} p(o_i|\bar{e_i}, \bar{f_{a_i}}, a_{i-1}, a_i) \quad (3)$$

where **f** is the input sentence; $\mathbf{e} = (\bar{e_1}, \ldots, \bar{e_n})$ is the target language phrases; $\mathbf{a} = (a_1, \ldots, a_n)$ is phrase alignments; $\bar{f_{a_i}}$ is a source phrase which has a translated phrase $\bar{e_i}$ defined by an alignment $a_i$. **O** is the orientation sequence of phrase where each $o_i$ has a value over three possible orientations, (*M*) monotone, (*S*) swap with previous phrase, or (*D*) discontinuous. **O**=$\{M, S, D\}$ and is defined as follows:

$$o_i = \begin{cases} M & \text{if } a_i - a_{i-1} = 1 \\ S & \text{if } a_i - a_{i-1} = -1 \\ D & \text{if } |a_i - a_{i-1}| \neq 1 \end{cases} \quad (4)$$

### 2.1 Models

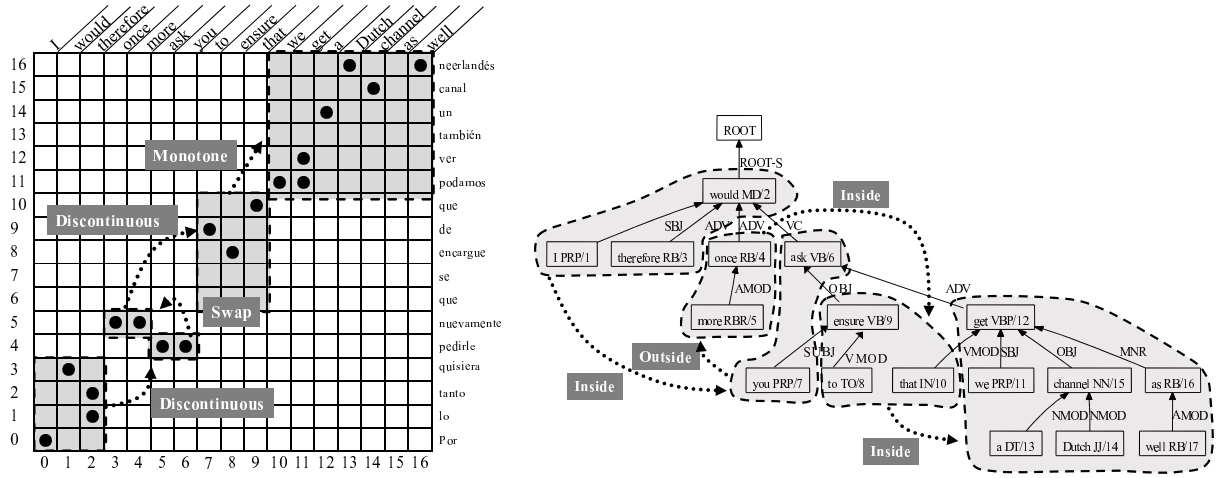
A lexicalized reordering model is defined in terms of transitions between phrases - two phrases in sequence, $previous$ and $next$, have a specific relationship to each other, such as $monotone$, $swap$ or $discontinuous$. Statistics on those relationships make up the model.

Lexicalized reordering models are well-defined for flat word surface structures. However, the models do not leverage source-side syntactic structures which are always available during the decoding time. Previous studies, such as Cherry (2008), show improvements when using source-side dependency structures as soft cohesive constraints. Cohesion constraints tell the decoder which cohesive movements are available, but the decoder has no opinion on the likelihood of these moves.

In a source-tree reordering model, we would condition monolingually and syntactically phrase movements on the source dependency tree. A source-tree reordering model considers in terms of previous source dependency structures. One can think about the phrase movements as the movement of the subtree $inside$ or $outside$ a source subtree when the decoder is leaving from the *previous* source state to the current source state. The notions of moving $inside$ (**I**) and $outside$ (**O**) a subtree can be interpreted as tracking facts about the subtree-to-subtree transitions observed in the source side of word-aligned training data. With extra guidance on subtree movements, our expectation is that source-tree reordering models will help the decoder make smarter distortion decisions.

An example of the source-tree reordering movements is illustrated in Figure 1 that contains a word/phrase alignment matrix of a English-Spanish sentence pair, source-dependency tree and reordering movements. The lexicalized orientation sequence is {D, S, D, M} while the subtree movement sequence is {I, O, I, I}. The lexicalized reordering model assigned $D$ for phrase "*ask you*" because the previous extracted phrase "*I would therefore*" was not continuous with "*ask you*". At the same time, the source-tree movement assigned *I* since "*ask you*" is moving $inside$ the subtree rooted at "*would*". In addition, "*once more*" received *O* from the source-tree reordering model since it is $swap$ with "*ask you*" and moving $outside$ the subtree rooted at "*ask*".

Let $T$ denote the source dependency tree and $T(n)$ stands for the subtree rooted at node $n$. A span $\bar{f}$ indicates the last source phrase translated to create the current state and each $\bar{f}$ has a dependency structure $s_h$. A subtree $T(n)$ covers a span of contiguous source words is constructed by dependency structures $s_h$; for subspan $\bar{f}$




(a) Alignment matrix with lexicalized orientation events

(b) Inside/Outside subtree movements on the source dependency tree

Figure 1: Source-tree reordering extraction examples for the English-Spanish sentence pair "*I would therefore once more ask you to ensure that we get a Dutch channel as well*"- "*Por lo tanto quisiera pedirle nuevamente que se encargue de que podamos ver tambin un canal neerlands*"

covered by $T(n)$, we say $\bar{f} \in T(n)$. We define a subtree that has begun translation but not yet complete, an *open* subtree. On the other hand, when all words under a node have been translated then we call a *completed* subtree. A phrase $\bar{f}$ is moving ***inside (I)*** a $T(n)$ if $\bar{f}$ helps $T(n)$ to be completed, in other words, $T(n)$ covers more contiguous words. A phrase $\bar{f}$ is moving ***outside (O)*** a $T(n)$ if $\bar{f}$ leaves $T(n)$ to be open, in other words, $T(n)$ contains some words which have not been covered yet. $inside$ and $outside$ are the two subtree movements we are going to model.

Mathematically speaking, a source-tree reordering model is defined as follows:

$$p(D|e,f) = \prod_{i=1}^{n} p(d_i|\bar{e_i}, \bar{f}_{a_i}, a_i, s_{i-1}, s_i) \qquad (5)$$

where $s_i$ and $s_{i-1}$ are dependency structures of source phrases $\bar{f}_{a_i}$ and $\bar{f}_{a_{i-1}}$ respectively; **D** is a random variable which represents the sequence of syntactical phrase movements over the source dependency tree; each $d_i$ takes a value either *inside* (**I**) or *outside* (**O**). $p(D|e,f)$ is the probability of the subtree movement likelihood over the source phrase sequence and their target movements. Since the model essentially constraints phrase movements on the source dependency tree however it does not explicitly provide orientations for a phrase-based decoder. Therefore, we combine our model with the lexicalized reordering model, as a result, a set of events contains $D = o_k\_d_j$ = {*M_I, S_I, D_I, M_O, S_O, D_O*}. The source dependency tree is used here to refine the reordering events provided by a lexicalized reordering model. Finally, the source-tree reordering model is derived as follows:

$$p(D|e,f) = \prod_{i=1}^{n} p((o\_d)_i|\bar{e_i}, \bar{f}_{a_i}, a_{i-1}, a_i, s_{i-1}, s_i) \quad (6)$$

## 2.2 Training

To train the model, the system needs to extract $o_k\_d_j$ events for phrase pairs. First, the source side dependency trees of the bilingual training data are provided by using a dependency parser. Given a sentence pair and source dependency tree, when performing the phrase-extract algorithm (Och and Ney, 2004) we also extract the source dependency structure of each phrase pair. The values of $o_k$ are obtained by lexicalized reordering models. To determine whether the current source phrase is moving $inside$ or $outside$ a subtree $T(n)$ with respect to previously extracted phrases we apply the exhaustive interruption check algorithm (Bach et al., 2009). This algorithm essentially walks through the dependency subtrees of previously extracted phrases and checks whether the subtree is open or completed. The value of $d_j$ is $I$ when the exhaustive interruption check algorithm returns false and $O$ otherwise.

Table 1 is a snapshot of the output of the reordering extraction procedure. The third column shows source-tree reordering features.

Table 2 displays the overall event distributions of source-tree reordering models. It appears clearly that occurrences of $S\_I$ and $S\_O$ are too sparsely seen in the training data which assigns nearly 98% of its probability mass to other events. The table strongly suggests that from training data the source-tree reordering models

| | Lexicalized | Source-tree |
|---|---|---|
| ... | | |
| ask you # pedirle | dis swap | D_I $*$ |
| ask you # pedirle | mono mono | M_I |
| ask you # pedirle | mono mono | M_O |
| once more # nuevamente | swap dis | S_O $*$ |
| once more # nuevamente | dis swap | D_O |
| once more # nuevamente que | swap dis | S_O |
| ... | | |

Table 1: Extracted reordering events; $*$ indicates events extracted from the example in Figure 1

observed $monotone$ and $inside$ movements more often than other categories.

| | *M_I* | *S_I* | *D_I* | *M_O* | *S_O* | *D_O* |
|---|---|---|---|---|---|---|
| En-Es | 0.38 | 0.01 | 0.14 | 0.3 | 0.01 | 0.15 |
| En-Ir | 0.62 | 0.01 | 0.13 | 0.16 | 0.01 | 0.07 |

Table 2: Distributions of the six source-tree reordering events estimated from English-Spanish and English-Iraqi training data

After having all extracted phrase pairs with dependency features, we need to estimate parameters of source-tree reordering models for a particular pair $p((o_j\_d_k)_i|\bar{e_i}, \bar{f}_{a_i})$. An event, such as *M_I*, can be interpreted by three possibilities. First, *M_I* is a joint probability of $monotone$ and $inside$ given a phrase pair. Second, *M_I* can be a conditional probability of $monotone$ given a phrase pair and it is $inside$. Finally, *M_I* can be a conditional probability of $inside$ given a phrase pair and it is $monotone$. The parameter $p((o_j\_d_k)_i|\bar{e_i}, \bar{f}_{a_i})$ is estimated by the maximum likelihood estimation criteria with a smoothing factor $\gamma$ as

$$p((o_j\_d_k)_i|\bar{e_i}, \bar{f}_{a_i}, o_j, d_k) = \frac{count(o_k\_d_j) + \gamma}{\sum_k \sum_j (count(o_k\_d_j) + \gamma)} \quad (7)$$

if it is a joint probability of subtree movements and lexicalized orientations (*DO*) or

$$p((o_j\_d_k)_i|\bar{e_i}, \bar{f}_{a_i}, d_k) = \frac{count(o_k\_d_j) + \gamma}{\sum_k (count(o_k\_d_j) + \gamma)} \quad (8)$$

if it is conditioned on subtree movements (*DOD*) or

$$p((o_j\_d_k)_i|\bar{e_i}, \bar{f}_{a_i}, o_j) = \frac{count(o_k\_d_j) + \gamma}{\sum_j (count(o_k\_d_j) + \gamma)} \quad (9)$$

if it is conditioned on lexicalized orientations (*DOO*).

Table 3 displays source-tree reordering estimated probabilities for a phrase pair "*ask you*"- "*pedirle*". Each probability was put under one of the three parameter estimation methods.

| | *M_I* | *S_I* | *D_I* | *M_O* | *S_O* | *D_O* |
|---|---|---|---|---|---|---|
| DO | 0.691 | 0.003 | 0.142 | 0.119 | 0.009 | 0.038 |
| DOD | 0.827 | 0.003 | 0.170 | 0.719 | 0.053 | 0.228 |
| DOO | 0.854 | 0.250 | 0.790 | 0.146 | 0.750 | 0.210 |

Table 3: $inside$ and $outside$ probabilities for phrase "*ask you*"- "*pedirle*" according to three parameter estimation methods

### 2.3 Decoding

The beam search strategy is unchanged from the phrase-based system. Our proposed source-tree reordering models concern monolingualy and syntactically movements in the source sentence. However, computing source-tree reordering model scores can be done in two scenarios 1) not using and 2) using cohesive constraints. Cohesive constraints can be enforced by the interruption check algorithm (Cherry, 2008; Bach et al., 2009). One can consider the first scenario as the decoder does not have any information about the source dependency tree during decoding time, therefore, we allow the decoder to consider both events $inside$ and $outside$. The decision of selecting a preferable feature is made by the tuning procedure. On the other hand, when the source dependency tree is available, subtree movements are informed to the decoder via cohesive constraints, as a result, we are able to allow the decoder to make a harder choice to consider either $inside$ or $outside$.

More specifically, if the decoder chooses to decode without cohesive constraints then after detecting the orientation of the current phrase, for example $swap$, the decoder will trigger two subtree movement features *S_I* and *S_O* and sum up both features in the log-linear combination. In other words, the decoder considers both events that the current phrase is moving $inside$ and $outside$ a subtree $T(n)$ given it is $swap$ orientation on flat word structures.

In the second scenario, the decoder uses cohesive constraints after detecting the orientation of the current phrase, for example $swap$. The decoder only considers one source-tree reordering feature. The choice of feature depends on the output of the interruption check algorithm on the current phrase. If the return is $inside$ then *S_I* will be used otherwise *S_O*.

## 3 Experimental Results

We built baseline systems using GIZA++ (Och and Ney, 2003), Moses' phrase extraction with the grow-diag-final-and heuristic (Koehn et al., 2007), a standard phrase-based decoder (Vogel, 2003), the SRI LM toolkit (Stolcke, 2002), the suffix-array language model (Zhang and Vogel, 2005), a lexicalized reordering model with a reordering window of 3, and the maximum number of target phrases restricted to 5. Results are reported using

lowercase BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). All model weights were trained on development sets via minimum-error rate training (MERT) (Venugopal and Vogel, 2005) with an unique 200-best list and optimizing toward BLEU. To shorten the training time, a multi-threaded GIZA++ version was used to utilize multi-processor servers (Gao and Vogel, 2008). We used the MALT parser (Nivre et al., 2006) to get English dependency trees. We perform experiments on English→Spanish and English→Iraqi tasks. Detailed corpus statistics are shown in Table 4.

| | English→Spanish | | English→Iraqi | |
|---|---|---|---|---|
| | English | Spanish | English | Iraqi |
| sent. pairs | 1,310,127 | | 654,556 | |
| uniq. pairs | 1,287,016 | | 510,314 | |
| avg. sent. length | 27.4 | 28.6 | 8.4 | 5.9 |
| # words | 35.8 M | 37.4 M | 5.5 M | 3.8 M |
| vocabulary | 117 K | 173 K | 34 K | 109 K |

Table 4: Corpus statistics of English→Spanish and English→Iraqi systems

We experiment systems in different configurations of the source-tree reordering model such as DO, DOD and DOO means parameters estimation using Equation 7, 8 and 9 respectively. Moreover, Coh means the decoder triggers cohesive constraints for source-tree reordering models (Cherry, 2008). Bold type is used to indicate highest scores.

Our first step in validating the proposed approach is to check with the English→Spanish system. We used the Europarl and News-Commentary parallel corpora for English→Spanish as provided in the ACL-WMT 2008[1] shared task evaluation. We built the baseline system using the parallel corpus restricting sentence length to 100 words for word alignment and a 4-gram SRI LM with modified Kneyser-Ney smoothing. We used nc-devtest2007(ncd07) as the development set; nc-test2007 (nct07) as in-domain and newstest2008 (net08) as out-domain held-out evaluation sets. Each test set has 1 translation reference. Table 5 shows that the best obtained improvements are **+0.8** BLEU point and **-1.4** TER score on the held-out evaluation sets. Moreover, the proposed methods also obtained improvements on the out-domain test set (net08).

We also validated the proposed approach on English→Iraqi. However, we have a smaller training corpus which comes from force protection domains and is spoken language style. This data is used in the DARPA TransTac program. The English→Iraqi pair also differs according to the language family. English is an Indo-European language while Iraqi is a Semitic language of the Afro-Asiatic language family.

[1]http://www.statmt.org/wmt08

| | nct07 | | net08 | |
|---|---|---|---|---|
| | BLEU | TER | BLEU | TER |
| Baseline | 32.89 | 65.25 | 20.11 | 83.09 |
| Coh | 33.33 | 64.72 | 19.80 | 82.84 |
| DO | 32.99 | 65.05 | 20.27 | 82.65 |
| DO+Coh | 33.28 | 64.77 | 20.61 | 82.35 |
| DOD | 33.17 | 64.54 | 20.33 | 82.12 |
| DOD+Coh | 33.46 | 64.41 | 20.58 | 82.05 |
| DOO | 33.10 | 64.51 | 20.51 | 82.12 |
| DOO+Coh | **33.67** | **64.03** | **20.71** | **81.70** |

Table 5: Scores of baseline and improved baseline systems with source-tree reordering models on English→Spanish

| | june08 | | nov08 | |
|---|---|---|---|---|
| | BLEU | TER | BLEU | TER |
| Baseline | 25.18 | 56.70 | 18.40 | 62.91 |
| Coh | 25.34 | 57.30 | 18.01 | 61.52 |
| DO | 25.31 | 57.30 | 18.43 | 60.98 |
| DO+Coh | 25.53 | 57.20 | 19.13 | 61.45 |
| DOD | 25.34 | 57.53 | 18.90 | 61.81 |
| DOD+Coh | 25.50 | **56.29** | **19.15** | 60.93 |
| DOO | 25.25 | 56.76 | 18.40 | **60.64** |
| DOO+Coh | **25.58** | 56.37 | 18.59 | 61.45 |

Table 6: Scores of baseline and improved baseline systems with source-tree reordering models on English→Iraqi

We used 429 sentences of TransTac T2T July 2007 (july07) as the development set; 656 sentences of TransTac T2T June 2008 (june08) and 618 sentences of November 2008 (nov08) as the held-out evaluation sets. Each test set has 4 reference translations. We used a suffix-array LM up to 6-gram with Good-Turing smoothing. In Table 6, source-tree reordering models produced the best improvements of **+0.8** BLEU point and **-2.3** TER score on the held-out evaluation sets.

## 4 Discussion and Analysis

In this section we perform detail error analysis from where different scenarios emerge and questions arise for our assumptions.

### 4.1 Breakdown improvement analysis

As we can see from the results, there are improvements on all the different test sets. However, one could expect that the methods may work for a portion of the data but not others. We divide the test sets into three portions based on sentence-level TER of the baseline system. Let $\mu$ and $\sigma$ be the mean and standard deviation of the sentence-level TER of the whole test set. We define three subsets $head$, $tail$ and $mid$ as the sentence whose TER score is lower than $\mu - \frac{1}{2}\sigma$, higher than $\mu + \frac{1}{2}\sigma$ and the rest, respectively. We then fix the division of the three subsets, and calculate the BLEU and TER scores on them

| | | En-Ir | | | | En-Es | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | jun08 | | nov08 | | nc07 | | nt08 | |
| System | | BLEU | TER | BLEU | TER | BLEU | TER | BLEU | TER |
| Baseline | $tail$ | 29.45 | 76.50 | 24.41 | 87.69 | 23.36 | 92.93 | 24.41 | 134.04 |
| | $mid$ | 38.61 | 53.60 | 35.89 | 61.07 | 31.08 | 66.75 | 22.61 | 86.32 |
| | $head$ | 61.38 | 25.80 | 60.90 | 28.16 | 44.58 | 47.45 | 35.34 | 59.54 |
| Coh | $tail$ | **+0.56** | **+1.35** | **+1.29** | **+5.27** | **+0.67** | **+1.80** | **+0.07** | **+1.27** |
| | $mid$ | **+0.14** | -0.91 | **+0.48** | **+1.08** | **+0.22** | **+0.07** | -0.02 | -0.19 |
| | $head$ | **+0.37** | -1.69 | -3.11 | -4.68 | -0.17 | -0.73 | -0.48 | **+1.27** |
| DO | $tail$ | **+0.28** | **+0.66** | **+1.91** | **+7.03** | **+0.49** | **+1.94** | **+0.87** | **+2.32** |
| | $mid$ | **+0.07** | -1.15 | **+0.58** | **+1.44** | **+0.24** | **+0.45** | **+0.12** | **+0.28** |
| | $head$ | -0.28 | -2.48 | -1.31 | -3.07 | -0.28 | -0.71 | -0.11 | -0.77 |
| DO+Coh | $tail$ | **+1.07** | **+1.95** | **+1.72** | **+5.19** | **+0.66** | **+1.78** | **+0.52** | **+1.60** |
| | $mid$ | **+0.80** | -0.85 | **+0.92** | **+1.32** | **+0.19** | **+0.21** | **+0.13** | **+0.25** |
| | $head$ | -0.37 | -2.41 | -1.59 | -3.62 | -0.25 | -0.75 | -0.01 | -1.11 |
| DOD | $tail$ | **+0.46** | **+0.06** | **+1.96** | **+4.84** | **+0.35** | **+1.91** | **+0.75** | **+2.84** |
| | $mid$ | **+0.53** | -1.35 | **+0.43** | **+0.29** | **+0.01** | -0.15 | **+0.05** | **+0.41** |
| | $head$ | **+0.27** | -1.03 | -0.61 | -2.33 | -0.79 | -1.33 | -0.37 | -1.37 |
| DOD+Coh | $tail$ | **+1.19** | **+2.70** | **+2.10** | **+5.89** | **+0.49** | **+0.43** | **+0.27** | **+1.30** |
| | $mid$ | **+0.44** | -0.37 | **+0.42** | **+1.16** | **+0.01** | -0.85 | **+0.12** | **+0.99** |
| | $head$ | **+0.32** | -1.25 | -0.66 | -2.02 | -0.37 | -1.35 | -0.26 | -2.05 |
| DOO | $tail$ | **+1.18** | **+2.41** | **+2.37** | **+7.36** | **+0.35** | **+1.92** | **+0.59** | **+0.39** |
| | $mid$ | **+0.13** | -0.62 | **+0.28** | **+1.83** | **+0.01** | -0.15 | **+0.06** | -0.38 |
| | $head$ | -0.50 | -2.13 | -0.58 | -2.63 | -0.79 | -1.34 | -0.47 | -1.52 |
| DOO+Coh | $tail$ | **+1.28** | **+2.70** | **+2.03** | **+5.88** | **+0.65** | **+1.61** | **+0.69** | **+1.10** |
| | $mid$ | **+0.74** | -0.52 | **+0.19** | **+0.82** | **+0.18** | -0.02 | **+0.12** | -0.05 |
| | $head$ | **+0.22** | -1.02 | -1.61 | -4.16 | -0.40 | -1.07 | -0.22 | -1.00 |

Table 7: Distribution of improvements over different portions of the test sets, where for TER the sign is reversed so that positive numbers means improve in TER, i.e., lower TER score. The improvements are marked by bold text.

for every system. From Table 7, the proposed methods tend to output better TER and BLEU for the $tail$ subsets, the improvements on the $mid$ subsets are smaller, and loss can be observed on the $head$ subsets. The splitting of different sets also reflects on the length of sentences, as shown in Table 8, the tail parts are generally long sentences. The breakdown analysis suggests a more subtle model taking into account the sentence lengths could bring in more improvements, especially, on the $tail$ set in which the baseline model loses.

| | jun08 | nov08 | nc07 | nt08 |
|---|---|---|---|---|
| $head$ | 7.92 | 6.27 | 20.39 | 13.07 |
| $mid$ | 12.31 | 11.09 | 28.07 | 22.78 |
| $tail$ | 13.91 | 14.08 | 35.29 | 25.33 |

Table 8: Average reference lengths

### 4.2 Interactions of reordering models

To further investigate the impact of the proposed models, we perform several analyses to examine whether there are significant differences in 1) the average phrase length that the decoder outputs; 2) the total number of reorderings occurred in the hypothesis and 3) the average reordering distance for all the reordering events. Table 9 shows the statistics on the four aspects for all the test sets. For the average phrase length, we can observe a smaller value when applying the proposed models on English-Spanish tasks. However, on English-Iraqi the picture is contradicting when on one set the phrase length is generally longer and on the other set both longer and shorter statistics can be observed in different systems. Generally, there is no evidence to support a claim that the proposed models have consistent impact on the length of phrases chosen by the decoder. The observation is not surprising since the proposed reordering models are more likely to affect the decoder's behavior on reorderings.

When analyzing the average reordering distance, a more consistent picture can be discovered. The average reordering distance is larger than the corresponding systems with only inside/outside subtree movements. Whereas we cannot observe similar phenomenon comparing the system with only cohesive constraints and the baseline, which indicates that the cohesive constraints actually have the effect of restricting long distance reorder generated by the inside/outside subtree movements. The most interesting observation is the *number of reorderings* in the hypothesis. To make it easier to think about how sparse the reordering events are, we present the occurrence rate of reorderings, i.e. the number of words divided by the number of reorderings, as listed in the parentheses inside Table 9. An interesting phenomenon is that in English-Iraqi tasks, the output is generally monotone in the baseline, and the number of reorderings increases dramatically by applying the inside/outside subtree movements. However, solely applying cohesive constraints does not increase the number of reorderings. In English-Spanish tasks, although all the features generate more reordering events than the baseline, applying only the cohesion constraints also increases the number of reorderings dramatically.

When combining the statistics of Table 9 the most

| | Number of Reorderings | | | | Frequency of Reordering | | | | Average Phrase Length | | | | Average Reordering Distance | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | En-Es | | En-Ir | | En-Es | | En-Ir | | En-Es | | En-Ir | | En-Es | | En-Ir | |
| | nc07 | nt08 | jun08 | nov08 | nc07 | nt08 | jun08 | nov08 | nc07 | nt08 | jun08 | nov08 | nc07 | nt08 | jun08 | nov08 |
| Baseline | *1507* | **1684** | *39* | *24* | 16.3 | 16.4 | 119 | 164 | 2.02 | 1.80 | 2.20 | 2.34 | 2.61 | 2.44 | 2.79 | 2.17 |
| Coh | 2045 | **2903** | 46 | 21 | 10.0 | 12.8 | 99 | 178 | 1.90 | 1.71 | 2.25 | 2.48 | 2.67 | 2.58 | 2.81 | 2.50 |
| DO | **2189** | 2113 | 97 | 58 | 11.6 | 13.4 | 47 | 64 | 1.95 | 1.76 | 2.25 | 2.47 | 2.57 | 2.46 | 2.88 | 3.05 |
| DO+Coh | 1929 | 1900 | 155 | 88 | 13.6 | 15.3 | 30 | 44 | 1.89 | 1.71 | 2.17 | 2.37 | 2.47 | 2.33 | 2.74 | 2.88 |
| DOD | 1735 | 2592 | 123 | 60 | 14.9 | 10.7 | 38 | 65 | 1.92 | 1.88 | 2.17 | 2.36 | 2.73 | 2.57 | 2.79 | 2.93 |
| DOD+Coh | 2070 | 2021 | 148 | **90** | 12.8 | 14.5 | 32 | 43 | 1.88 | 1.70 | 2.18 | 2.37 | 2.50 | 2.39 | 2.64 | 2.81 |
| DOO | 1735 | 1785 | 164 | 49 | 14.9 | 16.1 | 30 | 79 | 1.92 | 1.73 | 2.10 | 2.37 | 2.73 | 2.60 | 2.72 | 2.98 |
| DOO+Coh | 1818 | 1959 | **247** | 66 | 14.1 | 14.6 | 19 | 59 | 1.93 | 1.74 | 2.15 | 2.37 | 2.53 | 2.42 | 2.64 | 2.88 |

Table 9: Statistics on four aspects of the final hypothesis over different systems; 1. the number of reorderings, 2. the number of words in the hypotheses divided by the number of reordering, i.e. a larger number means more sparse observation of reorderings, 3. the average phrase length and, 4. the average reordering distance

significant effect the source-tree reordering models contribute is the number of reorderings. Instead of constraining the reordering, the models enable more reorderings to be generated. As shown in Table 2, in the training data there are generally more reorderings than we observed in the decoding results. It indicates the baseline reordering model is not subtle enough to encode accurately information in a more generalized way, so that more reorderings can be generated without losing performance. The source-tree reordering models provide a more discriminative mechanism to estimate reordering events. For example, in Table 2 the probability mass of monotone and discontinuous events are different given that the phrase is encoded with inside or outside subtree movements. Moreover, the reordering issue is more language-specific than general translation models, and the conditions for a reordering event to happen vary among languages. Providing more features that are conditioned on different information, such as include inside/outside subtree movements and cohesive constraints presented in this paper, could benefit the system performance by enabling MERT to choose the most prominent ones from a larger basis.

### 4.3 The effect of inside/outside events

| | En-Es | | En-Ir | |
|---|---|---|---|---|
| | nc07 | nt08 | jun08 | nov08 |
| Baseline | 29.35 | 38.52 | 9.30 | 9.39 |
| Coh | 20.23 | 29.40 | 8.23 | 8.90 |
| DO | 30.34 | 32.57 | 12.35 | 11.65 |
| DO+Coh | 12.26 | 13.07 | 15.40 | 13.11 |
| DOD | 32.39 | 37.64 | 12.65 | 11.00 |
| DOD+Coh | 15.94 | 23.99 | 11.89 | 11.97 |
| DOO | 28.75 | 32.08 | 12.35 | 11.65 |
| DOO+Coh | 18.44 | 25.50 | 16.77 | 10.68 |

Table 10: The percentage of sentences having $outside$ subtree events

All the analysis above inspired us to carry out a more direct analysis of the decoder behaviors. As the main motivation of the proposed approach is to model the behavior of $inside$/$outside$ subtree events, natural assumptions could be that 1) different target languages should have different probabilities of generating a sequence that has outside subtree events on the same source language and 2) whether the model could change the behavior of generating outside subtree events. Further more, comparing to baseline system, do the changes, i.e. generating more or less outside subtree events than baseline, bring improvements to those sentences? From Table 10, the number of sentences having outside subtree events has not changed much when decoding with subtree movement features in English-Spanish tasks, while this number generally increases in English-Iraqi tasks. Moreover, when decoding with both subtree movements and cohesive constraints, we observe that the number of sentences having outside subtree events sharply decreases, whereas it increases in English-Iraqi. This result shows an interesting correlation with the performance improvements in Table 5 and 6, where the systems with cohesive constraints generally outperform those without. If we consider the cohesive constraints as hard constraints, then the outside subtree events are considered as violations, however in English-Iraqi tasks, the performance becomes better with more "violations". The observation further consolidates our suggestion that subtle models should be preferred for future developments, because the features may encode the information that the violation of constraints is actually preferred, no matter whether it is because of the nature of the particular language or the style of the source (spoken, written, etc.).

## 5 Conclusions and Future Work

In this study, our major contribution is a novel source-tree reordering model that exploits dependency subtree movements and constraints. These movements and constraints enable us to efficiently capture the subtree-to-subtree transitions observed both in the source of word-aligned training data and in decoding time. Representing subtree movements as features allows MERT to train the corresponding weights for these features relative to others in the model. We show that this model provides improvements for four held-out evaluation sets and for two language pairs. In future work, we plan to extend the parameterization of our models to explicitly represent source-side subtree movements during the decoding time.

We also plan to combine our models with the hierarchical phrase reordering model (Galley and Manning, 2008). We believe such extensions will generalize more subtle reordering events on source dependency trees.

## Acknowledgments

This work is in part supported by the US DARPA TransTac and GALE programs. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA. We would like to thank Colin Cherry for fruitful discussions and anonymous reviewers for helpful comments.

## References

Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of ACL-COLING'06*, pages 529–536, Sydney, Australia.

Nguyen Bach, Stephan Vogel, and Colin Cherry. 2009. Cohesive constraints in a beam search phrase-based decoder. In *Proceedings of NAACL-HLT'09*, Boulder, Colorado, May/June. Association for Computational Linguistics.

Colin Cherry. 2008. Cohesive phrase-based decoding for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 72–80, Columbus, Ohio, June. Association for Computational Linguistics.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL'05*, pages 263–270, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL'05*, pages 531–540, Ann Arbor, USA, June.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of EMNLP'02*, pages 304–311, Philadelphia, PA, July 6-7.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of EMNLP'08*, Hawaii, USA.

Michel Galley, Mark Hopkins, Kevin Knight, and Marcu Daniel. 2004. What's in a translation rule? In *Proceedings of HLT-NAACL'04*, Boston, USA, May.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Proceedings of the ACL 2008 Software Engineering, Testing, and Quality Assurance Workshop*, Columbus, Ohio, USA.

Kevin Knight. 1999. Squibs and discussions: Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615, December.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based machine translation. In *Proceedings of HLT-NAACL'03*, pages 48–54, Edmonton, Canada.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the IWSLT'05*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris 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 ACL'07*, pages 177–180, Prague, Czech Republic, June.

Roland Kuhn, Denis Yuen, Michel Simard, Patrick Paul, George Foster, Eric Joanis, and Howard Johnson. 2006. Segment choice models: Feature-rich models for global distortion in statistical machine translation. In *Proceedings of HLT-NAACL'06*, pages 25–32, New York, NY.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of ACL'06*, pages 609–616, Morristown, NJ, USA. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC'06*, Genoa, Italy.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 1:29, pages 19–51.

Franz J. Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. In *Computational Linguistics*, volume 30, pages 417–449.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL'02*, pages 311–318, Philadelphia, PA, July.

Chris Quirk, Aruk Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of ACL'05*, pages 271–279, Ann Arbor, USA, June.

Kay Rottmann and Stephan Vogel. 2007. Word reordering in statistical machine translation with a pos-based distortion model. In *Proceedings of TMI-11*, pages 171–180, Skvde, Sweden.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-08: HLT*, pages 577–585, Columbus, Ohio, June. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA'06*, pages 223–231, August.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 2, pages 901–904, Denver.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL'04*, pages 101–104.

Ashish Venugopal and Stephan Vogel. 2005. Considerations in maximum mutual information and minimum classification error training for statistical machine translation. In *Proceedings of EAMT-05*, Budapest, Hungary.

Stephan Vogel. 2003. SMT decoder dissected: Word reordering. In *Proceedings of NLP-KE'03*, pages 561–566, Bejing, China, Oct.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP'07*, pages 737–745.

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

Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of COLING'04*, pages 508–514, Geneva, Switzerland, August.

Peng Xu, Jaeho Kang, Michael Ringgaard, and Franz Och. 2009. Using a dependency parser to improve smt for subject-object-verb languages. In *Proceedings of NAACL-HLT'09*, pages 245–253, Boulder, Colorado, June. Association for Computational Linguistics.

Hirofumi Yamamoto, Hideo Okuma, and Eiichiro Sumita. 2008. Imposing constraints from the source tree on ITG constraints for SMT. In *Proceedings of ACL-08:HLT, SSST-2*, pages 1–9, Columbus, Ohio, June. Association for Computational Linguistics.

Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of COLING'04*, pages 205–211, Geneva, Switzerland, August.

Ying Zhang and Stephan Vogel. 2005. An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora. In *Proceedings of EAMT'05*, Budapest, Hungary, May. The European Association for Machine Translation.