Discriminative Word Alignment with a Function Word Reordering Model

Hendra Setiawan
UMIACS
University of Maryland
hendra@umiacs.umd.edu

Chris Dyer
Language Technologies Institute
Carnegie Mellon University
cdyer@cs.cmu.edu

Philip Resnik
Linguistics and UMIACS
University of Maryland
resnik@umd.edu

Abstract

We address the modeling, parameter estimation and search challenges that arise from the introduction of reordering models that capture non-local reordering in alignment modeling. In particular, we introduce several reordering models that utilize (pairs of) function words as contexts for alignment reordering. To address the parameter estimation challenge, we propose to estimate these reordering models from a relatively small amount of manually-aligned corpora. To address the search challenge, we devise an iterative local search algorithm that stochastically explores reordering possibilities. By capturing non-local reordering phenomena, our proposed alignment model bears a closer resemblance to state-of-the-art translation model. Empirical results show significant improvements in alignment quality as well as in translation performance over baselines in a large-scale Chinese-English translation task.

1 Introduction

In many Statistical Machine Translation (SMT) systems, alignment represents an important piece of information, from which translation rules are learnt. However, while translation models have evolved from word-based to syntax-based modeling, the *de facto* alignment model remains word-based (Brown et al., 1993; Vogel et al., 1996). This gap between alignment modeling and translation modeling is clearly undesirable as it often generates tensions that would prevent the extraction of many useful translation rules (DeNero and Klein, 2007). Recent work, e.g. by Blunsom et al. (2009) and Haghihi et al. (2009) just to name a few, show that alignment models that bear closer resemblance to state-of-the-art translation model consistently yields not only a better alignment quality but also an improved translation quality.

In this paper, we follow this recent effort to narrow the gap between alignment model and translation model to improve translation quality. More concretely, we focus on the reordering component since we observe that the treatment of reordering remains significantly different when comparing alignment versus translation: the reordering component in state-of-the-art translation models has focused on long-distance reordering, but its counterpart in alignment models has remained focused on local reordering, typically modeling distortion based entirely on positional information. This leaves most alignment decisions to association-based scores.

Why is employing stronger reordering models more challenging in alignment than in translation? One answer can be attributed to the fact that alignment points are unobserved in parallel text, thus so are their reorderings. As such, introducing stronger reordering often further exacerbates the computational complexity to do inference over the model. Some recent alignment models appeal to external linguistic knowledge, mostly by using monolingual syntactic parses (Cherry and Lin, 2006; Pauls et al., 2010), which at the same time, provides an approximation of the bilingual syntactic divergences that drive the reordering. To our knowledge, however, this approach has been used mainly to constrain reordering possibilities, or to add to the generalization ability of association-based scores, not to directly model reordering in the context of alignment.

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 534–544,
MIT, Massachusetts, USA, 9-11 October 2010.

In this paper, we introduce a new approach to improving the modeling of reordering in alignment. Instead of relying on monolingual parses, we condition our reordering model on the behavior of *function words* and the phrases that surround them. Function words are the "syntactic glue" of sentences, and in fact many syntacticians believe that functional categories, as opposed to substantive categories like noun and verb, are primarily responsible for cross-language syntactic variation (Ouhalla, 1991). Our reordering model can be seen as offering a reasonable approximation to more fully elaborated bilingual syntactic modeling, and this approximation is also highly practical, as it demands no external knowledge (other than a list of function words) and avoids the practical issues associated with the use of monolingual parses, e.g. whether the monolingual parser is robust enough to produce reliable output for every sentence in training data.

At a glance, our reordering model enumerates the function words on both source and target sides, modeling their reordering relative to their neighboring phrases, their neighboring function words, and the sentence boundaries. Because the frequency of function words is high, we find that by predicting the reordering of function words accurately, the reordering of the remaining words improves in accuracy as well. In total, we introduce six sub-models involving function words, and these serve as features in a log linear model. We train model weights discriminatively using Minimum Error Rate Training (MERT) (Och, 2003), optimizing F-measure.

The parameters of our sub-models are estimated from manually-aligned corpora, leading the reordering model more directly toward reproducing human alignments, rather than maximizing the likelihood of unaligned training data. This use of manual data for parameter estimation is a reasonable choice because these models depend on a small, fixed number of lexical items that occur frequently in language, hence only small training corpora are required. In addition, the availability of manually-aligned corpora has been growing steadily.

The remainder of the paper proceeds as follows. In Section 2, we provide empirical motivation for our approach. In Section 3, we discuss six sub-models based on function word relationships and how their parameters are estimated; these are combined with additional features in Section 4 to produce a single discriminative alignment model. Section 5 describes a simple decoding algorithm to find the most probable alignment under the combined model, Section 6 describes the training of our discriminative model and Section 7 presents experimental results for the model using this algorithm. We wrap up in Sections 8 and 9 with a discussion of related work and a summary of our conclusions.

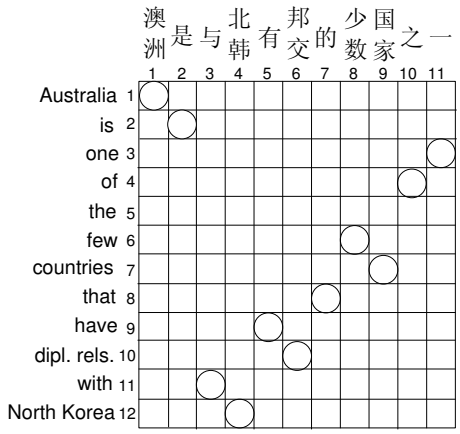

Figure 1: An aligned Chinese-English sentence pair.

2 Empirical Motivation

Fig. 1 shows an example of a Chinese-English sentence pair together with correct alignment points. Predicting the alignment for this particular Chinese-English sentence pair is challenging, since the significantly different syntactic structures of these two languages lead to non-monotone reordering. For example, an accurate alignment model should account for the fact that prepositional phrases in Chinese appear in a different order than in English, as illustrated by the movement of the phrase "与北韩/with North Korea" from the beginning of the Chinese noun phrase to the end of the corresponding English.

The central question that concerns us here is how to define and infer regularities that can be useful to predict alignment reorderings. The approach we take here is supported by empirical results from a pilot study, conducted as an inquiry into the idea of focusing on function words to model alignment reordering, which we briefly describe.

We took a Chinese-English manually-aligned corpus of approximately 21 thousand sentence pairs,

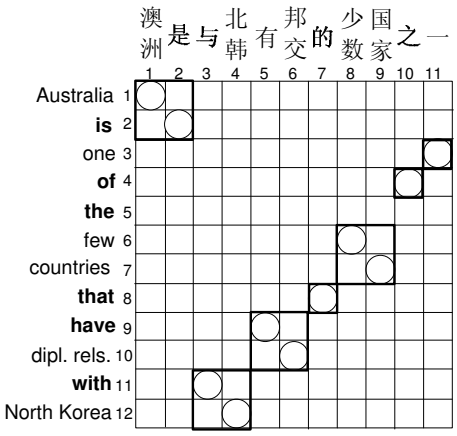

Figure 2: The all-monotone phrase pairs, indicated as rectangular areas in **bold**, that can be extracted from the Fig. 1 example.

and divided each sentence pair into *all-monotone phrase pairs*. Visually, an all-monotone phrase pair corresponds to a maximal block in the alignment matrix for which internal alignment points appear in monotone order from the top-left corner to the bottom-right corner. Fig. 2 illustrates seven such pairs that can be extracted from the example in Fig. 1. In total, there are 154,517 such phrase pairs in our manually-aligned corpus.

The alignment configuration internal to all-monotone phrase pair blocks is, obviously, monotonic, which is a configuration that is effectively modeled by traditional alignments models. On the other hand, the reordering between two adjacent blocks is the focus of our efforts since existing models are less effective at modeling non-monotonic alignment configurations. To measure the function words' potential to predict non-monotone reorderings, we examined the *border* words where two adjacent blocks meet. In particular, we are interested in how many adjacent blocks whose border words are function words.

The results of this pilot study were quite encouraging. If we consider only the Chinese side of the phrase pairs, 88.35% adjacent blocks have function words as their boundary words. If we consider only the English side, function words appear at the borders of 93.91% adjacent blocks. If we consider both the Chinese and English sides, the percentage increases to 95.53%. Notice that in Fig. 2, function words appear at the borders of *all* adjacent all-monotone phrase pairs, if both Chinese and English sides are considered. Clearly with such high coverage, function words are central in predicting non-monotone reordering in alignment.

3 Reordering with Function Words

The reordering models we describe follow our previous work using function word models for translation (Setiawan et al., 2007; Setiawan et al., 2009). The core hypothesis in this work is that function words provide robust clues to the reordering patterns of the phrases surrounding them. To make this insight useful for alignment, we develop features that score the alignment configuration of the neighboring phrases of a function word (which functions as an anchor) using two kinds of information: 1) the relative ordering of the phrases with respect to the function word anchor; and 2) the span of the phrases. This section provides a high level overview of our reordering model, which attempts to leverage this information.

To facilitate subsequent discussions, we introduce the notion of *monolingual* function word phrase FW_i, which consists of the tuple (Y_i, L_i, R_i), where Y_i is the i-th function word and L_i,R_i are its left and right neighboring phrases, respectively. Note that this notion of "phrase" is defined only for reordering purposes in our model, and does not necessarily correspond to a linguistic phrase. We define such phrases on both sides to cover as many non-monotone reorderings as possible, as suggested by the pilot study. To denote the side, we append a subscript: $FW_{i,S} = (Y_{i,S}, L_{i,S}, R_{i,S})$ refers to a function word phrase on the source side, and $FW_{i,T} = (Y_{i,T}, L_{i,T}, R_{i,T})$ to one on the target side. In our subsequent discussion, we will mainly use $FW_{i,S}$, and we will omit subscripts S or T if they are clear from context.

The primary objective of our reordering model is to predict the projection of monolingual function word phrases from one language to the other, inferring *bilingual* function word phrase pairs $FW_{i,S\rightarrow T} = (Y_{i,S\rightarrow T}, L_{i,S\rightarrow T}, R_{i,S\rightarrow T})$, which encode the two aforementioned pieces of information.[1] To infer these phrases, we take a probabilis-

[1]The subscript $S \rightarrow T$ denotes the projection direction from source to target. The subscript for the other direction is $T \rightarrow S$.

tic approach. For instance, to estimate the spans of $L_{i,S\rightarrow T}, R_{i,S\rightarrow T}$, our reordering model assumes that any span to the left of $Y_{i,S}$ is a possible $L_{i,S}$ and any span to the right of $Y_{i,S}$ is a possible $R_{i,S}$, deciding which is most probable via features, rather than committing to particular spans (e.g. as defined by a monolingual text chunker or parser). We only enforce one criterion on $L_{i,S\rightarrow T}$ and $R_{i,S\rightarrow T}$: they have to be the *maximal* alignment blocks satisfying the consistent heuristic (Och and Ney, 2004) that end or start with $Y_{i,S\rightarrow T}$ on the source S side respectively.[2]

To infer these phrases, we decompose $L_{i,S\rightarrow T}$ into $(o(L_{i,S\rightarrow T}),\ d(FW_{i-1,S\rightarrow T}),\ b(\langle s\rangle))$; similarly, $R_{i,S\rightarrow T}$ into $(o(R_{i,S\rightarrow T}),d(FW_{i+1,S\rightarrow T}),\ b(\langle/s\rangle)\ ))$. Taking the decomposition of $L_{i,S\rightarrow T}$ as a case in point, here $o(L_{i,S\rightarrow T})$ describes the reordering of the left neighbor $L_{i,S\rightarrow T}$ with respect to the function word $Y_{i,S\rightarrow T}$, while $d(FW_{i-1,S\rightarrow T})$ and $b(\langle s\rangle))$ probe the span of $L_{i,S\rightarrow T}$, i.e. whether it goes beyond the preceding function word phrase pairs $FW_{i-1,S\rightarrow T}$ and up to the beginning-of-sentence marker $\langle s\rangle$ respectively. The same definition applies to the decomposition of $R_{i,S\rightarrow T}$, where $FW_{i+1,S\rightarrow T}$ is the succeeding function word phrase pair and $\langle/s\rangle$ is the end-of-sentence marker.

3.1 Six (Sub-)Models

To model $o(L_{i,S\rightarrow T})$, $o(R_{i,S\rightarrow T})$, i.e. the reordering of the neighboring phrases of a function word, we employ the *orientation* model introduced by Setiawan et al. (2007). Formally, this model takes the form of probability distribution $P_{ori}(o(L_{i,S\rightarrow T}), o(R_{i,S\rightarrow T})|Y_{i,S\rightarrow T})$, which conditions the reordering on the lexical identity of the function word alignment (but independent of the lexical identity of its neighboring phrases). In particular, o maps the reordering into one of the following four orientation values (borrowed from Nagata et al. (2006)) with respect to the function word: Monotone Adjacent (MA), Monotone Gap (MG), Reverse Adjacent (RA) and Reverse Gap (RG). The Monotone/Reverse distinction indicates whether the projected order follows the original order, while the Adjacent/Gap distinction indicates whether the projections of the function word and the neighboring phrase are adjacent or separated by an intervening phrase.

To model $d(FW_{i-1,S\rightarrow T})$, $d(FW_{i+1,S\rightarrow T})$, i.e. whether $L_{i,S\rightarrow T}$ and $R_{i,S\rightarrow T}$ extend beyond the neighboring function word phrase pairs, we utilize the *pairwise dominance* model of Setiawan et al. (2009). Taking $d(FW_{i-1,S\rightarrow T})$ as a case in point, this model takes the form $P_{dom}(d(FW_{i-1,S\rightarrow T})|Y_{i-1,S\rightarrow T}, Y_{i,S\rightarrow T})$, where d takes one of the following four dominance values: leftFirst, rightFirst, dontCare, or neither. We will detail the exact formulation of these values in the next subsection. However, to provide intuition, the value of either leftFirst or neither for $d(FW_{i-1,S\rightarrow T})$ would suggest that the span of $L_{i,S\rightarrow T}$ doesn't extend to $Y_{i-1,S\rightarrow T}$; the further distinction between leftFirst and neither concerns with whether the span of $R_{i-1,S\rightarrow T}$ extends to $FW_{i,S\rightarrow T}$.

To model $b(\langle s\rangle)$, $b(\langle/s\rangle)$, i.e. whether the span of $L_{i,S\rightarrow T}$ and $R_{i,S\rightarrow T}$ extends up to sentence markers, we introduce the *borderwise dominance* model. Formally, this model is similar to the pairwise dominance model, except that we use the sentence boundaries as the anchors instead of the neighboring phrase pairs. This model captures longer distance dependencies compared to the previous two models; in the Chinese-English case, in particular, it is useful to discourage word alignments from crossing clause or sentence boundaries. The sentence boundary issue is especially important in machine translation (MT) experimentation, since the Chinese side of English-Chinese parallel text often includes long sentences that are composed of several independent clauses joined together; in such cases, words from one clause should be discouraged from aligning to words from other clauses. In Fig. 1, this model is potentially useful to discourage words from crossing the copula "是/is".

We define each model for all (pairs of) function word phrase pairs, forming features over a set of word alignments (A) between source (S) and target

[2]This heuristic is commonly used in learning phrase pairs from parallel text. The maximality ensures the uniqueness of L and R.

(T) sentence pair, as follows:

$$f_{ori} = \prod_{i=1}^{N} P_{ori}(o(L_i), o(R_i)|Y_i) \quad (1)$$

$$f_{dom} = \prod_{i=2}^{N} P_{dom}(d(FW_{i-1})|Y_{i-1}, Y_i) \quad (2)$$

$$f_{bdom} = \prod_{i=1}^{N} P_{bdom}(b(\langle s\rangle)|\langle s\rangle, Y_i) \cdot P_{bdom}(b(\langle /s\rangle)|Y_i, \langle /s\rangle) \quad (3)$$

where N is the number of function words (of the source side, in the $S \rightarrow T$ case). As the bilingual function word phrase pairs are uni-directional, we employ these three models in both directions, i.e. $T \rightarrow S$ as well as $S \rightarrow T$. As a result, there are six reordering models based on function words.

3.2 Prediction and Parameter Estimation

Given $FW_{i-1,S\rightarrow T}$ (and all other $FW_{\forall i'/i,S\rightarrow T}$), our reordering model has to decompose $L_{i,S\rightarrow T}$ into ($o(L_{i,S\rightarrow T})$, $d(FW_{i-1,S\rightarrow T})$, $b(\langle s\rangle)$); and $R_{i,S\rightarrow T}$ into ($o(R_{i,S\rightarrow T})$,$d(FW_{i+1,S\rightarrow T})$, $b(\langle /s\rangle)$)) during prediction and parameter estimation. In prediction mode (described in Section 5), it has to make the decomposition on the current state of alignment, while during parameter estimation, it has to make the same decomposition on the manually-aligned corpora. Since the process is identical, we proceed with the discussion in the context of parameter estimation, where the decomposition is performed to collect counts to estimate the parameters of our models.

Orientation model. Using $L_{i,S\rightarrow T}$ as a case in point and given ($Y_{i,S\rightarrow T}$=s_l^l/t_m^m, $L_{i,S\rightarrow T}$=$s_{l_1}^{l_2}/t_{m_1}^{m_2}$, $R_{i,S\rightarrow T}$=$s_{l_3}^{l_4}/t_{m_3}^{m_4}$)[3], the value of $o(L_{i,S\rightarrow T})$ in terms of Monotone/Reverse is:

$$\text{Monotone/Reverse} = \begin{cases} M, & m_2 < m, \\ R, & m < m_1. \end{cases} \quad (4)$$

while its value in terms of Adjacent/Gap values is:

$$\text{Adjacent/Gap} = \begin{cases} A, & |m - m_1| \vee |m - m_2| = 1, \\ G, & \text{otherwise.} \end{cases} \quad (5)$$

[3]We use subscripts to indicate the starting index, and superscripts the ending index.

By adjusting the indices, the computation of $o(R_{i,S\rightarrow T})$ follows similarly to the procedure above.

Suppose we want to estimate the probability of $L_{i,S\rightarrow T}$=MA for a particular Y_i. Note that here, we are interested in the lexical identity of Y_i, thus the index i is irrelevant. We first gather the counts of the orientation value for all $L_{i,S\rightarrow T}$ of Y_i in the corpus: $c(o(L_{i,S\rightarrow T}) \in \{MA, RA, MG, RG\}, Y_i)$. Then $P_{ori}(MA|Y_i)$ is estimated as follows:

$$P_{ori}(MA|Y_i) = \frac{c(MA, Y_i)}{c(Y_i)} \quad (6)$$

where $c(Y_i)$ is the frequency of Y_i in the corpus. The estimation of other orientation values as well as the $T \rightarrow S$ version of the model, follows the same procedure.

Pairwise and Borderwise dominance models. Given $R_{i,S\rightarrow T} = s_{l_1}^{l_2}/t_{m_1}^{m_2}$ and $L_{i+1,S\rightarrow T} = s_{l_3}^{l_4}/t_{m_3}^{m_4}$, i.e. the spans of the neighbors of a pair of neighboring function word phrase pairs ($Y_i = s_{l_5}^{l_5}/t_{m_5}^{m_5}, Y_{i+1} = s_{l_6}^{l_6}/t_{m_6}^{m_6}$), the value of $d(FW_{i+1,S\rightarrow T})$ is:

$$= \begin{cases} \mathsf{leftFirst}, & l_2 \geq l_6 \bigwedge l_3 > l_5 \\ \mathsf{rightFirst}, & l_2 < l_6 \bigwedge l_3 \leq l_5 \\ \mathsf{dontCare}, & l_2 \geq l_6 \bigwedge l_3 \leq l_5 \\ \mathsf{neither}, & l_2 < l_6 \bigwedge l_3 > l_5 \end{cases} \quad (7)$$

Note that the neighbors of the sentence markers for the borderwise models span the whole sentence, thus value of neither is impossible for these models.

Suppose we want to estimate the probability of Y_i and Y_{i+1} having a dontCare dominance value. Note that here we are interested in the lexical identity of Y_i and Y_{i+1}, thus the models are insensitive to the indices. We first gather the counts of the Y_i and Y_{i+1} having the dontCare value $c(\mathsf{dontCare}, Y_i, Y_{i+1})$; then $P_{dom}(\mathsf{dontCare}|Y_i, Y_{i+1})$ is estimated as follows:

$$P_{dom}(\mathsf{dontCare}|Y_i, Y_{i+1}) = \frac{c(\mathsf{dontCare}, Y_i, Y_{i+1})}{c(Y_i, Y_{i+1})} \quad (8)$$

where $c(Y_i, Y_{i+1})$ is the count of Y_i appears after Y_{i+1} in the training corpus without any other function word comes in between.

4 Alignment Model

To use the function word alignment features described in the previous section to predict alignments, we use a linear model of the following form:

$$\hat{A} = \arg \max_{A \in \mathcal{A}(S,T)} \boldsymbol{\theta} \cdot \mathbf{f}(A, S, T) \quad (9)$$

where $\mathcal{A}(S, T)$ is the set of all possible alignments of a source sentence S and target sentence T, and $\mathbf{f}(A, S, T)$ is a vector of feature functions on A, S, and T, and $\boldsymbol{\theta}$ is a parameter vector.

In addition to the six reordering models, our model employs several association-based scores that look at alignments in isolation. These features include:

1. Normalized log-likelihood ratio (LLR). This feature represents an association score, derived from statistical testing statistics. LLR (Dunning, 1993) has been widely used especially to measure lexical association. Since the values of LLR are unnormalized, we normalize them on a per-sentence basis, so that the normalized LLRs of, say, a particular source word to the target words in a particular sentence sum up to one.

2. Translation table from IBM model 4. This feature represents another association score, derived from a generative model, in particular the word-based IBM model 4. The use of this feature is widespread in recent alignment models, since it provides a relatively accurate initial prediction.

3. Translation table from manually-aligned corpora. This feature represents a gold-standard association score, based on human annotation. While attractive, this feature suffers from data sparseness issues since the lexical coverage of manually-aligned corpora, especially over content words, is very low. To overcome this issue, we design this feature to have two levels of granularity; as such, a fine-grained one is applied for function words and the coarse-grained one for content words.

4. Grow-diag-final alignments bonus. This feature encourages our alignment model to reuse alignment points that are part of the alignments created by the grow-diag-final heuristic, which we used as the baseline of our machine translation experiments.

5. Fertility model from IBM model 4. This feature, which is another by-product of IBM model 4, measures the probability of a certain word aligning to zero, one, or two or more words.

6. Null-alignment probability. This binomial feature models preference towards not aligning words, i.e. aligning to the NULL token. The intuition is to penalize NULL alignments depending on word class, by assigning lower probability mass to unaligned content words than to unaligned function words. In our experiment, we assign feature value 10^{-3} for a function word aligning to NULL, and 10^{-5} for a content word aligning to NULL.

Note that with the exception of the alignment bonus feature (4), all features are uni-directional, and therefore we employ these features in both directions just as was done for the reordering models.

5 Search

To find $\hat{A}$ using the model in Eq. 9, it is necessary to search $2^{|S|\times|T|}$ different alignment configurations, and, because of the non-local dependencies in some of our features, it is not possible to use dynamic programming to perform this search efficiently. We therefore employ an approximate search for the best alignment. We use a local search procedure which starts from some alignment (in our case, a symmetrized Model 4 alignment) and make local changes to it. Rather than taking a pure hill-climbing approach which greedily moves to locally better configurations (Brown et al., 1993), we use a stochastic search procedure which can move into lower-scoring states with some probability, similar to the Monte Carlo techniques used to draw samples from analytically intractable probability distributions.

5.1 Algorithm

To find $\hat{A}$, our search algorithm starts with an initial alignment $A^{(1)}$ and iteratively draws a new set by making a few small changes to the current set. For each step $i = [1, n]$, with alignment $A^{(i)}$, a set of neighboring alignments $\mathcal{N}(A^{(i)})$ is induced by applying small transformations (discussed below) to the current alignment. The next alignment $A^{(i+1)}$

is sampled from the following distribution:

$$p(A^{(i+1)}|S,T,A^{(i)}) = \frac{\exp \boldsymbol{\theta} \cdot \mathbf{f}(A^{(i+1)},S,T)}{Z(A^{(i)},S,T)}$$

$$\text{where } Z(A^{(i)},S,T) = \sum_{A' \in \mathcal{N}(A^{(i)})} \exp \boldsymbol{\theta} \cdot \mathbf{f}(A',S,T)$$

In addition to the current 'active' alignment configuration $A^{(i)}$, the algorithm keeps track of the highest scoring alignment observed so far, $A^{\max}$. After n steps, the algorithm returns $A^{\max}$ as its approximation of $\hat{A}$. In the experiments reported below, we initialized $A^{(1)}$ with the Model 4 alignments symmetrized by using the *grow-diag-final-and* heuristic (Koehn et al., 2003).

5.2 Alignment Neighborhoods

We now turn to a discussion of how the alignment neighborhoods used by our stochastic search algorithm are generated. We define three local transformation operations that apply to single columns of the alignment grid (which represent all of the alignments to the l^{th} source word), rows, or existing alignment points (l,m). Our three neighborhood generating operators are ALIGN, ALIGNEXCLUSIVE, and SWAP. The ALIGN operator applies to the l^{th} column of A and can either add an alignment point (l,m') or move an existing one (including to *null*, thus deleting it). ALIGNEXCLUSIVE adds an alignment point (l,m) and deletes all other points from row m. Finally, the SWAP operator swaps (l,m) and (l',m'), resulting in new alignment points (l,m') and (l',m). We increase the decoder's mobility by traversing the target side and applying the same steps above for each target word. Fig. 3 illustrates the three operators. By iterating over all columns l and rows m, the full alignment space $\mathcal{A}(S,T)$ can be explored.[4]

To further reduce the search space, an alignment point (l,m) is only admitted into a neighborhood if it is found in the high-recall alignment set $\mathcal{R}(S,T)$, which we define to be the model 4 union alignments (bidirectional model 4 symmetrized via union) plus the 5 best alignments according to the log-likelihood ratio.

[4]Using only the ALIGN operator, it is possible to explore the full alignment space; however, using all three operators increases mobility.

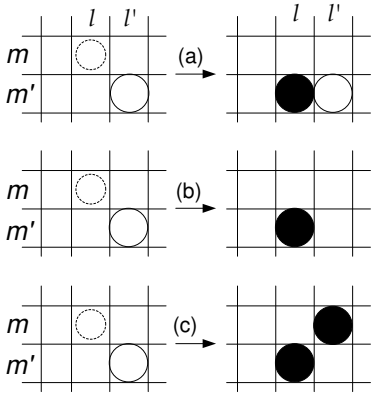

Figure 3: Illustrations for (a) ALIGN, (b) ALIGNEXCLUSIVE, and (c) SWAP operators, as applied to align the dotted, smaller circle (l,m) to (l,m'). The left hand side represents $\mathcal{A}^{(i)}$, while the right hand side represents a candidate for $\mathcal{A}^{(i+1)}$. The solid circles represent the new alignment points added to $\mathcal{A}^{(i+1)}$.

6 Discriminative Training

To set the model parameters $\boldsymbol{\theta}$, we used the minimum error rate training (MERT) algorithm (Och, 2003) to maximize the F-measure of the 1-best alignment of the model on a development set consisting of sentence pairs with manually generated alignments. The candidate set used by MERT to approximate the model is simply the set of alignments $\{A^{(1)}, A^{(2)}, \ldots, A^{(n)}\}$ encountered in the stochastic search.

While MERT does not scale to large numbers of features, the scarcity of manually aligned training data also means that models with large numbers of sparse features would be difficult to learn discriminatively, so this limitation is somewhat inherent in the problem space. Additionally, MERT has several advantages that make it particularly useful for our task. First, we can optimize F-measure of the alignments directly, which has been shown to correlate with translation quality in a downstream system (Fraser and Marcu, 2007b). Second, we are optimizing the quality of the 1-best alignments under the model. Since translation pipelines typically use only a single word alignment, this criterion is appropriate. Finally, and very importantly for us, MERT requires only an approximation of the model's hypothesis space to carry out optimization. Since we are using a stochastic search, this is crucial, since sub-

sequent evaluations of the same sentence pair (even with the same weights) may result in a different candidate set.

Although MERT is a non-probabilistic optimizer, we explore the alignment space stochastically. This is necessary to make sure that the weights we use correspond to a probability distribution that is not overly peaked (which would result in a greedy hill-climbing search) or flat (which would explore the model space without information from the model). We found that normalizing the weights by the Euclidean norm resulted in a distribution that was well-balanced between the two extremes.

7 Experiments

We evaluated our proposed alignment model intrinsically on an alignment task and extrinsically on a large-scale translation task, focusing on Chinese-English as the language pair. Our training data consists of manually aligned corpora available from LDC (LDC2006E93 and LDC2008E57) and unaligned corpora, which include FBIS, ISI, HKNews and Xinhua. In total, the manually aligned corpora consist of more than 21 thousand sentence pairs, while the unaligned corpora consist of more than 710 thousand sentence pairs. The manually-aligned corpora are primarily used for training the reordering models and for discriminative training purposes. For translation experiments, we used cdec (Dyer et al., 2010), a fast implementation of hierarchical phrase-based translation models (Chiang, 2005), which represents a state-of-the-art translation system.

We constructed the list of function words in English manually and in Chinese from (Howard, 2002). Punctuation marks were added to the list, resulting in 883 and 359 tokens in the Chinese and English lists, respectively. For the alignment experiments, we took the first 500 sentence pairs from the newswire genre of the manually-aligned corpora and used the first 250 sentences as the development set, with the remaining 250 as the test set. To ensure blind experimentation, we excluded these sentence pairs from the training of the features, including the reordering models.

7.1 Alignment Quality

We used GIZA++, the implementation of the *defacto* standard IBM alignment model, as our baseline alignment model. In particular, we used GIZA++ to align the concatenation of the development set, the test set, and the unaligned corpora, with 5, 5, 3 and 3 iterations of model 1, HMM, model 3, and model 4 respectively. Since the IBM model is asymmetric, we followed the standard practice of running GIZA++ twice, once in each direction, and combining the resulting outputs heuristically. We chose to use the grow-diag-final-and heuristic as it worked well for hierarchical phrase-based translation in our early experiments. We recorded the alignment quality of the test set as our baseline performance.

For our alignment model, we used the same set of training data. To align the test set, we first tuned the weights of the features in our discriminative alignment model using minimum error rate training (MERT) (Och, 2003) with $F_{\alpha=0.1}$ as the optimization criterion. At each iteration, our aligner outputs k-best alignments under current set of weights, from which MERT proceeds to compute the next set of weights. MERT terminates once the improvement over the previous iteration is lower than a predefined value. Once tuned, we ran our aligner on the test set and measured the quality of the resulting alignment as the performance of our model.

Model	P	R	$F_{0.5}$	$F_{0.1}$
gdfa	70.97	63.83	67.21	64.48
association	73.70	76.85	75.24	76.52
+ori	74.09	78.29	76.13	77.85
+dom	75.06	78.98	76.97	78.57
+bdom	**75.41**	**80.53**	**77.89**	**79.99**

Table 1: Alignment quality results ($F_{0.1}$) for our discriminative reordering models with various features (lines 2-5) versus the baseline IBM word-based Model 4 symmetrized using the grow-diag-final-and heuristic. The balanced $F_{0.5}$ measure is reported for reference. The best scores are **bolded**.

Table 1 reports the results of our experiments, which are conducted in an incremental fashion primarily to highlight the role of reordering modeling. The first line (gdfa) reports the baseline perfor-

mance. In the first experiment (association), we employed only the association-based features described in Section 4. As shown, we obtain a significant improvement over baseline. This result is consistent with recent literature (Fraser and Marcu, 2007a) that shows that a discriminatively trained model outperforms baseline unsupervised models like GIZA++. In the second set of experiments, we added the reordering models into our discriminative model one by one, starting with the orientation models, then the pairwise dominance model and finally the borderwise dominance model, reported in lines +ori, +dom and +bdom respectively. As shown, each additional reordering model provides a significant additional improvement. The best result is obtained by employing all reordering models. These results empirically confirm our hypothesis that we can improve alignment quality by employing reordering models that capture non-local reordering phenomena.

7.2 Translation Quality

For translation experiments, we used the products from our intrinsic experiments to learn translation rules for the hierarchical phrase-based decoder, i.e. the features weights of the +bdom experiment to align the MT training data using our discriminative model. For our translation model, we used the standard features based on the relative frequency counts, including a 5-gram language model feature trained on the English portion of the whole training data plus portions of the Gigaword v2 corpus. Specifically, we tuned the weights of these features via MERT on the NIST MT06 set and we report the result on the NIST MT02, MT03, MT04 and MT05 sets.

	MT02	MT03	MT04	MT05
gdfa	25.61	32.05	31.80	29.34
this work	**26.56**	**33.79**	**32.61**	**30.47**

Table 2: The translation performance (BLEU) of hierarchical phrase-based translation trained on training data aligned by IBM model 4 symmetrized with the grow-diag-final-and heuristic, versus being trained on alignments by our discriminative alignment model. **Bolded** scores indicate that the improvement is statistically significant.

Table 2 shows the result of our translation experiments. In our alignment model, we employed the whole set of reordering models, i.e. the one reported in the +bdom line in Table 1. As shown, our discriminative alignment model produces a consistent and significant improvement over the baseline IBM model 4 ($p < 0.01$), ranging between 0.81 and 1.71 BLEU points.

8 Related Work

The focus of our work is to strengthen the reordering component of alignment modeling. Although the *de facto* standard, the IBM models do not generalize well in practice: the IBM approach employs a series of reordering models based on the word's position, but reordering depends on syntactic context rather than absolute position in the sentence. Over the years, there have been many proposals to improve these reordering models, most notably Vogel et al. (1996), which adds a first-order dependency. Nevertheless, the use of these distortion-based models remains widespread (Marcu and Wong, 2002; Moore, 2004).

Alignment modeling is challenging because it often has to consider a prohibitively large alignment space. Efforts to constrain the space generally comes from the use of Inversion Transduction Grammar (ITG) (Wu, 1997). Recent proposals that use ITG constraints include (Haghighi et al., 2009; Blunsom et al., 2009) just to name a few. More recent models have begun to use linguistically-motivated constraints, often in combination with ITG, primarily exploiting monolingual syntactic information (Burkett et al., 2010; Pauls et al., 2010).

Our reordering model is closely related to the model proposed by Zhang and Gildea (2005; 2006; 2007a), with respect to conditioning the reordering predictions on lexical items. These related models treat their lexical items as latent variables to be estimated from training data, while our model uses a fixed set of lexical items that correspond to the class of function words. With respect to the focus on function words, our reordering model is closely related to the UALIGN system (Hermjakob, 2009). However, UALIGN uses deep syntactic analysis and hand-crafted heuristics in its model.

9 Conclusions

Languages exhibit regularities of word order that are preserved when projected to another language. We use the notion of function words to infer such regularities, resulting in several reordering models that are employed as features in a discriminative alignment model. In particular, our models predict the reordering of function words by looking at their dependencies with respect to their neighboring phrases, their neighboring function words, and the sentence boundaries. By capturing such long-distance dependencies, our proposed alignment model contributes to the effort to unify alignment and translation. Our experiments demonstrate that our alignment approach achieves both its intrinsic and extrinsic goals.

Acknowledgements

This research was supported in part by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-001. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of the sponsors.

References

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A gibbs sampler for phrasal synchronous grammar induction. In *ACL*, pages 782–790, Suntec, Singapore, August. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

David Burkett, John Blitzer, and Dan Klein. 2010. Joint parsing and alignment with weakly synchronized grammars. In *HLT-NAACL*, pages 127–135, Los Angeles, California, June. Association for Computational Linguistics.

Colin Cherry and Dekang Lin. 2006. Soft syntactic constraints for word alignment through discriminative training. In *COLING/ACL*, pages 105–112, Sydney, Australia, July. Association for Computational Linguistics.

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

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *ACL*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, March.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *ACL*, Uppsala, Sweden.

Alexander Fraser and Daniel Marcu. 2007a. Getting the structure right for word alignment: LEAF. In *EMNLP-CoNLL*, pages 51–60, Prague, Czech Republic, June. Association for Computational Linguistics.

Alexander Fraser and Daniel Marcu. 2007b. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised itg models. In *ACL*, pages 923–931, Suntec, Singapore, August. Association for Computational Linguistics.

Ulf Hermjakob. 2009. Improved word alignment with statistics and linguistic heuristics. In *EMNLP*, pages 229–237, Singapore, August. Association for Computational Linguistics.

Jiaying Howard. 2002. *A Student Handbook for Chinese Function Words*. The Chinese University Press.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HTL-NAACL*, pages 127–133, Edmonton, Alberta, Canada, May. Association for Computational Linguistics.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *EMNLP*, July 23.

Robert C. Moore. 2004. Improving ibm word alignment model 1. In *ACL*, pages 518–525, Barcelona, Spain, July.

Masaaki Nagata, Kuniko Saito, Kazuhide Yamamoto, and Kazuteru Ohashi. 2006. A clustered global phrase reordering model for statistical machine translation. In *ACL*, pages 713–720, Sydney, Australia, July. Association for Computational Linguistics.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*, pages 160–167.

Jamal Ouhalla. 1991. *Functional Categories and Parametric Variation*. Routledge.

Adam Pauls, Dan Klein, David Chiang, and Kevin Knight. 2010. Unsupervised syntactic alignment with inversion transduction grammars. In *HLT-NAACL*, pages 118–126, Los Angeles, California, June. Association for Computational Linguistics.

Hendra Setiawan, Min-Yen Kan, and Haizhou Li. 2007. Ordering phrases with function words. In *ACL*, pages 712–719, Prague, Czech Republic, June. Association for Computational Linguistics.

Hendra Setiawan, Min Yen Kan, Haizhou Li, and Philip Resnik. 2009. Topological ordering of function words in hierarchical phrase-based translation. In *ACL*, pages 324–332, Suntec, Singapore, August. Association for Computational Linguistics.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING*, pages 836–841, Copenhagen.

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

Hao Zhang and Daniel Gildea. 2005. Stochastic lexicalized inversion transduction grammar for alignment. In *ACL*. The Association for Computer Linguistics.

Hao Zhang and Daniel Gildea. 2006. Inducing word alignments with bilexical synchronous trees. In *ACL*. The Association for Computer Linguistics.