Discriminative Phrase-based Lexicalized Reordering Models using Weighted Reordering Graphs

Wang Ling, João Graça, David Martins de Matos, Isabel Trancoso, Alan Black
L^{2}F Spoken Systems Lab, INESC-ID, Lisboa, Portugal
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
{wang.ling,joao.graca,isabel.trancoso,david.matos}@inesc-id.pt
awb@cs.cmu.edu

Abstract

Lexicalized reordering models play a central role in phrase-based statistical machine translation systems. Starting from the distance-based reordering model, improvements have been made by considering adjacent words in word-based models, adjacent phrases pairs in phrase-based models, and finally, all phrases pairs in a sentence pair in the reordering graphs. However, reordering graphs treat all phrase pairs equally and fail to weight the relationships between phrase pairs. In this work, we propose an extension to the reordering models, named weighted reordering models, that allows discriminative behavior to be defined in the estimation of the reordering model orientations. We apply our extension using the weighted alignment matrices to weight phrase pairs, based on the consistency of their alignments, and define a distance metric to weight relationships between phrase pairs, based on their distance in the sentence. Experiments on the IWSLT 2010 evaluation dataset for for the Chinese-English language pair yields an improvement of 0.38 (2%) and 0.94 (3.7%) BLEU points over the state-of-the-art work's results using weighted alignment matrices.

1 Introduction

Reordering in Machine Translation (MT) is the task of word-order redistribution of translated words. An early reordering paradigm uses a simple distance based reordering model, which penalizes words that diverge from their original position after being translated (Koehn et al., 2003). This works moderately well for language pairs where reordering distances are small, but performs poorly for language pairs such as Chinese-English, where the opposite occurs. One of many approaches to implement improved reordering models is to use the lexical information during the phrase extraction algorithm to predict reordering orientations, using word-aligned bilingual sentences. However, the fact that spurious word alignments might occur leads to the use of alternative representations for word alignments that allow multiple alignment hypotheses, rather than the 1-best alignment (Venugopal et al., 2009; Mi et al., 2008; Christopher Dyer et al., 2008). More recently, a more efficient representation of multiple alignments was proposed in (Liu et al., 2009) named weighted alignment matrices, which represents the alignment probability distribution over the words of each parallel sentence. The method for building a word-based lexicalized reordering model using these matrices is proposed in (Ling et al., 2011). However, phrase-based reordering models have been shown to perform better than word-based models for several language pairs (Tillmann, 2004; Su et al., 2010; Galley and Manning, 2008), such as Chinese-English and Arabic-English.

In this work, we propose an extension to the phrase-based lexicalized model approach using reordering graphs presented in (Su et al., 2010), which allows phrase pairs to be weighted differently, rather than uniformly as in the original proposal. Then, we will present a phrase-based approach to estimate the orientations of the reordering model from the weighted alignment matrices using this extension.

2 Lexicalized Reordering models

In this section we will present the lexicalized reordering models approaches that are relevant for this work.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 47–55,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

2.1 Word-based Reordering

The lexicalized reordering model is possibly the most used lexicalized reordering model and it calculates, as features, the reordering orientation for the previous and the next word, for each phrase pair. In the word-based reordering model (Axelrod et al., 2005), during the phrase extraction, given a source sentence S and a target sentence T, the alignment set A, where a_i^j is an alignment from i to j, the phrase pair with words in positions between i and j in S, S_i^j, and n and m in T, T_n^m, can be classified with one of three orientations with respect to the previous word. The orientation is

- Monotonous - if only the previous word in the source is aligned with the previous word in the target, or, more formally, if $a_{i-1}^{n-1} \in A \wedge a_{j+1}^{n-1} \notin A$.
- Swap - if only the next word in the source is aligned with the previous word in the target, or more formally, if $a_{j+1}^{n-1} \in A \wedge a_{i-1}^{n-1} \notin A$.
- Discontinuous - if neither of the above are true, which means, $(a_{i-1}^{n-1} \in A \wedge a_{j+1}^{n-1} \in A) \vee (a_{i-1}^{n-1} \notin A \wedge a_{j+1}^{n-1} \notin A)$.

The orientations with respect to the next word are given analogously. The reordering model is generated by grouping the phrase pairs that are equal, and calculating the probabilities of the grouped phrase pair being associated each orientation type and direction, based on the orientations for each direction that are extracted. Formally, the probability of the phrase pair p having a monotonous orientation is given by:

$$P(p, mono) = \frac{C(p,mono)}{C(p,mono)+C(p,swap)+C(p,disc)} \quad (1)$$

Where $C(p, o)$ is the number of times a phrase is extracted with the orientation o in that group of phrase pairs.

2.2 Word-based Reordering using alignment matrices

The work in (Ling et al., 2011) adapts the word-based reordering model to extract the reordering orientations from the weighted alignment matrices. This is done by changing the $C(p, o)$ from a count function over a given set of phrase pairs P to a weighted sum given by:

$$C(p, o) = \sum_{p \in P} Sc(p) P_c(p, o) \quad (2)$$

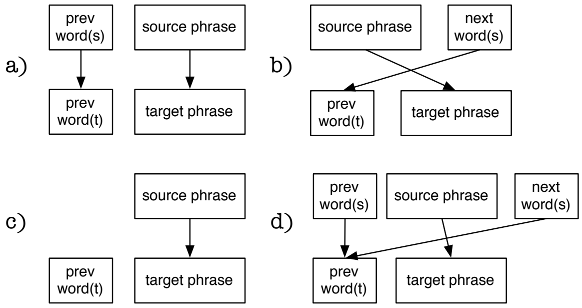

Figure 1: Enumeration of possible reordering cases with respect to the previous word. Case a) is classified as monotonous, case b) is classified as swap and cases c) and d) are classified as discontinuous.

The score $Sc(p)$ of a phrase pair p is given by the algorithm described in (Liu et al., 2009), which is based on its alignments. This score is higher if the alignment points in the phrase pair have high probabilities, and if the alignment is consistent. Thus, if a phrase pair has better quality, its orientation is given more weight than phrase pairs with worse quality. Rather than classifying each phrase pair with either monotonous (M), swap (S) or discontinuous (D), a probability distribution for the orientations is calculated. Thus, for the previous word, given a weighted alignment matrix W, the phrase pair between the indexes i and j in S, S_i^j, and n and m in T, T_n^m, the probability values for each orientation are given by:

- $P_c(p, M) = W_{i-1}^{n-1}(1 - W_{j+1}^{n-1})$
- $P_c(p, S) = W_{j+1}^{n-1}(1 - W_{i-1}^{n-1})$
- $P_c(p, D) = W_{i-1}^{n-1} W_{j+1}^{n-1}$ $+ (1 - W_{i-1}^{n-1})(1 - W_{j+1}^{n-1})$

2.3 Phrase-based Reordering

The problem with the word-based lexicalized reordering model is that it is assumed that the adjacent words are translated by themselves, which is not always true in phrase-based SMT. The reordering model presented in (Tillmann, 2004) uses adjacent phrases to generate the phrase orientations. In this model, the previous orientation of a phrase pair p:

- Monotonous if there is a phrase pair with the source S_x^{i-1} that ends at $i-1$ and starts at any x, that is aligned to the target phrase T_y^{n-1}, that ends at $n-1$ and starts at any y. In another words, if there is an adjacent phrase pair that occurs before p, both in the source and in the target sentences.

- Swap if there is a phrase pair with the source S_{j+1}^{x} that starts at $j+1$ and ends at any x, that is aligned to the target phrase T_{y}^{n-1}, that ends at $n-1$ and starts at any y. In another words, if there is an adjacent phrase pair that occurs before p in the target sentence and after p in the source sentence.
- Discontinuous - if neither of the above are true.

The work presented in (Tillmann, 2004) only considers phrases that are smaller than a fixed size, since the possible phrases for each bilingual sentence are generated and kept in memory making the time and memory needed to store and lookup all possible phrases grows rapidly as the size grows. The work in (Galley and Manning, 2008) implements a shift-reduce parsing algorithm that updates the previously extracted phrase pair orientations when a new phrase pair is extracted, which allows arbitrary sized phrase pairs to be considered.

2.4 Phrase Reordering using Reordering Graphs

The phrase based model considers the existence of a single adjacent phrase, which is not ideal, since many possible adjacent phrases exist for each extracted phrase pair, which can generate different orientations. In work done in (Su et al., 2010) the orientation is computed by considering all possible reorderings of phrase pairs that are extracted in the sentence pair. Once again $C(p,o)$ is given by the weighted count:

$$C(p,o) = \sum_{p \in Ph} P_c(p,o) \tag{3}$$

Where $P_c(p,o)$ is extracted by structuring the extracted phrase pairs into a reordering graph. This graph is created for each sentence pair, using extractable phrase pairs as nodes and connecting adjacent phrase pairs in the target side with an edge. Each edge is associated with a reordering orientation, dependent on the source side of the connected phrase pairs. If these are not adjacent in the source side, the edge is given a discontinuous orientation, otherwise, the edge is given a monotonous or swap orientation depending on whether they are in the same order or not in the source side, respectively.

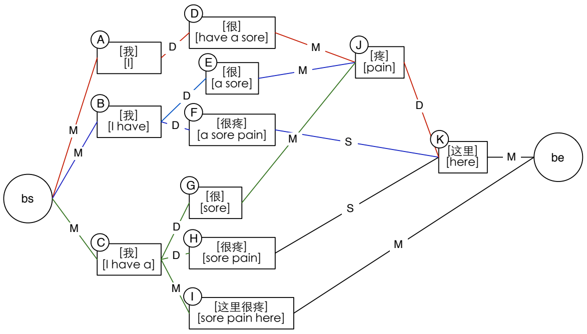

Figure 2: Reordering Graph for the sentence pair with source “我这里很疼 (wo zhe li hen teng)” and target ‘I have a sore pain here”. Each extracted phrase pair is denoted as a rectangle, with the source and target phrases inside. Each phrase pair is also labeled for reference. The edges are labeled with its orientation (the colors are only for easier visualization).

Furthermore, phrase pairs with no adjacent phrase pairs are linked to the nearest phrase pair. These arcs are also given a discontinuous orientation.

A start node bs is added and is linked to all phrase pairs at the start of the target sentence , and a end node be, which is linked by all phrase pairs that end the target sentence.

For the previous orientation weight, the probability $P_c(p,o)$ of the phrase pair p having a given orientation o, considering the set $Prev(o)$, with all phrase pairs that are linked to p that would lead to a orientation o, is given by:

$$P_c(p,o) = \sum_{p' \in Prev(o)} \frac{\alpha(p')\beta(p)}{\beta(bs)} \tag{4}$$

In this equation, $\alpha(p')$ is the number of paths to the p' node from the first phrase, $\beta(p)$ is the number of paths from p to the last phrase and $\beta(bs)$ results in the number of possible paths. We can see from this equation that $\alpha(p') \times \beta(p)$ results in the number of paths containing the arc from p' to p, thus not only multiple adjacent phrase pairs are considered, but these are also weighted by the number of possible translation segmentations that would use that phrase pair.

An example of a reordering graph is illustrated in figure 2. We can see in that for the phrase pair “这里 (zhe li)”→“here”, instead of simply giving the swap orientation due to the adjacent phrase pair “很疼 (hen teng)”→“sore pain”, it also considers the case where “疼 (teng)” is translated by itself to “pain”, in which case the orientation would be discontinuous. In this case, the weight is evenly distributed between the 2 orientations since there are 2 paths that contain an edge with each orientation.

3 Phrase-based reordering using Weighted Reordering Graphs

We see in the example given in section 2.4 that the phrase pair "这里 (zhe li)"→"here" is given an equal weight to the swap and discontinuous orientations. However, if we translate "疼 (teng)" by itself, we would have to translate "很 (hen)" without "疼 (teng)". The translation for "很 (hen)" by itself to "sore" is not very probable, since "很 (hen)" without context is generally translated to "very", "much" or "quite". Thus, it is more probable during decoding that the segmentation "很疼 (hen teng)" is used. Although the phrase-based reordering model presented in section 2.3 gives a better reordering estimate in this case, in cases where there is an equal probability of both segmentations the graph-based approach would be better. Hence, we argument that by treating phrase pairs discriminatively, we can improve reordering orientations estimate in both cases. In this work, we propose an extension to the reordering graphs to allow the definition of discriminative behavior during training. We will start by describing our model for the Weighted Reordering Graph and then we proceed into the definition of the algorithm to extract the word orientations from the Weighted Reordering Matrices.

3.1 Weighted Reordering Graph Model

We define a weighted reordering graph for a given sentence pair S as $G_S = (V, E, W_v, W_e)$, where V is the set of all vertices, which are phrase pairs $p_1, p_2, ..., p_n$, and E is the set of all edges and we denote a edge from p_1 to p_2 as $e(p_1, p_2)$. W_v and W_e are functions that map each element in V and E to a weight.

We define a path $PH(bs, p_1, p_2, be)$ as a path starting in bs and through p_1, then p_2 and ending in be. The weight of a path $PH(p_1, p_2, ..., p_{k-1}, p_k)$ is given by the product of the weights of its phrase pairs $W_v(p_1)W_v(p_2)...W_v(p_{k-1}), W_v(p_k)$ and the weights of the edges connecting the phrases in the path $e(p_1, p_2)...e(p_{k-1}, p_k)$. If both weight functions W_v and W_e are set to return 1 we define the same behavior as a reordering graph described in section 2.4, since all paths will have the weight of 1.

Following equation 4, we define the probability of a given orientation, $P_c(p, o)$, for a weighted reordering graph as:

$$P_c(p, o) = \sum_{p' \in Prev(o)} \frac{\alpha_w(p')W_e(e(p', p))\beta_w(p)}{\beta_w(bs)} \quad (5)$$

Where $\alpha_w(p')$ is the sum of weights of paths from bs to p' and $\beta_w(p)$ is the sum of the weights of the paths from p to be. It is also crucial to consider the weight of the edge $e(p', p)$, since it is not weighted in neither $\alpha_w(p')$ nor $\beta_w(p)$. This is not present in equation 4 since all edges are weighted as 1.

The functions α_w and β_w can be defined as:

$$\alpha_w(p) = W_v(p) \sum_{p' \in Prev(p)} \alpha_w(p')W_e(e(p', p))) \quad (6)$$

$$\beta_w(p) = W_v(p) \sum_{p' \in Next(p)} \beta_w(p')W_e(e(p', p))) \quad (7)$$

Where $Prev(p)$ and $Next(p)$ are sets of all phrase pairs that are linked to and linked from p, respectively. We also initiate $\alpha_w(bs) = 1$ and $\beta_w(be) = 1$. These two values can be initialized with any value, as this will not affect the normalized result from equation 5.

The pre-computation of α_w and β_w can be performed using an approach similar to the forward-backward algorithm and calculating the forward probabilities for α_w and backward probabilities for β_w, with time complexity in the order $O(N^2T)$, where N is the number of different phrase pairs and T is the length of sequences. To compute α_w, we need to take into account that the vertices/phrase pairs can be ordered topologically in an array so that a vertice in the index i does not have edges pointing to any vertice at any index at or before $i - 1$. This can be done by ordering the phrase pairs by the ending position of the target phrase n, starting with bs, since we know that the phrase pairs ending at n can only have links to phrase pairs ending at least at $n + 1$, according to the definition of an edge in a reordering graph.

The algorithm for computing β_w can be done analogously, by starting from be and sorting according to the target phrase's starting position.

3.2 Choosing W_v and W_e

In this work, we use the information given by the Weighted Alignment Matrices to define $W_v(p)$. We set $W_v(p)$ using the phrase pair scoring algorithm presented in (Liu et al., 2009), which calculates the weight of a phrase pair based on its alignment points. This weight can be seen as the

Algorithm 1 Compute α_w

```
Require: sorted phrase pairs $P$
  $\alpha_w(P) = \{0, ..., 0\}$
  for $p$ in $P$ do
    if $p = bs$ then
      $\alpha_w(bs) := W_v(bs)$
    end if
    for $p'$ in followingNodes($p$) do
      $w = W_e(e(p, p'))W_v(p')$
      $\alpha_w(p') := \alpha_w(p') + w\alpha_w(p)$
    end for
  end for
  return extractedPhrasePairs
```

probability of that phrase pair being extractable according to the heuristic proposed in (Koehn et al., 2003).

We also set $W_e(e(p^a, p^b))$ to a distance function between the phrases p^a and p^b in the target, to better cope with the situation where there are no adjacent phrase pairs to a phrase pair and an edge is added to the closest adjacent phrase. An example of such a case instance is displayed in figure 3. In this case, the phrase ending with the target word "wrap" can not be extracted, since it is aligned with "把 (ba)" and "包 (bao)". Thus, consistent phrase pairs that can be extracted must at least contain these 2 words to have "wrap" in the target. Another situation where this might also happen is the case where a phrase is not extracted due to size constraints. In both cases the reordering orientation of the edge is always discontinuous since the target phrases are not adjacent. We find that the reordering orientations that originates from these edges are not very precise. In the first case, the next orientation for the phrase pair "不需要 (bu xu yao)"→"need not" would be divided between monotonous from the edge to the phrase pair "把它包 (ba ta bao)"→"wrap it" and discontinuous from the edge to the phrase pair from "它 (ta)" to "it". However, in an actual translation the only way to translate this sentence correctly would be to treat "把它包 (ba ta bao)" as a segment, which favors the monotonous orientation. Furthermore, even if we could translate "把 (ba)" and "包 (bao)" to "wrap", the orientation would still be monotonous. In the second case, an edge between phrase pairs p^a and p^b that is created because a phrase pair p^c between p^a and p^b was not extracted due to size constraints, would mean that there is a missing edge from p^a to p^c and another from p^b to p^c, and no edge from p^a to p^b. Thus, the orientation of this edge is likely to be spurious, and should be given a lower weight.

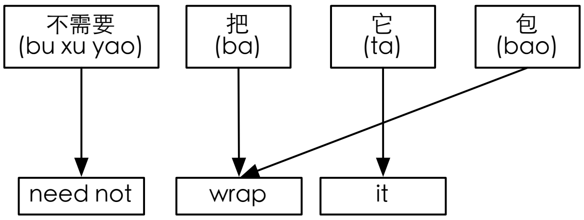

Figure 3: Illustration of a case where no adjacent phrase pairs exits for a phrase pair. In this case, the phrase pair "不需要 (bu xu yao)"→"need not" and the phrase pair "它 (ta)"→"it" are linked by an edge even though they are not adjacent.

This cannot be done by setting this weight to 0 (removing the arc), since it will render all paths that contain that edge to 0. For instance, if we set the maximum phrase pair size to 7 and the first 8 words of the source and target are aligned in a way that no smaller translation units can be extracted, there would be an edge from bs to the 9th word. Therefore, if we set the weight of the edge to 0, any path we take would have 0 weight, rendering the whole sentence pair useless. Hence, we define the W_e function as:

$$W_e(e(p, p')) = \frac{1}{(1 + \lambda)^d} \tag{8}$$

Where d is the distance between p and p' defined by the number of words in the target sentence between p and p', and λ is a positive value defining the penalty as d increases. In this work, we empirically set $\lambda = 0.5$, and leave the optimization of this parameter as future work.

3.3 Reordering Model Comparison

In order to illustrate the performance of the different reordering models, we consider two training sentences taken from the IWSLT 2010 DIALOG task. The weighted alignment matrices for these sentences are illustrated in tables 1 and 2. For simplicity in terms of illustration, we assume that algorithms that do not use the alignment matrices, consider all non-zero cells as alignment points. The probability distribution for different previous orientations of each reordering model for the phrase pair "这里 (zhe li)→here" from sentence 1 and the phrase pair "今天 (jin tian)→today" from sentence 2 are calculated in tables 3 and 4, respectively.

Sentence 1	I	have	a	sore	pain	here
我 (wo)	0.90					
这里 (zhe li)						0.75
很 (hen)				0.50		
疼 (teng)					0.80	

Table 1: Weighted alignment matrix for a training sentence pair from DIALOG training corpus from IWSLT 2010.

Sentence 2	Are	there	any	baseball	games	today
今天 (jin tian)						1
有 (you)		0.60	0.90			
棒球 (bang qiu)				1		
比赛 (bi sai)					0.65	
吗 (ma)						

Table 2: Weighted alignment matrix for a training sentence pair from DIALOG training corpus from IWSLT 2010.

We can see that the word-based reordering models classify the word orientation as discontinuous, since the previous word in the target is not aligned to adjacent words in the source. This leads to inaccurate orientations for the first phrase pair, since the words "很 (hen)" and "疼 (teng)" have a high probability of being translated together. In the second phrase pair, it gives a good approximation of the correct orientation, since "棒球 (bang qiu)" and "比赛 (bi sai)" are good translations even when translated without "有 (you)".

The opposite occurs with the phrase-based reordering model, since it considers the source phrase segmentation "很疼 (hen teng)" and "有棒球比赛 (you bang qiu bi sai)", respectively. Thus, the estimation of the orientation is better for the former and worse for the latter phrase pair.

这里 (zhe li)→here	Mono	Swap	Disc
Word-based	0	0	1
Weighted-Word-based	0	0	1
Phrase-based	0	1	0
Graph-based	0	0.333	0.333
Weighted-graph-based	0	0.271	0.180

Table 3: Previous orientation probabilities for different lexicalized reordering models for the phrase pair "这里 (zhe li)"→"here", taken from sentence 1.

今天 (jin tian)→today	Mono	Swap	Disc
Word-based	0	0	1
Weighted-Word-based	0	0	1
Phrase-based	0	1	0
Graph-based	0	0.166	0.500
Weighted-graph-based	0	0.187	0.416

Table 4: Previous orientation probabilities for different lexicalized reordering models for the phrase pair "今天 (jin tian)"→"today", taken from sentence 2.

Using the reordering graph, orientations are estimated for different adjacent phrase pairs. In the first phrase pair, both cases where the source phrase "很疼 (hen teng)" and "疼 (teng)" are used as translation units are taken into account, and the same happens with the second phrase pair. As it was already referred in section 3, the problem with this estimation is that it fails to consider that "疼 (teng)" is more likely to be translated with "很 (hen)", otherwise the translation of "很 (hen)" is less likely to be accurate.

Finally, by using weighted-reordering-graph, using phrase scores calculated from weighted alignment matrices, paths in the graph that contain phrase pairs that are better aligned are given more weight.

4 Experiments

We implemented both word-based and phrase-based lexicalized reordering models described above, and compared the translation results with our algorithm.

4.1 Corpus

Our experiments were performed over two datasets, the BTEC and the DIALOG parallel corpora from the latest IWSLT evaluation in 2010 (Paul et al., 2010). The experiments performed with the BTEC corpus used the French-English subset, while the ones perfomed with the DIALOG corpus used the Chinese-English subset. The training corpora contains about 19K and 30K sentences, respectively.

The development corpus for the BTEC task was the CSTAR03 test set composed by 506 sentences, and the test set was the IWSLT04 test set composed by 500 sentences and 16 references. As for the DIALOG task, we performed 2 tests, one using the evaluation datasets from IWSLT evaluation in 2006 (IWSLT06) and in 2008 (IWSLT08). The development from the IWSLT06 evaluation is composed by 489 sentences, and the test set was composed by 500 sentences and 7 references. The development from IWSLT08 evaluation is composed by 245 sentences, and the test set was composed by 506 sentences and 7 references.

4.2 Setup

We use two setups for the different corpora that are used. Word-based (word-based), phrase-based (phrase-based), and phrase-based using reordering graphs (graph-based) reordering models are generally used with the regular phrase extraction algorithm, with alignment constraints defined in (Koehn et al., 2003). Thus, we compare our method with the methods above in this environment. The weighted word-based reordering model (weighted word-based) described in section 2.2 was tested using the phrase extraction algorithm described in (Liu et al., 2009), where phrase pairs are filtered out based on their scores, which are calculated from their alignment probabilities. So, we also test our algorithm under these conditions. In our work, we set the threshold for filtering out phrase pairs to 0.1, which was the threshold used in (Ling et al., 2011). We also test separately our implementation of W_v (graph-based W_v), which weights phrase pairs according to their scores, and W_e (graph-based W_e), a penalty based on the distance between phrase pairs. Then, we test both approaches combined (graph-based W_vW_e).

The word alignments and the weighted alignment matrices are generated using the Geppetto toolkit[1], using a regular HMM-based word alignment model (V. Graça et al., 2010), without restraints. The optimum alignment is found using posterior decoding, and the weighted alignment matrices are obtained from the same alignment posteriors.

The optimization of the translation model weights was done using MERT tuning. Each experiment was run three times, and the final scores are calculated as the average of the three runs in order to stabilize the results. The results were evaluated using BLEU-4 and METEOR, and computed with 16 references.

4.3 Results

Tables 5, 6 and 7 show the scores using the different reordering models. Consistent improvements

BTEC	BLEU	METEOR
regular phrase extraction		
word-based	**57.97**	**63.82**
phrase-based	57.56	63.57
graph-based	57.53	63.63
graph-based W_v	57.30	63.42
graph-based W_e	57.47	63.49
graph-based W_vW_e	57.66	63.63
weighted phrase extraction		
weighted word-based	**62.01**	**66.31**
graph-based W_v	61.10	65.75
graph-based W_vW_e	61.75	66.19

Table 5: Results for the BTEC task.

IWSLT06 DIALOG	BLEU	METEOR
regular phrase extraction		
word-based	14.88	36.61
phrase-based	15.32	36.90
graph-based	15.28	37.14
graph-based W_v	15.65	37.40
graph-based W_e	15.39	37.09
graph-based W_vW_e	**15.81**	**37.66**
weighted phrase extraction		
weighted word-based	17.58	40.33
graph-based W_v	17.84	40.53
graph-based W_vW_e	**17.96**	**40.90**

Table 6: Results for the DIALOG task using the test set from IWSLT06.

IWSLT08 DIALOG	BLEU	METEOR
regular phrase extraction		
word-based	23.30	40.39
phrase-based	23.42	40.27
graph-based	23.52	40.37
graph-based W_v	23.97	**40.74**
graph-based W_e	23.67	40.70
graph-based W_vW_e	**24.13**	40.69
weighted phrase extraction		
weighted word-based	24.53	44.59
graph-based W_v	25.34	**45.38**
graph-based W_vW_e	**25.47**	44.62

Table 7: Results for the DIALOG task using the test set from IWSLT08.

[1] http://code.google.com/p/geppetto/

in the DIALOG corpus scores over state-of-the-art reordering models when using the weighted reordering graphs. We can see that both W_v and W_e generate improved results, and their combination generally performs better than both when used individually. The METEOR score is not always higher using our algorithm, but we believe this roots from the fact that the MERT tuning was set to optimize the BLEU score.

For the BTEC corpus, we observe that phrase-based models do not perform as well, although the difference in BLEU is only 0.26 (0.4%) in the weighted case with respect to the weighted word-based model. We believe that this is because a large percentage of translation units that are used during decoding is one-to-one, as it was reported in (Ling et al., 2010). It is highly likely that this roots from the fact that the training set is small, so the probability of finding large sequences of strings in the training set that matches the ones in the test set is rather low. In this case the word-based reordering models yield better reordering estimates, since considering longer training phrase-pairs that will not even be present in the test set will only degenerate the orientation probabilities. This suggests for future work that adding prior knowledge about the probability of a phrase-pair given its size could improve the translation quality. In the extreme case, we can give more weight to phrase pairs with the source that is present in the test set, for estimating the orientations, generating a reordering model that is specific for each test set.

5 Conclusions

In this work, we presented the current state of the art in improving the lexicalized model orientation estimates. The basic word based lexicalized reordering model uses neighboring words to perform the orientation estimates. However, since words are not always translated by themselves, these estimates can be improved by considering neighboring phrases rather than words. Another way to improve the phrase based reordering model is to consider multiple adjacent phrases, which can be done using a reordering graph. Finally, another improvement can be made by addressing the limitations of the lexicalized reordering models extracted from a single alignment, and generate the orientation estimates using weighted alignment matrices.

We extend the reordering graph model to allow the discriminative treatment of different paths. Using scores extracted from weighted alignment matrices to weight phrase pairs and a distance penalty function to penalize paths with phrase pairs that are not adjacent, improvements of 0.38 (2%) and 0.94 (3.4%) in BLEU over the state of the art algorithms using weighted alignment matrices for the Chinese-English language pair can be achieved.

As future work, we will experiment with different types of phrase pair and edge scorers and extending the weighted reordering graphs to allow multiple scorers to be combined and optimized. Additionally, we will evaluate the impact of our algorithm in larger corpora, since we believe that using a larger training corpora will result in bigger improvements over the word-based approaches, since longer sequences of words in the test set will be found in the training set, resulting in longer translation units.

The code used in this work is currently integrated with the Geppetto toolkit[2] , and it will be made available in the next version for public use.

6 Acknowledgements

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through projects CMU-PT/HuMach/0039/2008 and CMU-PT/0005/2007.

The PhD thesis of Wang Ling is supported by FCT grant SFRH/BD/51157/2010.

The authors also wish to thank the anonymous reviewers for many helpful comments.

References

Amittai Axelrod, Ra Birch Mayne, Chris Callison-burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *In Proc. International Workshop on Spoken Language Translation (IWSLT*.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing Word Lattice Translation. Technical Report LAMP-TR-149, University of Maryland, College Park, February.

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

[2] http://code.google.com/p/geppetto/

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 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*, NAACL '03, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Wang Ling, Tiago Luís, João Graça, Luísa Coheur, and Isabel Trancoso. 2010. Towards a general and extensible phrase-extraction algorithm. In *IWSLT '10: International Workshop on Spoken Language Translation*, pages 313–320, Paris, France.

Wang Ling, Tiago Luís, João Graça, Luísa Coheur, and Isabel Trancoso. 2011. Reordering modeling using weighted alignment matrices. In *Proceedings of ACL-2011: HLT*, Portland, Oregon, June. Association for Computational Linguistics.

Yang Liu, Tian Xia, Xinyan Xiao, and Qun Liu. 2009. Weighted alignment matrices for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 1017–1026, Morristown, NJ, USA. Association for Computational Linguistics.

Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proceedings of ACL-08: HLT*, pages 192–199, Columbus, Ohio, June. Association for Computational Linguistics.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the iwslt 2010 evaluation campaign. In *IWSLT '10: International Workshop on Spoken Language Translation*, pages 3–27.

Jinsong Su, Yang Liu, Yajuan Lü, Haitao Mi, and Qun Liu. 2010. Learning lexicalized reordering models from reordering graphs. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 12–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 101–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

João V. Graça, Kuzman Ganchev, and Ben Taskar. 2010. Learning Tractable Word Alignment Models with Complex Constraints. *Comput. Linguist.*, 36:481–504.

Ashish Venugopal, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. 2009. Wider pipelines: N-best alignments and parses in MT training.