A Simple and Effective Hierarchical Phrase Reordering Model

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

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

Abstract

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

1 Introduction

Statistical phrase-based systems (**?**; **?**) have consistently delivered state-of-the-art performance in recent machine translation evaluations, yet these systems remain weak at handling word order changes. The re-ordering models used in the original phrase-based systems penalize phrase displacements proportionally to the amount of nonmonotonicity, with no consideration of the fact that some words are far more likely to be displaced than others (e.g., in English-to-Japanese translation, a verb should typically move to the end of the clause).

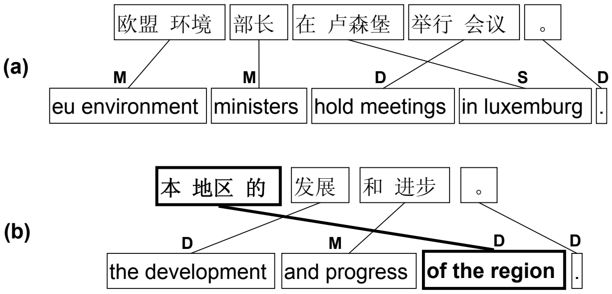

Figure 1: Phase orientations (monotone, swap, discontinuous) for Chinese-to-English translation. While previous work reasonably models phrase reordering in simple examples (a), it fails to capture more complex reorderings, such as the swapping of “of the region” (b).

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

While these lexicalized re-ordering models have shown substantial improvements over unlexicalized phrase-based systems, these models only have a limited ability to capture sensible long distance reorderings, as can be seen in Fig. 1(b). The phrase *of the region* should swap with the rest of the noun phrase, yet these previous approaches are unable to model this movement, and assume the orientation of this phrase is discontinuous (*D*). Observe that, in a shortened version of the same sentence (without *and progress*), the phrase orientation would be different (*S*), even though the shortened version has essentially the same sentence structure. Coming from the other direction, such observations about phrase reordering between different languages are precisely the kinds of facts that parsing approaches to machine translation are designed to handle and do successfully handle (**?**; **?**; **?**).

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

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

2 Lexicalized Reordering Models

We compare our re-ordering model with related work (**?**; **?**) using a log-linear approach common to many state-of-the-art statistical machine translation systems (**?**). Given an input sentence **f**, which is to be translated into a target sentence **e**, the decoder searches for the most probable translation **ê** according to the following decision rule:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \Big\{ p(\mathbf{e}|\mathbf{f}) \Big\} \quad (1)$$

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

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

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

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

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

[1] We note here that the parameterization and terminology in (**?**) is slightly different. We purposely ignore these differences in order to enable a direct comparison between Tillman's, Moses', and our approach.

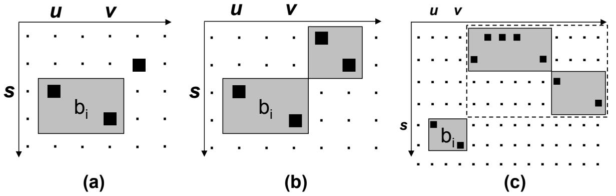

Figure 2: Occurrence of a swap according to the three orientation models: word-based, phrase-based, and hierarchical. Black squares represent word alignments, and gray squares represent blocks identified by phrase-extract. In (a), block $b_i = (e_i, f_{a_i})$ is recognized as a swap according to all three models. In (b), b_i is not recognized as a swap by the word-based model. In (c), b_i is recognized as a swap only by the hierarchical model.

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

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

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

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

3 Training

We present here three approaches for computing $p(o_i|\overline{e}_i, \overline{f}_{a_i}, a_{i-1}, a_i)$ on word-aligned data using relative frequency estimates. We assume here that phrase $\overline{e}_i$ spans the word range $s, \ldots, t$ in the target sentence **e** and that the phrase $\overline{f}_{a_i}$ spans the range $u, \ldots, v$ in the source sentence **f**. All phrase pairs in this paper are extracted with the phrase-extract algorithm (**?**), with maximum length set to 7.

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

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

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

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

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

Table 1 displays overall class distributions according to the three models. It appears clearly that occurrences of M and S are too sparsely seen in the word-based model, which assigns more than 80% of its probability mass to D. Conversely, the hierarchical model counts considerably less discontinuous cases, and is the only model that accounts for the fact that real data is predominantly monotone.

Since D is a rather uninformative default category that gives no clue how a particular phrase should be displaced, we will also provide MT evaluation scores (in Section **??**) for a set of classes

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