# Which Words Matter in Defining Phrase Reorderings in Statistical Machine Translation?

**Hamidreza Ghader** h.ghader@uva.nl
**Christof Monz** c.monz@uva.nl
Informatics Institute, University of Amsterdam, The Netherlands

**Abstract**

Lexicalized and hierarchical reordering models use relative frequencies of fully lexicalized phrase pairs to learn phrase reordering distributions. This results in unreliable estimation for infrequent phrase pairs which also tend to be longer phrases. There are some smoothing techniques used to smooth the distributions in these models. But these techniques are unable to address the similarities between phrase pairs and their reordering distributions. We propose two models to use shorter sub-phrase pairs of an original phrase pair to smooth the phrase reordering distributions. In the first model we follow the classic idea of backing off to shorter histories commonly used in language model smoothing. In the second model, we use syntactic dependencies to identify the most relevant words in a phrase to back off to. We show how these models can be easily applied to existing lexicalized and hierarchical reordering models. Our models achieve improvements of up to 0.40 BLEU points in Chinese-English translation compared to a baseline which uses a regular lexicalized reordering model and a hierarchical reordering model. The results show that not all the words inside a phrase pair are equally important in defining phrase reordering behavior and shortening towards important words will decrease the sparsity problem for long phrase pairs.

## 1 Introduction

The introduction of lexicalized reordering models (LRMs) (Tillmann, 2004; Koehn et al., 2005) was a significant step towards better reordering by modeling the orientation of the current phrase pair with respect to the previously translated phrase. LRMs score the order in which phrases are translated by using a distribution of distinguished orientations conditioned on phrase pairs. Typically, the set of orientations consists of: monotone (M), swap (S) and discontinuous (D). However, LRMs are limited to reorderings of neighboring phrases only. Galley and Manning (2008) proposed a hierarchical phrase reordering model (HRM) for more global reorderings.

LRMs and HRMs both use relative frequencies observed in a parallel corpus to estimate the distribution of orientations conditioned on phrase pairs. As a result, they both suffer from the same problem of estimating reliable distributions for cases that occur rarely during training and therefore have to resort to smoothing methods to alleviate sparsity issues.

Cherry (2013) builds on top of HRMs and proposes a sparse feature approach which uses word clusters instead of fully lexicalized forms for infrequent words to decrease the effect of sparsity on the estimated model.

In this paper, we propose two types of approaches to use the most influential words from inside the original phrase pairs to estimate better orientation distributions for infrequent phrase pairs that takes phrase pair similarity more into account. In the first approach, we define a back-off model to shorten towards important words inside the original phrase pairs following the idea

| | | | | Dirichlet Smoothed | | | |
|---|---|---|---|---|---|---|---|
| | Source | Target | Freq | M | S | DL | DR |
| a | 中国 政府 | chinese government | 2834 | 0.216 | 0.034 | 0.315 | 0.433 |
| b | 日本 政府 | japanese government | 580 | 0.157 | 0.039 | 0.299 | 0.503 |
| c | 尼泊尔 政府 | nepalese government | 11 | 0.525 | 0.001 | 0.101 | 0.370 |

| | | | | Recursive Map Smoothed | | | |
|---|---|---|---|---|---|---|---|
| | Source | Target | Freq | M | S | DL | DR |
| a | 中国 政府 | chinese government | 2834 | 0.216 | 0.034 | 0.315 | 0.432 |
| b | 日本 政府 | japanese government | 580 | 0.158 | 0.039 | 0.300 | 0.501 |
| c | 尼泊尔 政府 | nepalese government | 11 | 0.400 | 0.009 | 0.202 | 0.388 |

Table 1: Examples of similar phrase pairs and their orientation probabilities using Dirichlet (Equation 1) and Recursive MAP (Equation 2) smoothing. M=monotone, S=swap, DL=discontinous left, and DR=discontinous right.

of back-off models in language model smoothing. This is, to some extent, complementary to the HRM in the sense of using smaller phrase pairs to make better prediction. The difference is that within HRMs smaller phrase pairs are merged into longer blocks when possible, while we propose to use shorter forms of phrase pairs when possible. In the second approach, we propose to produce generalized forms of original, fully lexicalized phrase pairs by including important words and marginalizing others allowing for smoothed distributions that better capture the true distributions of orientations. Here, we use syntactic dependencies from the original phrase pair to generalize and shorten in a more linguistically informed way.

The main contribution of this paper includes new methods to use shortened and generalized forms of a phrase pair to smooth the original phrase orientation distributions. We show that our smoothing approaches result in improvements in a phrase-based machine translation system, even when compared against a strong baseline using both LRM and HRM together. These methods do not require any changes to the decoder and do not lead to any additional computations during the decoding.

Our second contribution is a deeper analysis showing that orientation distributions conditioned on long phrase pairs typically depend on a few words within phrase pairs and not the whole lexicalized form. This supports and adds to the sparse reordering features (Cherry, 2013).

## 2 Problem Definition

In order to smooth the original maximum likelihood estimation, LRMs originally back off to the general distribution over orientations:

$$P(o \mid \bar{f}, \bar{e}) = \frac{C(o, \bar{f}, \bar{e}) + \sigma P(o)}{\sum_{o'} C(o', \bar{f}, \bar{e}) + \sigma} \quad (1)$$

which is also known as Dirichlet smoothing, where $\sigma P(o)$ denotes the parameters of the Dirichlet prior that maximizes the likelihood of the observed data, $C(o, \bar{f}, \bar{e})$ refers to the number of times a phrase pair cooccurs with orientation $o$, and $\sigma$ is the *equivalent sample size*, i.e., the number of samples required from $P(o)$ to reflect the observed data (Smucker and Allan, 2005). Cherry (2013) and Chen et al. (2013) introduce recursive MAP smoothing, which makes use of more specific priors by recursively backing off to orientation priors, see Equation 2. While recursive MAP smoothing factorizes phrase pairs into source and target phrases, it still considers the phrases themselves as fixed units.

$$
\begin{aligned}
P(o \mid \bar{f}, \bar{e}) &= \frac{C(o, \bar{e}, \bar{f}) + \alpha_s P_s(o \mid \bar{f}) + \alpha_t P_t(o \mid \bar{e})}{\sum_{o'} C(o', \bar{e}, \bar{f}) + \alpha_s + \alpha_t} \\
P_s(o \mid \bar{f}) &= \frac{\sum_{\bar{e}} C(o, \bar{f}, \bar{e}) + \alpha_g P_g(o)}{\sum_{o', \bar{e}} C(o', \bar{f}, \bar{e}) + \alpha_g} \\
P_t(o \mid \bar{e}) &= \frac{\sum_{\bar{f}} C(o, \bar{f}, \bar{e}) + \alpha_g P_g(o)}{\sum_{o', \bar{f}} C(o', \bar{f}, \bar{e}) + \alpha_g} \\
P_g(o) &= \frac{\sum_{\bar{f}, \bar{e}} C(o, \bar{f}, \bar{e}) + \alpha_u / 3}{\sum_{o', \bar{f}, \bar{e}} C(o', \bar{f}, \bar{e}) + \alpha_u}
\end{aligned}
\tag{2}
$$

To better understand what kind of information is ignored by both of the aforementioned smoothing methods, consider the phrase pairs and their corresponding distributions given in Table 1, for which we would expect similar distributions. The phrase pairs in rows (a) and (b) are frequently observed during training, resulting in reliable estimates. On the other hand, the phrase pair in row (c) is infrequent, leading to a very different distribution, due to the smoothing prior, while being semantically and syntactically close to (a) and (b). In Table 1, we can also observe that recursive MAP smoothing results in slightly more similar distributions compared to plain Dirichlet smoothing but the overall differences remain noticeable.

In this paper, we argue that in order to obtain smoother reordering distributions for phrase-pairs such as the ones in Table 1, one has to take phrase-internal information into account.

## 3 Related Work

The problem of data sparsity of training LRMs has first been addressed by Nagata et al. (2006) who propose to use POS tags and word clustering methods and distinguish the first or last word of a phrase, based on the language, as the head of a phrase.

Somewhat complementary to our work, Galley and Manning (2008) introduced hierarchical reordering models that group phrases occurring next to the current phrase into blocks, ignoring the internal derivation within a block, which biases orientations more towards monotone and swap. At the same time, orientations are still conditioned on entire phrase pairs, which means that their approach suffers from the same sparsity problems as LRMs. This problem has been more directly addressed by Cherry (2013) who uses unsupervised word classes for infrequent words of the phrase pairs in the form of sparse features. Like (Nagata et al., 2006), the first and last words of phrase pair are used as features in his model. Unfortunately, this approach also introduces thousands of additional sparse features, many of which have to be extracted *during* decoding, requiring changes to the decoder as well as a sizable tuning set.

Durrani et al. (2014) investigate the effect of generalized word forms on reordering in an n-gram-based operation sequence model, where they use different generalized representations including POS tags and unsupervised word classes to generalize reordering rules to similar cases with unobserved lexical operations.

While the approaches above use discrete representations, (Li et al., 2014) propose a discriminative model using continuous space representations of phrase pairs to address data sparsity problems. They train a neural network classifier based on recursive auto-encoders to generate vector space representations of phrase pairs and base reordering decision on those representations. They apply their model as an additional hypergraph reranking step since direct integration into the decoder would make hypothesis recombination more complex and substantially increase the size of the search space.

In addition to reordering models, several approaches have used word classes to improve other models within a statistical machine translation system, including translation (Wuebker

et al., 2013) and language models, where the problem of data sparsity is particularly exacerbated for morphologically rich target languages (Chahuneau et al., 2013; Bisazza and Monz, 2014).

## 4 Model Definition

In this section, we propose two different models which use different words as source of information to better estimate reordering distributions of sparse phrase pairs. Each model uses a different generalization scheme to obtain less sparse but still informative representations.

### 4.1 Interpolated Back-off Sub-phrases

In n-gram language modeling shorter n-grams have been used to smooth the probability distributions of higher order sparse n-grams. Lower order n-grams form the basis of Jelinek-Mercer, Katz, Witten-Bell and absolute discount smoothing methods (Chen and Goodman, 1999). For instance, Jelinek-Mercer smoothing linearly interpolates distributions of lower orders to smooth the distributions of higher order n-grams.

We use this as a motivation that shorter phrase pairs in lexicalized reordering models could play the role of lower-order n-grams in language model smoothing. But while backing off is obvious in language modeling, it is not straightforward in the context of lexicalized reordering models as there are several plausible ways to shorten a phrase pair, which is further complicated by the internal word alignments of the phrase pairs.

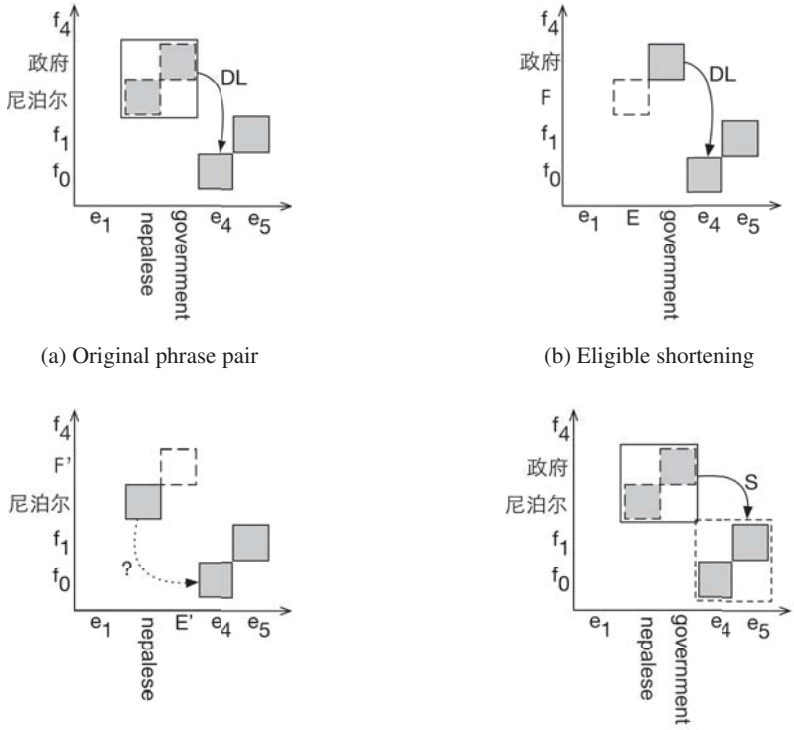


Figure 1: Backing off to shorter phrase pairs using eligible sub-phrase pairs (b). For comparison, we also include an example of a grouping of phrases as done in HRM (d).

The example in Figure 1 illustrates how sub-phrase pairs (Figure 1b and 1c) of the longer

phrase pair (Figure 1a) can be used to estimate the *discontinuous left* orientation for a longer, infrequent phrase pair. Following the strategy within language modeling to back off to shorter n-grams, we back off to the sub-phrase pairs that are consistent with the inside alignment of the longer phrase pair and provide a shorter and less sparse history. In this example, the number of times that sub-phrase pair (政府, government), see Figure 1b, appears with a *discontinuous left* jump of the length of the previous phrase pair for the next translation is considered when estimating the *discontinuous left* orientation for the longer phrase pair. On the other hand, the sub-phrase pair (尼泊尔, nepalese), see Figure 1c, cannot be used to predict a future *discontinuous left* of the long phrase pair, as there is no direct way to connect it to ($f_0$, $e_4$). The difference between this model and HRM can be seen in Figure 1d. HRM groups small phrase pairs from the context into longer blocks and determines the orientation with respect to grouped block, while our model looks into the phrase pair itself and uses possible shortenings to better estimate the orientation distribution conditioned on the original phrase pair. Our model can be applied to HRMs as well as LRMs to estimate a better distribution for long infrequent phrase pairs.

In order to provide a formal definition which sub-phrase pairs to consider when backing off, let us assume that $A$ is the set of alignment connections between the source $\bar{f}$ and target $\bar{e}$ side of a longer phrase pair. The set of eligible sub-phrase pairs, $E_{\bar{f},\bar{e}}$, is defined as follows:

$$E_{\bar{f},\bar{e}} = \{(\bar{f}^{[l,k]}, \bar{e}^{[l',n]}) | \; 1 \leqslant k \leqslant m, \; 0 \leqslant l \leqslant k, 0 \leqslant l' \leqslant n \text{ and } (\bar{f}^{[l,k]}, \bar{e}^{[l',n]}) \text{ consistent with } A \text{ if } l > 0 \wedge l' > 0 \; \} \tag{3}$$

where $\bar{f}^{[l,k]}$ is a sub-phrase of $\bar{f}$ with length $l$ which ends at the $k$th word of $\bar{f}$, $m$ and $n$ are the lengths of $\bar{f}$ and $\bar{e}$ respectively and the consistency with the alignment is ensured by the following three conditions (Koehn et al., 2005):

1. $\exists e_i \in \bar{e}^{[l',n]}, f_j \in \bar{f}^{[l,k]} : (i,j) \in A$
2. $\forall e_i \in \bar{e}^{[l',n]} : (i,j) \in A \Rightarrow f_j \in \bar{f}^{[l,k]}$
3. $\forall f_i \in \bar{f}^{[l,k]} : (i,j) \in A \Rightarrow e_i \in \bar{e}^{[l',n]}$

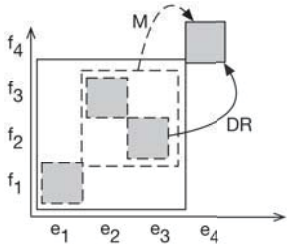


Figure 2: The distributions needed to estimate the conditional probability $\hat{p}(M \mid f_1 f_2 f_3, e_1 e_2 e_3)$ include $p(DR \mid f_2, e_3)$ and $p(M \mid f_2 f_3, e_2 e_3)$

Considering Figure 2, it is clear why ($\bar{f}^{[1,2]}$, $\bar{e}^{[1,3]}$) and ($\bar{f}^{[2,3]}$, $\bar{e}^{[2,3]}$) are considered eligible shortenings. Other possible shortenings such as ($\bar{f}^{[2,2]}$, $\bar{e}^{[3,3]}$) and ($\bar{f}^{[1,3]}$, $\bar{e}^{[1,2]}$) either violate the consistency conditions or do not run up to the end of the target side of the original phrase pair as in the definition above. Note that $n$ is a constant here and $\bar{e}^{[l',n]}$ means that all sub-phrase pairs must finish at the end of the target side of the original phrase pair. Otherwise one cannot directly determine the orientation with respect to the next phrase pair.

In our model, we compute the smoothed orientation distribution conditioned on a phrase pair by linearly interpolating the distribution of all eligible sub-phrases:

$$\hat{P}(o \mid \bar{f}, \bar{e}) = \sum_{(\bar{f}^{[l,k]}, \bar{e}^{[l',n]}) \in E_{\bar{f},\bar{e}}} \lambda_{l,l'} P(\Omega(\bar{f}^{[l,k]}, \bar{f}, o) \mid \bar{f}^{[l,k]}, \bar{e}^{[l',n]}) \quad (4)$$

where $E_{\bar{f},\bar{e}}$ is the set of eligible sub-phrase pairs and $\bar{f}^{[l,k]}$ indicates a sub-phrase of $\bar{f}$ with length $l$ ending at the $k$th word of $\bar{f}$, $\bar{e}^{[l',n]}$ is a sub phrase of $\bar{e}$ with the length of $l'$ which ends at the last word of $\bar{e}$ and the function $\Omega(\bar{f}^{[l,k]}, \bar{f}, o)$ returns the correct orientation considering the position of source sub-phrase $\bar{f}^{[l,k]}$ with respect to either end of the source phrase $\bar{f}$ and orientation $o$.

In order to compute the linear interpolation over the conditional distributions of the sub-phrase pairs, we have to determine the weight of each term in the linear interpolation (Equation 4). Here, we use expectation-maximization (EM) over a held-out data set, which is word-aligned using GIZA++ (Och and Ney, 2003). We extract phrase pairs using a common phrase extraction algorithm (Koehn et al., 2005) and count the number of occurrences of orientations for each phrase pair. These counts are used with unsmoothed reordering probabilities learned over the training data to compute the likelihood over the held-out data. We designed the EM algorithm to learn a set of lambda parameters for each length combination of the original phrase pairs. To reduce the number of parameters, we assume that all sub-phrases with the same length on source and target side share the same weight. This model is referred to as the BackOff model in the remainder of this paper.

### 4.2 Recursive Back-off MAP Smoothing

Above we used linear interpolation to estimate the final distribution from the orientation distributions of shorter sub-phrase pairs. Here we investigate another method aiming to affect the distributions of frequent phrase pairs to a lesser extent than those of non-frequent ones.

To this end, we use recursive MAP smoothing to estimate the distribution of the original phrase pair. In linear interpolation, all phrase pairs with the same length will get the same portion of their estimated distribution from their sub-phrases. On the other hand, for more frequent phrase pairs, the maximum likelihood distribution of the phrase pair itself is more reliable than the distributions of its sub-phrase pairs. Thus, a model relying more on the distribution of the original phrase pairs for frequent phrase pairs would be desirable.

To achieve this, we use a formulation similar to recursive MAP smoothing (Equation 2) with recursively backing off to the distributions of shorter sub-phrase pairs. At each recursion step we use the distribution of the longest sub-phrase pair as the prior distribution. Taking our definition for eligible sub-phrase pairs into account (Equation 3), all other sub-phrase pairs of the original phrase pair are sub-phrase pairs of the longest sub-phrase pair, in the case that the original phrase pair does not include unaligned words. For cases including unaligned words like the example in Figure 3, there could be sub-phrases where none of them is the sub-phrase of the other. In these cases we include the distributions of all those sub-phrases with the same *equivalent sample size* as the prior distributions. The estimated probability distribution of a phrase pair ($\bar{f}$,$\bar{e}$) is defined as follows:

$$\hat{P}(o \mid \bar{f}, \bar{e}) = \frac{C(o, \bar{f}, \bar{e}) + \sum_{(\bar{f}_L, \bar{e}_L) \in L_{\bar{f},\bar{e}}} \alpha \hat{P}(\Omega(\bar{f}_L, \bar{f}, o) \mid \bar{f}_L, \bar{e}_L)}{\sum_{o \in O} C(o, \bar{f}, \bar{e}) + \sum_{(\bar{f}_L, \bar{e}_L) \in L_{\bar{f},\bar{e}}} \alpha} \quad (5)$$

where $L_{\bar{f},\bar{e}}$ refers to the set of eligible sub-phrase pairs of $(\bar{f}, \bar{e})$ that are not sub-phrase pairs of each other and which is defined as follows:

$$L_{\bar{f},\bar{e}} = \{(\bar{f}', \bar{e}') \in E_{\bar{f},\bar{e}} \backslash \{(\bar{f}, \bar{e})\} \mid \neg \exists (\bar{f}'', \bar{e}'') \in E_{\bar{f},\bar{e}} \backslash \{(\bar{f}, \bar{e}), (\bar{f}', \bar{e}')\} : \bar{f}' \sqsubseteq \bar{f}'' \wedge \bar{e}' \sqsubseteq \bar{e}''\} \quad (6)$$

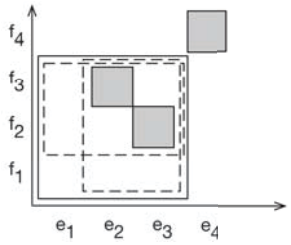


Figure 3: Illustration of sub-phrases where neither of the two pairs $(f_1 f_2 f_3, e_2 e_3)$ and $(f_2 f_3, e_1 e_2 e_3)$ of the original phrase pair $(f_1 f_2 f_3, e_1 e_2 e_3)$ is a sub-phrase of the other.

Here, $\bar{f}' \sqsubseteq \bar{f}''$ means that $\bar{f}'$ is a sub-phrase of or equal to $\bar{f}''$. As a result, $L_{\bar{f},\bar{e}}$ is the set of longest eligible sub-phrase pairs of original phrase pair where none of them is a sub-phrase of the others. For $E_{\bar{f},\bar{e}}$ see Equation 3. $O$ is the set of possible orientations. Note that we refer to this model shortly as RecursiveBackOff model from now on. We also refer to both this model and the BackOff model described in the previous section as back-off models.

### 4.3 Dependency Based Generalization

The methods described so far generalize the original phrase pairs by shortening towards the last aligned words as the most important words to define the reordering behavior of a phrase pair. In the remainder of this section, we use dependency parses to define how to generalize the original phrase pair and shorten towards important words.

Head-driven hierarchical phrase based translation (Li et al., 2012) suggests that using heads of phrases can be beneficial for better reordering in general. In our work, we define the heads of a phrase to be its *exposed heads*. Given a dependency parse, the exposed heads are all words inside a subsequence that are modifying a word outside it. Figure 4 shows an example of exposed heads in a phrase pair. The highlighted words are the exposed heads of the phrase pair. Exposed heads have been used in multiple linguistically motivated approaches as strong predictors of the next word in structured language models (Chelba and Jelinek, 2000; Garmash and Monz, 2015) and the next rule in a hierarchical translation system (Li et al., 2012).

In our model, besides training a regular lexicalized or hierarchical reordering model on surface forms of phrases, we train another reordering model which keeps the exposed heads lexicalized and replaces the remaining words in a phrase pair by a generalized representation. Assume that $RE$ is the set of dependency relations in the dependency parse tree of sentence $S$.

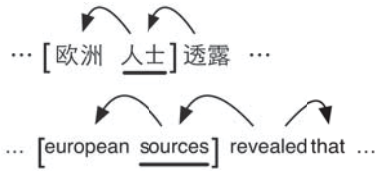


Figure 4: Examples of exposed heads in a Chinese-English phrase pair (between square brackets). The underlined words are the exposed heads since they have an incoming dependency originating outside of the phrase.

We consider each relation as an ordered pair $(w'_l, w_k)$ which means $w$ at index $k$ of sentence $S$ modifies $w'$ at index $l$. In addition, assuming $f_i^j$ is a phrase in $S$, starting from the $i$th and

ending with the $j$th word in sentence $S$, then the generalization $w_G$ of a word is:

$$w_G = \begin{cases} w_k & \text{if } w_k \in f_i^j, \exists (w'_l, w_k) \in RE : \\ & l < i \text{ or } l > j \text{ or } w' = ROOT \\ Gen(w_k) & \text{otherwise} \end{cases}$$

Here, $k$ and $l$ are indices of words $w$ and $w'$ in sentence $S$ and $ROOT$ is the root of the dependency parse of sentence $S$. The function $Gen(w)$ returns a generalization form for word $w$. We define this function in three different ways to create three different models.

1. $Gen(w_k)$ = POS_tag($w_k$)
2. $Gen(w_k)$ = <mod> if $w_{k-1}$ is not equal to <mod> and nothing otherwise
3. $Gen(w_k)$ remove $w_k$.

The question is how to use these generalizations to improve the estimation of the reordering distribution for each phrase pair. Our first model applies a generalization to the bilingual training data and creates a reordering model similar to the regular lexicalized reordering model, but based on relative frequency of generalized phrase pairs. In practice, a phrase pair may have multiple generalizations due to different dependency parses in different contexts. Since it is difficult to produce a dependency parse for the target side during decoding, we assume that a phrase pair will always have one possible generalization. Under this assumption, we can approximate the orientation distribution of a phrase pair to be:

$$\hat{P}(o \mid \bar{f}, \bar{e}) = P(o \mid \bar{f_G}, \bar{e_G}) \tag{7}$$

We can use the orientation distribution of a generalization as our estimate of the distribution of a phrase pair that produces the generalization. Here, $\bar{f_G}$ and $\bar{e_G}$ are word by word generalizations of $\bar{f}$ and $\bar{e}$. In case of multiple generalizations we use the one maximizing $P(\bar{f_G}, \bar{e_G} \mid \bar{f}, \bar{e})$. Depending on which of the three definitions for $Gen(w_k)$ we use, we name our models as PMLH (POS Modifiers Lexicalized Heads), MMLH (Merged Modifiers Lexicalized Heads), and LH (Lexicalized Heads) respectively.

As an alternative model, we propose to use the generalized distributions as a prior distribution in Dirichlet smoothing, where the distribution of each phrase pair is smoothed with the distribution of its generalized form. This should result in more accurate distributions since it affects the distributions of frequent phrase pairs to a lesser extent.

## 5 Experiments

We evaluate our models for Chinese-to-English translation. Our training data consists of the parallel and monolingual data released by NIST's OpenMT campaign, with MT04 used for tuning and news data from MT05 and MT06 for testing, see Table 2. Case-insensitive BLEU (Papineni et al., 2002) and translation error rate (TER) (Snover et al., 2006) are used as evaluation metrics.

### 5.1 Baseline

We use an in-house implementation of a phrase-based statistical machine translation system similar to Moses (Koehn et al., 2007), including the commonly used translation, lexical weighting, language, lexicalized reordering, and hierarchical reordering models. We use both lexicalized and hierarchical reordering models together, since this is the best model reported in (Galley and Manning, 2008) and our smoothing methods can be easily applied to the both models. Word alignments are produced using GIZA++ (Och and Ney, 2003), using grow-diag-final-and (Koehn et al., 2003). A 5-gram language model is trained on the English Gigaword corpus

| Corpus | Lines | Tokens(ch) | Tokens(en) |
|---|---|---|---|
| train | 937K | 22.3M | 25,9M |
| MT04 (dev) | 1,788 | 49.6K | 59.2K |
| MT05 (test) | 1,082 | 30.3K | 35.8K |
| MT06 (test) | 1,181 | 29.7K | 33.5K |

Table 2: Statistics for the Chinese-English bilingual corpora used in all experiments. Token counts for the English side of dev and test sets are averaged across all references.

with 1.6B tokens using interpolated, modified Kneser-Ney smoothing. The lexicalized and the hierarchical reordering models are trained with relative and smoothed frequencies using Dirichlet smoothing (Equation 1), for both left-to-right and right-to-left directions distinguishing four orientations: monotone (M), swap (S), discontinuous left (DL), and discontinuous right (DR). Feature weights are tuned using PRO (Hopkins and May, 2011) and statistical differences are computed using approximate randomization (Riezler and Maxwell, 2005).

In addition to the baseline, we reimplemented the 2POS model by Nagata et al. (2006), which uses the POS tag of the first and last words of a phrase pair to smooth the reordering distributions. The 2POS model is used in combination with the baseline lexicalized and hierarchical reordering models. Comparing our models to the 2POS model allows us to see whether backed-off sub-phrases and exposed heads of phrase pairs yield better performance than simply using the first and last words.

### 5.2 Comparison Systems and Results

We compare the baseline to the models described in Section 4. For all systems other than the baseline, the lexicalized and the hierarchical reordering models are replaced by the corresponding smoothed models. When computing the RecursiveBackOff model (Section 4.2), using Equation 5, we set the value of $\alpha$ to 10, following Cherry (2013) and Chen et al. (2013).

For the dependency-based model, we use the dependency parses of the source and the target side of the training corpus. The Stanford Neural-network dependency parser (Chen and Manning, 2014) is used to generate parses for both sides of the training corpus. From a dependency parse, we extract the smallest subtree that includes all incoming and outgoing relations of the words of a phrase. This is done for both the source and the target side phrases. Considering these subtrees, all words with an incoming connection from outside are exposed heads.

The experimental results for all models are shown in Table 3. As one can see, all our models achieve improvements in terms of BLEU on the test sets. The improvements for our back-off models are only significant for RecursiveBackOff over MT06 and MT05+MT06. The improvements over MT05 by our dependency-based shortenings are statistically significant for all models except PMLH. In the case of MT06, only the improvements resulting from MMLH are not statistically significant. However, both the PMLH and the MMLH model achieve the same improvements over MT05 and MT06 combined, and both are statistically significant. The LH model performs better than these models and also achieves higher improvement on the merged data. This model generalizes much more than the other models and is the only model that changes the distributions of single word phrases which are among the most frequently used phrase pairs. However, for frequent phrase pairs, being mapped to the same generalization form can be potentially harmful. In order to be able to control the effect of the model on the phrase pairs based on their frequency, we use the distributions in LH as a prior distribution with Dirichlet smoothing (Equation 1). This results in the LHSmoothed model shown in Table 3 which achieves the best improvements over both MT05 and MT06.

| Model | MT05 | | MT06 | | MT05 + MT06 | | |
|---|---|---|---|---|---|---|---|
| | BLEU↑ | TER↓ | BLEU↑ | TER↓ | BLEU↑ | TER↓ | RIBES↑ |
| Lex+Hrc | 32.25 | 60.13 | 33.00 | 57.17 | 32.84 | 58.62 | 79.24 |
| Nagata's 2POS | 32.20 | 60.31 | 33.13 | 57.07 | 32.87 | 58.66 | 79.11 |
| BackOff | *32.40* | 60.45 | *33.10* | *57.00* | *33.00* | 58.69 | 79.07 |
| RecursiveBackOff | *32.37* | 60.29 | *33.34*$^{\triangle,-}$ | *57.08* | *33.05*$^{\triangle,-}$ | 58.65 | *79.28* |
| PMLH | *32.41* | *60.08* | *33.26*$^{\triangle,-}$ | *57.05* | *33.04*$^{\triangle,-}$ | *58.53* | *79.26* |
| MMLH | *32.62*$^{\blacktriangle,\triangle}$ | *59.84*$^{-,\blacktriangle}$ | *33.18* | *56.91*$^{\triangle,-}$ | *33.04*$^{\triangle,-}$ | *58.35*$^{\blacktriangle,\triangle}$ | *79.43* |
| LH | *32.64*$^{\blacktriangle,\triangle}$ | *59.85*$^{\triangle,\blacktriangle}$ | *33.26*$^{\triangle,-}$ | *56.85*$^{\blacktriangle,-}$ | *33.11*$^{\blacktriangle,\triangle}$ | *58.32*$^{\blacktriangle,\blacktriangle}$ | *79.34* |
| LHSmoothed | **32.65**$^{\blacktriangle,\blacktriangle}$ | **59.80**$^{\triangle,\blacktriangle}$ | **33.38**$^{\triangle,-}$ | **56.77**$^{\blacktriangle,\triangle}$ | **33.20**$^{\blacktriangle,\triangle}$ | **58.25**$^{\blacktriangle,\blacktriangle}$ | *79.35* |

Table 3: Model comparison using BLEU, TER (lower is better), and RIBES over news data, which is combination of newswire and broadcast_news in case of MT06 and just newswire for MT05. Scores better than the baseline are in italics. ▲ and △ indicate statistically significant improvements at $p < 0.01$ and $p < 0.05$, respectively. The left hand side ▲ or △ is with respect to Lex+Hrc and the right hand side ones with respect to Nagata's 2POS. PMLH refers to the model using POS tags for modifiers and keeps exposed heads lexicalized. MMLH merges modifiers and keeps exposed heads lexicalized. LH removes modifiers. LHSmoothed uses the LH model with Dirichlet smoothing.

In addition to BLEU, we also report results using TER. Results for TER are in line with BLEU. BLEU and TER are general translation quality metrics, which are known to be not very sensitive to reordering changes (Birch et al., 2010). To this end we also include RIBES (Isozaki et al., 2010), a reordering-specific metric that is designed to directly address word-order differences between hypothesis and reference translation in translation tasks with long distant reordering language pairs and is highly sensitive to word-order mistakes.

### 5.3 Analysis

The improvements achieved by our BackOff and RecursiveBackOff methods show that these models capture some useful generalizations by shortening the phrase pairs towards the last aligned words in the target side as the most important words. The difference between the two models indicates that shortening is less beneficial for frequent phrase pairs and shorter phrase pairs which are less affected by lower-order distributions.

The improvements achieved by our generalization models support our hypothesis that not all words inside a phrase pair have the same impact on the reordering properties of the phrase pair as a whole. The experimental results for the PMLH model show that the lexicalized form of modifier words inside a phrase pair may just have the negative effect of increasing data sparsity.

Observing the improvements achieved by MMLH, we can go further and say that even the number of modifiers of an exposed head in a phrase does not influence reordering properties of a phrase pair. The improvements achieved by our LH model show not only that the number of modifiers but also the mere presence or absence of them does not significantly influence the reordering properties of a phrase pair.

One thing to bear in mind is that the PMLH and MMLH models do not change the distribution of single word phrase pairs which are mostly frequent phrase pairs, while the LH model does change these distributions as well. With the LHSmoothed model we have controlled this effect and decreased it to be negligible for frequent single word phrase pairs. However, it still changes the distributions of infrequent single word phrase pairs. Comparing the results of LH and LHSmoothed in Table 3, we suspect that the difference between the models for MT06 is

| Source Length | Target Length | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 8232 | 2719 | 879 | 269 | 89 | 18 | 7 |
| 2 | 2344 | 1777 | 1055 | 410 | 158 | 58 | 18 |
| 3 | 316 | 390 | 376 | 252 | 100 | 61 | 29 |
| 4 | 42 | 46 | 97 | 63 | 29 | 28 | 14 |
| 5 | 2 | 3 | 11 | 11 | 10 | 8 | 12 |
| 6 | 0 | 1 | 1 | 3 | 3 | 5 | 2 |
| 7 | 0 | 0 | 0 | 3 | 1 | 1 | 1 |

Table 4: Number of times that phrase pairs with different lengths and a frequency of less than 10 in the training data have been used during test on MT06 by the baseline. Phrase pairs occurring less than 10 times account for 72% of all phrases used during decoding of MT06.

| | |
|---|---|
| Source | ... 由于 东京 和 汉城 当局 均 寄望 能 于 二〇〇五年 底 前 签署 自由 贸易 协定 ... |
| Baseline | ... tokyo and seoul authorities are to be placed in 2005 before the end of the signing of a free trade agreement ... |
| LHSmoothed | ... tokyo and seoul authorities both in the hope of signing a free trade agreement before the end of 2005 ... |
| Ref | ... tokyo and seoul both hoped to sign a fta agreement by the end of 2005 ... |
| Source | 俄罗斯 多 次 指控 西方 插手 东欧 事务 ... |
| Baseline | russia has repeatedly accused of meddling in the affairs of the western and eastern europe ... |
| LHSmoothed | russia has repeatedly accused western intervention in the eastern european affairs ... |
| Ref | russia has been accusing the west of interfering in the affairs of eastern europe ... |

Table 5: Examples from MT05 illustrating reordering improvements over the baseline.

due to the effect of frequent single word phrase pairs.[1] However, comparing the results of LHSmoothed with other models we can say that even infrequent single word phrase pairs have benefited from the higher generalization level offered by this model. The statistics of infrequent phrase pairs used during testing and their lengths are shown in Table 4, giving an indication of why this model achieves the highest improvements. The table is showing that 41% of the infrequent phrase pairs (frequency $< 10$) used during translating MT06 have length of one in the both sides. So our models other than LH and LHSmoothed can not have neither positive nor negative effect on almost half of the infrequent phrase pairs. This also probably could explain why the back-off models achieve such a little improvements when 75% of infrequent phrase pairs have length $< 3$ in both sides.

In general, our dependency based models change the distributions to a larger extent than our back-off models, where not all long phrase pairs have eligible sub-phrase pairs, while the dependency models more often result in shorter generalizations. Table 5 provides some examples where our LHSmoothed model has improved the translations by better modeling of reorderings. To further the understanding of how the reordering distributions of infrequent phrase pairs used in the examples in Table 5 are affected by our models, we show some examples in Table 6, before and after applying our model. As a result of our model, the phrase pair (寄望, in the hope)

[1] We also used the distributions from PMLH and MMLH as priors in Dirichlet smoothing, but it did not lead to any noticeable changes in the results.

| 寄望 | in the hope | M | S | DL | DR |
|---|---|---|---|---|---|
| Monotone with previous | Baseline | 0.10 | 0.01 | 0.11 | 0.78 |
| | LHSmoothed | 0.28 | 0.01 | 0.01 | 0.70 |
| | Counts in training | 0 | 0 | 0 | 1 |
| 能 于 | of | | | | |
| Discontinuous right with next | Baseline | 0.10 | 0.01 | 0.12 | 0.77 |
| | LHSmoothed | 0.06 | 0.01 | 0.07 | 0.87 |
| | Counts in training | 0 | 0 | 0 | 1 |
| 签署 自由 贸易 协定 | signing a free trade agreement | | | | |
| Discontinuous right with previous | Baseline | 0.10 | 0.02 | 0.68 | 0.20 |
| | LHSmoothed | 0.21 | 0.03 | 0.30 | 0.46 |
| | Counts in training | 0 | 0 | 1 | 0 |

Table 6: Orientation distributions shift of some infrequent phrase pairs that has been used with the correct orientation to produce our translations using LHSmoothed model in Table 5

receives an increase in monotone orientation probability although it has frequency of zero for this orientation. The next phrase pair (能 于, of) receives an increase in discontinuous right probability resulting in the correct usage of this orientation during translation. The most interesting case is the substantial decrease in the probability of discontinuous left and the increase in discontinuous right for the phrase pair (签署 自由 贸易 协定, signing a free trade agreement), even though it has frequency 1 for the former orientation and 0 for the latter. These shifts within the probability distributions lead to the better translation generated by LHSmoothed model for the first example in Table 5.

## 6 Conclusions

We have introduced a novel method that builds on the established idea of backing off to shorter histories, commonly used in language model smoothing, and shown that it can be successfully applied to smoothing of lexicalized and hierarchical reordering models in statistical machine translation. Furthermore, we have shown that not all sub-phrase pairs are influential in that regard. The sub-phrase pair consisting of just exposed heads of a phrase pair tends to be the most important one and most other words inside a phrase pair have negligible influence on reordering behavior. Earlier approaches, such as (Nagata et al., 2006) and (Cherry, 2013), often assume that the last and the first word of a phrase pair are important, but our experiments indicate that exHGposed heads tend to be stronger predictors. We showed that generalized representations of phrase pairs based on exposed heads can help decrease sparsity and result in more reliable reordering distributions.

Considering the analysis of the length of infrequent phrase pairs used during translation, we also conclude that a smoothing model that would be able to further improve the distribution of single word phrase pairs is crucial for achieving higher improvements during translation.

For future work, we plan to investigate the effect surface word forms from outside of a phrase pair can have on reordering. This could further help improve reordering distributions of infrequent single word phrase pairs more effectively as they constitute a large portion of the phrases used during decoding.

**Acknowledgments**

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers 639.022.213 and 612.001.218.

## References


Birch, A., Osborne, M., and Blunsom, P. (2010). Metrics for MT evaluation: evaluating reordering. *Machine Translation*, 24(1):15–26.

Bisazza, A. and Monz, C. (2014). Class-based language modeling for translating into morphologically rich languages. In *Proceedings of the 25th Annual Conference on Computational Linguistics (COLING)*, pages 1918–1927.

Chahuneau, V., Schlinger, E., Smith, N. A., and Dyer, C. (2013). Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1687.

Chelba, C. and Jelinek, F. (2000). Structured language modeling. *Computer Speech & Language*, 14(4):283–332.

Chen, B., Foster, G., and Kuhn, R. (2013). Adaptation of reordering models for statistical machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 938–946.

Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 740–750.

Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Cherry, C. (2013). Improved reordering for phrase-based translation using sparse features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 22–31.

Durrani, N., Koehn, P., Schmid, H., and Fraser, A. (2014). Investigating the usefulness of generalized word representations in SMT. In *Proceedings of the 25th Annual Conference on Computational Linguistics (COLING)*, pages 421–432.

Galley, M. and Manning, C. D. (2008). A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856.

Garmash, E. and Monz, C. (2015). Bilingual structured language models for statistical machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2398–2408.

Hopkins, M. and May, J. (2011). Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (2010). Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952.

Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of International Workshop on Spoken Language Translation*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics*, pages 177–180.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 48–54.

Li, J., Tu, Z., Zhou, G., and van Genabith, J. (2012). Head-driven hierarchical phrase-based translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 33–37.

Li, P., Liu, Y., Sun, M., Izuha, T., and Zhang, D. (2014). A neural reordering model for phrase-based translation. In *Proceedings of the 25th Annual Conference on Computational Linguistics (COLING)*, pages 1897–1907.

Nagata, M., Saito, K., Yamamoto, K., and Ohashi, K. (2006). A clustered global phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 713–720.

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Riezler, S. and Maxwell, J. T. (2005). On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64.

Smucker, M. D. and Allan, J. (2005). An investigation of dirichlet prior smoothing's performance advantage. Technical Report IR-391, The University of Massachusetts, The Center for Intelligent Information Retrieval.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

Tillmann, C. (2004). A unigram orientation model for statistical machine translation. In *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 101–104.

Wuebker, J., Peitz, S., Rietig, F., and Ney, H. (2013). Improving statistical machine translation with word class models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1377–1381.