A Phrase Orientation Model for Hierarchical Machine Translation

Matthias Huck and **Joern Wuebker** and **Felix Rietig** and **Hermann Ney**
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{huck,wuebker,rietig,ney}@i6.informatik.rwth-aachen.de

Abstract

We introduce a lexicalized reordering model for hierarchical phrase-based machine translation. The model scores *monotone*, *swap*, and *discontinuous* phrase orientations in the manner of the one presented by Tillmann (2004). While this type of lexicalized reordering model is a valuable and widely-used component of standard phrase-based statistical machine translation systems (Koehn et al., 2007), it is however commonly not employed in hierarchical decoders.

We describe how phrase orientation probabilities can be extracted from word-aligned training data for use with hierarchical phrase inventories, and show how orientations can be scored in hierarchical decoding. The model is empirically evaluated on the NIST Chinese→English translation task. We achieve a significant improvement of +1.2 %BLEU over a typical hierarchical baseline setup and an improvement of +0.7 %BLEU over a syntax-augmented hierarchical setup. On a French→German translation task, we obtain a gain of up to +0.4 %BLEU.

1 Introduction

In hierarchical phrase-based translation (Chiang, 2005), a probabilistic synchronous context-free grammar (SCFG) is induced from bilingual training corpora. In addition to continuous *lexical* phrases as in standard phrase-based translation, *hierarchical* phrases with usually up to two nonterminals are extracted from the word-aligned parallel training data.

Hierarchical decoding is typically carried out with a parsing-based procedure. The parsing algorithm is extended to handle translation candidates and to incorporate language model scores via cube pruning (Chiang, 2007). During decoding, a hierarchical translation rule implicitly specifies the placement of the target part of a subderivation which is substituting one of its nonterminals in a partial hypothesis. The hierarchical phrase-based model thus provides an integrated reordering mechanism. The reorderings which are being conducted by the hierarchical decoder are a result of the application of SCFG rules, which generally means that there must have been some evidence in the training data for each reordering operation. At first glance one might be tempted to believe that any additional designated phrase orientation modeling would be futile in hierarchical translation as a consequence of this. We argue that such a conclusion is false, and we will provide empirical evidence in this work that lexicalized phrase orientation scoring can be highly beneficial not only in standard phrase-based systems, but also in hierarchical ones.

The purpose of a phrase orientation model is to assess the adequacy of phrase reordering during search. In standard phrase-based translation with continuous phrases only and left-to-right hypothesis generation (Koehn et al., 2003; Zens and Ney, 2008), phrase reordering is implemented by jumps within the input sentence. The choice of the best order for the target sequence is made based on the language model score of this sequence and a distortion cost that is computed from the source-side jump distances. Though the space of admissible reorderings is in most cases contrained by a maximum jump width or coverage-based restrictions (Zens et al., 2004) for efficiency reasons, the basic approach of arbitrarily jumping to uncovered positions on source side is still very permissive. Lexicalized reordering models assist the decoder in taking a good decision. Phrase-based decoding allows for a straightforward integration of lexicalized reordering models which assign

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 452–463,
Sofia, Bulgaria, August 8-9, 2013

different scores depending on how a currently translated phrase has been reordered with respect to its context. Popular lexicalized reordering models for phrase-based translation distinguish three orientation classes: *monotone*, *swap*, and *discontinuous* (Tillmann, 2004; Koehn et al., 2007; Galley and Manning, 2008). To obtain such a model, scores for the three classes are calculated from the counts of the respective orientation occurrences in the word-aligned training data for each extracted phrase. The left-to-right orientation of phrases during phrase-based search can be easily determined from the start and end positions of continuous phrases. Approximations may need to be adopted for the right-to-left scoring direction.

The utility of phrase orientation models in standard phrase-based translation is plausible and has been empirically established in practice. In hierarchical phrase-based translation, some other types of lexicalized reordering models have been investigated recently (He et al., 2010a; He et al., 2010b; Hayashi et al., 2010; Huck et al., 2012a), but in none of them are the orientation scores conditioned on the lexical identity of each phrase individually. These models are rather word-based and applied on block boundaries. Experimental results obtained with these other types of lexicalized reordering models have been very encouraging, though.

There are certain reasons why assessing the adequacy of phrase reordering should be useful in hierarchical search:

- Albeit phrase reorderings are always a result of the application of SCFG rules, the decoder is still able to choose from many different parses of the input sentence.
- The decoder can furthermore choose from many translation options for each given parse, which result in different reorderings and different phrases being embedded in the reordering non-terminals.
- All other models only weakly connect an embedded phrase with the hierarchical phrase it is placed into, in particular as the set of non-terminals of the hierarchical grammar only contains two generic non-terminal symbols.

We therefore investigate phrase orientation modeling for hierarchical translation in this work.

2 Outline

The remainder of the paper is structured as follows: We briefly outline important related publications in the following section. We subsequently give a summary of some essential aspects of the hierarchical phrase-based translation approach (Section 4). Phrase orientation modeling and a way in which a phrase orientation model can be trained for hierarchical phrase inventories are explained in Section 5. In Section 6 we introduce an extension of hierarchical search which enables the decoder to score phrase orientations. Empirical results are presented in Section 7. We conclude the paper in Section 8.

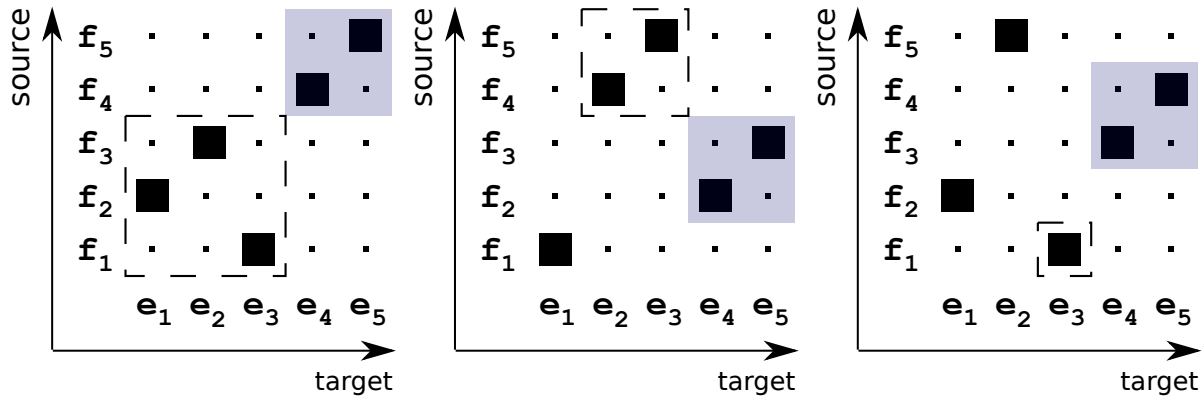
3 Related Work

Hierarchical phrase-based translation was proposed by Chiang (2005). He et al. (2010a) integrated a maximum entropy based lexicalized reordering model with word- and class-based features. Different classifiers for different rule patterns are trained for their model (He et al., 2010b). A comparable discriminatively trained model which applies a single classifier for all types of rules was developed by Huck et al. (2012a). Hayashi et al. (2010) explored the word-based reordering model by Tromble and Eisner (2009) in hierarchical translation.

For standard phrase-based translation, Galley and Manning (2008) introduced a hierarchical phrase orientation model. Similar to previous approaches (Tillmann, 2004; Koehn et al., 2007), it distinguishes the three orientation classes *monotone*, *swap*, and *discontinuous*. However, it differs in that it is not limited to model local reordering phenomena, but allows for phrases to be hierarchically combined into *blocks* in order to determine the orientation class. This has the advantage that probability mass is shifted from the rather uninformative default category *discontinuous* to the other two orientation classes, which model the location of a phrase more specifically. In this work, we transfer this concept to a hierarchical phrase-based machine translation system.

4 Hierarchical Phrase-Based Translation

The non-terminal set of a standard hierarchical grammar comprises two symbols which are shared by source and target: the initial symbol S and one generic non-terminal symbol X. The generic non-terminal X is used as a placeholder for the gaps



(a) Monotone phrase orientation. (b) Swap phrase orientation. (c) Discontinuous phrase orientation.

Figure 1: Extraction of the orientation classes *monotone*, *swap*, and *discontinuous* from word-aligned training samples. The examples show the left-to-right orientation of the shaded phrases. The dashed rectangles indicate how the predecessor words are merged into *blocks* with regard to their word alignment.

within the right-hand side of hierarchical translation rules as well as on all left-hand sides of the translation rules that are extracted from the parallel training corpus.

Extracted rules of a standard hierarchical grammar are of the form $X \rightarrow \langle \alpha, \beta, \sim \rangle$ where $\langle \alpha, \beta \rangle$ is a bilingual phrase pair that may contain X, i.e. $\alpha \in (\{X\} \cup V_F)^+$ and $\beta \in (\{X\} \cup V_E)^+$, where V_F and V_E are the source and target vocabulary, respectively. The non-terminals on the source side and on the target side of hierarchical rules are linked in a one-to-one correspondence. The $\sim$ relation defines this one-to-one correspondence. In addition to the extracted rules, a non-lexicalized *initial rule*

$$S \rightarrow \langle X^{\sim 0}, X^{\sim 0} \rangle \tag{1}$$

is engrafted into the hierarchical grammar, as well as a special *glue rule*

$$S \rightarrow \langle S^{\sim 0} X^{\sim 1}, S^{\sim 0} X^{\sim 1} \rangle \tag{2}$$

that the system can use for serial concatenation of phrases as in monotonic phrase-based translation. The initial symbol S is the start symbol of the grammar.

Hierarchical search is conducted with a customized version of the CYK+ parsing algorithm (Chappelier and Rajman, 1998) and cube pruning (Chiang, 2007). A hypergraph which represents the whole parsing space is built employing CYK+. Cube pruning operates in bottom-up topological order on this hypergraph and expands at most k derivations at each hypernode.

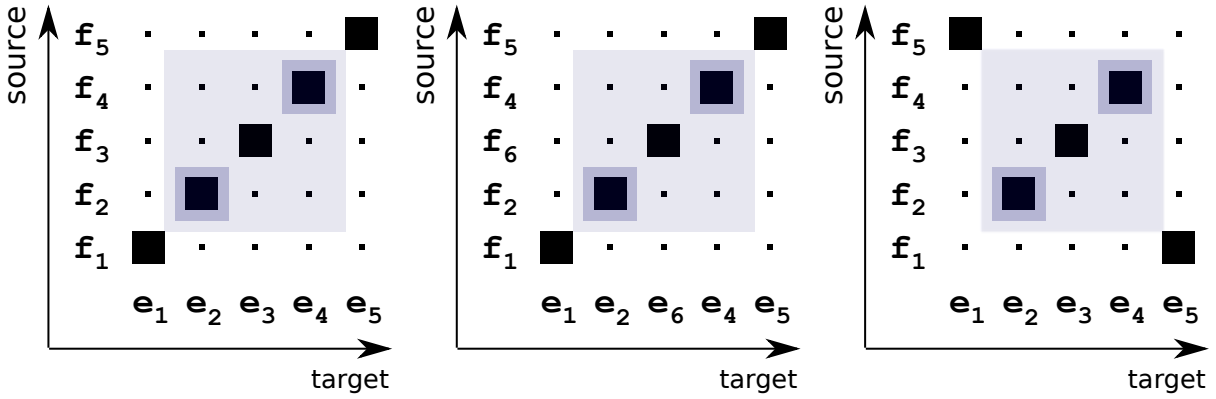
5 Modeling Phrase Orientation for Hierarchical Machine Translation

The phrase orientation model we are using was introduced by Galley and Manning (2008). To model the sequential order of phrases within the global translation context, the three orientation classes *monotone* (M), *swap* (S) and *discontinuous* (D) are distinguished, each in both left-to-right and right-to-left direction. In order to capture the global rather than the local context, previous phrases can be merged into *blocks* if they are consistent with respect to the word alignment. A phrase is in monotone orientation if a consistent monotone predecessor block exists, and in swap orientation if a consistent swap predecessor block exists. Otherwise it is in discontinuous orientation.

Given a sequence of source words f_1^J and a sequence of target words e_1^I, a block $\langle f_{j_1}^{j_2}, e_{i_1}^{i_2} \rangle$ (with $1 \leq j_1 \leq j_2 \leq J$ and $1 \leq i_1 \leq i_2 \leq I$) is *consistent* with respect to the word alignment $A \subseteq \{1, ..., I\} \times \{1, ..., J\}$ *iff*

$$\begin{aligned} &\exists (i,j) \in A : i_1 \leq i \leq i_2 \wedge j_1 \leq j \leq j_2 \\ &\wedge \forall (i,j) \in A : i_1 \leq i \leq i_2 \leftrightarrow j_1 \leq j \leq j_2. \end{aligned} \tag{3}$$

Consistency is based upon two conditions in this definition: (1.) At least one source and target position within the block must be aligned, and (2.) words from inside the source interval may only be aligned to words from inside the target interval and vice versa. These are the same conditions as those that are applied for the extraction of



(a) A monotone orientation.

Left-to-right orientation counts:

$$N(M|f_2X^{\sim 0}f_4, e_2X^{\sim 0}e_4) = 1$$
$$N(S|f_2X^{\sim 0}f_4, e_2X^{\sim 0}e_4) = 0$$
$$N(D|f_2X^{\sim 0}f_4, e_2X^{\sim 0}e_4) = 0$$

(b) Another monotone orientation.

Left-to-right orientation counts:

$$N(M|f_2X^{\sim 0}f_4, e_2X^{\sim 0}e_4) = 2$$
$$N(S|f_2X^{\sim 0}f_4, e_2X^{\sim 0}e_4) = 0$$
$$N(D|f_2X^{\sim 0}f_4, e_2X^{\sim 0}e_4) = 0$$

(c) A swap orientation.

Left-to-right orientation counts:

$$N(M|f_2X^{\sim 0}f_4, e_2X^{\sim 0}e_4) = 2$$
$$N(S|f_2X^{\sim 0}f_4, e_2X^{\sim 0}e_4) = 1$$
$$N(D|f_2X^{\sim 0}f_4, e_2X^{\sim 0}e_4) = 0$$

Figure 2: Accumulation of orientation counts for hierarchical phrases during extraction. The hierarchical phrase $\langle f_2X^{\sim 0}f_4, e_2X^{\sim 0}e_4\rangle$ (dark shaded) can be extracted from all the three training samples. Its orientation is identical to the orientation of the continuous phrase (lightly shaded) which the sub-phrase is cut out of, respectively. Note that the actual lexical content of the sub-phrase may differ. For instance, the sub-phrase $\langle f_3, e_3\rangle$ is being cut out in Fig. 2a, and the sub-phrase $\langle f_6, e_6\rangle$ is being cut out in Fig. 2b.

standard continuous phrases. The only difference is that length constraints are applied to phrases, but not to blocks.

Figure 1 illustrates the extraction of monotone, swap, and discontinuous orientation classes in left-to-right direction from word-aligned bilingual training samples. The right-to-left direction works analogously.

We found that this concept can be neatly plugged into the hierarchical phrase-based translation paradigm, without having to resort to approximations in decoding, which is necessary to determine the right-to-left orientation in a standard phrase-based system (Cherry et al., 2012). To train the orientations, the extraction procedure from the standard phrase-based version of the reordering model can be used with a minor extension. The model is trained on the same word-aligned data from which the translation rules are extracted. For each training sentence, we extract all phrases of unlimited length that are consistent with the word alignment, and store their corners in a matrix. The corners are distinguished by their location: top-left, top-right, bottom-left, and bottom-right. For each bilingual phrase, we determine its left-to-right and right-to-left orientation by checking for adjacent corners.

The lexicalized orientation probability for the orientation $O \in \{M, S, D\}$ and the phrase pair $\langle\alpha, \beta\rangle$ is estimated as its smoothed relative frequency:

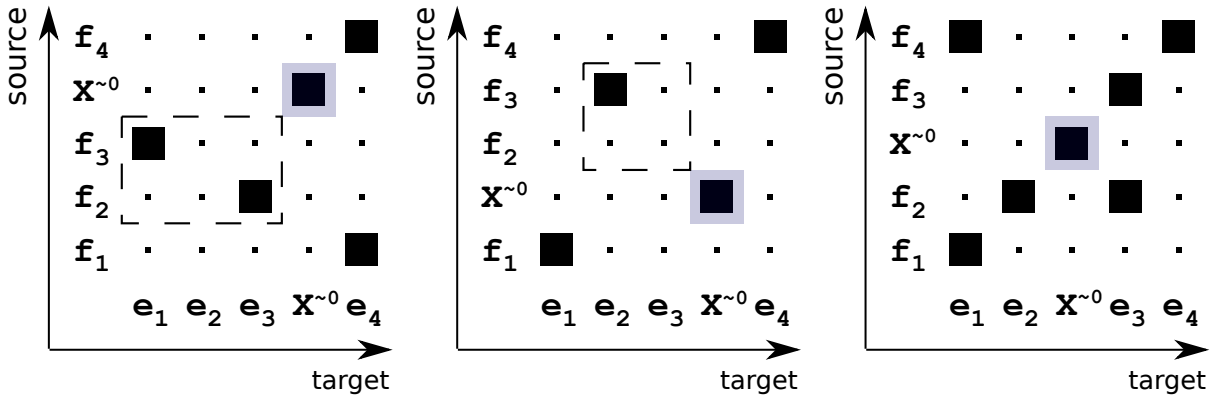
$$p(O) = \frac{N(O)}{\sum_{O' \in \{M,S,D\}} N(O')} \tag{4}$$

$$p(O|\alpha, \beta) = \frac{\sigma \cdot p(O) + N(O|\alpha, \beta)}{\sigma + \sum_{O' \in \{M,S,D\}} N(O'|\tilde{f}, \tilde{e})}. \tag{5}$$

Here, $N(O)$ denotes the global count and $N(O|\alpha, \beta)$ the lexicalized count for the orientation O. σ is a smoothing constant.

To determine the orientation frequency for a hierarchical phrase with non-terminal symbols, the orientation counts of all those phrases are accumulated from which a sub-phrase is cut out and replaced by a non-terminal symbol to obtain this hierarchical phrase. Figure 2 gives an example.

Negative logarithms of the values are used as costs in the log-linear model combination (Och and Ney, 2002). Cost 0 for all orientations is assigned to the special rules which are not extracted from the training data (initial and glue rule).



(a) Monotone non-terminal orientation. (b) Swap non-terminal orientation. (c) Discontinuous non-terminal orientation.

Figure 3: Scoring with the orientation classes *monotone*, *swap*, and *discontinuous*. Each picture shows exactly one hierarchical phrase. The block which replaces the non-terminal X during decoding is embedded with the orientation of this non-terminal X within the hierarchical phrase. The examples show the left-to-right orientation of the non-terminal. The left-to-right orientation can be detected from the word alignment of the hierarchical phrase, except for cases where the non-terminal is in boundary position on target side.

6 Phrase Orientation Scoring in Hierarchical Decoding

Our implementation of phrase orientation scoring in hierarchical decoding is based on the observation that hierarchical rule applications, i.e. the usage of rules with non-terminals within their right-hand sides, settle the target sequence order. Monotone, swap, or discontinuous orientations of blocks are each due to monotone, swap, or discontinuous placement of non-terminals which are being substituted by these blocks.

The problem of phrase orientation scoring can thus be mostly reduced to three steps which need to be carried out whenever a hierarchical rule is applied:

1. Determining the orientations of the non-terminals in the rule.

2. Retrieving the proper orientation cost of the topmost rule application in the sub-derivation which corresponds to the embedded block for the respective non-terminal.

3. Applying the orientation cost to the log-linear model combination for the current derivation.

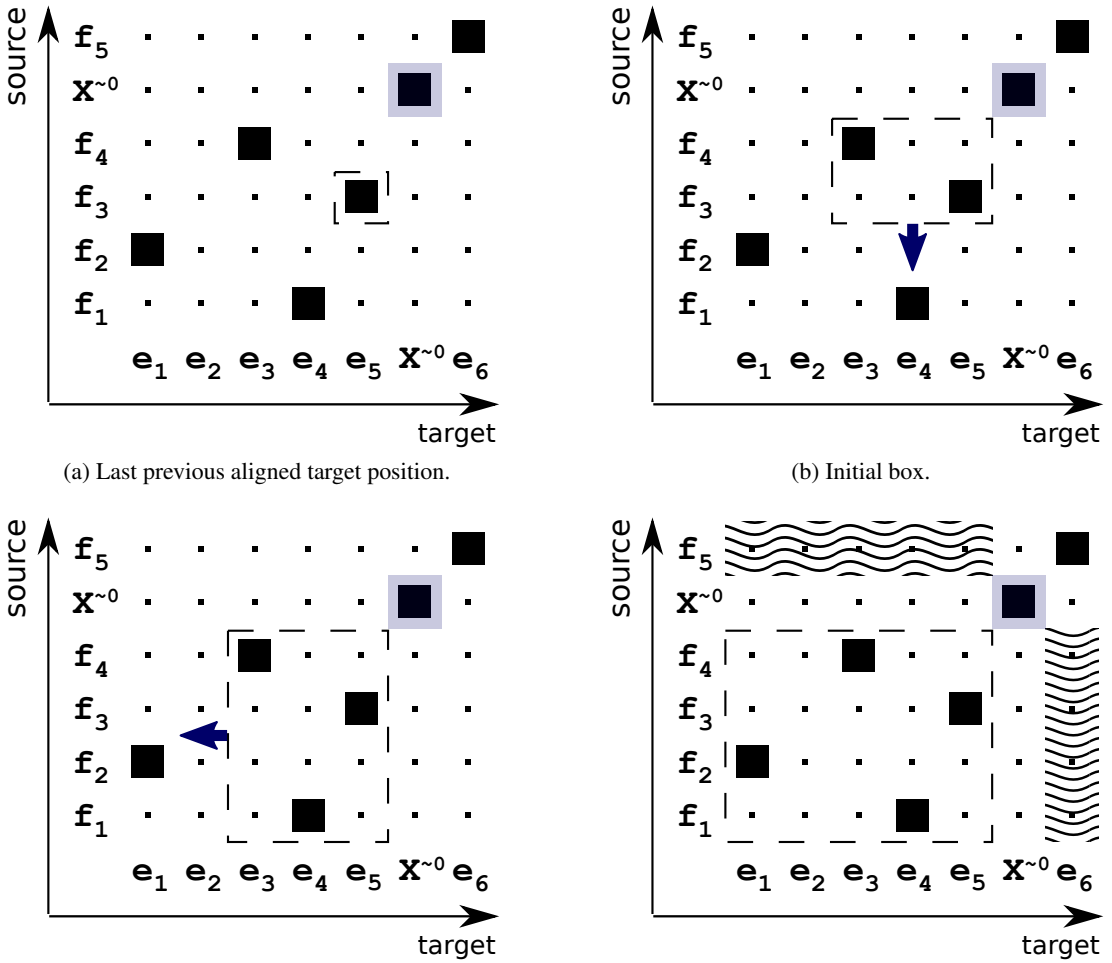
The orientation of a non-terminal in a hierarchical rule is dependent on the word alignments in its context. Figure 3 depicts three examples.[1] We however need to deal with special cases where a non-terminal orientation cannot be established at the moment when the hierarchical rule is considered. We first describe the non-degenerate case (Section 6.1). Afterwards we briefly discuss our strategy in the special situation of *boundary non-terminals* where the non-terminal orientation cannot be determined from information which is inherent to the hierarchical rule under consideration (Section 6.3).

We focus on left-to-right orientation scoring; right-to-left scoring is symmetric.

6.1 Determining Orientations

In order to determine the orientation class of a non-terminal, we rely on the word alignments within the phrases. With each phrase, we store the alignment matrix that has been seen most frequently during phrase extraction. Non-terminal symbols on target side are assumed to be aligned to the respective non-terminal symbols on source

[1] Note that even maximal consecutive lexical intervals (either on source or target side) are not necessarily aligned in a way which makes them consistent bilingual blocks. In Fig. 3a, e_4 is for instance aligned both below and above the non-terminal. In Fig. 3c, neither $\langle f_1 f_2, e_1 e_2 \rangle$ nor $\langle f_1 f_2, e_3 e_4 \rangle$ would be valid continuous phrases (the same holds for $\langle f_3 f_4, e_1 e_2 \rangle$ and $\langle f_3 f_4, e_3 e_4 \rangle$). We actually need the generalization of the phrase orientation model to hierarchical phrases as described in Section 5 for this reason. Otherwise we would be able to just score neighboring consistent sub-blocks with a model that does not account for hierarchical phrases with non-terminals.



(a) Last previous aligned target position.

(b) Initial box.

(c) Expansion of the initial box.

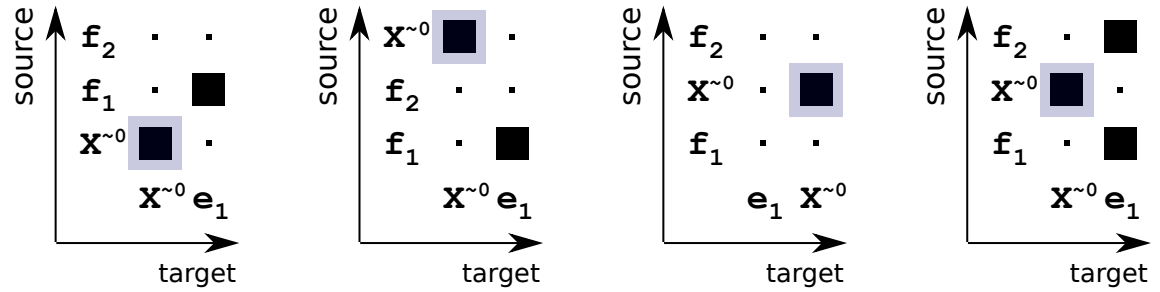
(d) The final box is a consistent left-to-right monotone predecessor block of the non-terminal.

Figure 4: Determining the orientation class during decoding. Starting from the last previous aligned target position, a box is spanned across the relevant alignment links onto the corner of the non-terminal. The box is then checked for consistency.

side according to the $\sim$ relation. In the alignment matrix, the rows and columns of non-terminals can obviously contain only exactly this one alignment link.

Starting from the last previous aligned target position to the left of the non-terminal, the algorithm expands a box that spans across the other relevant alignment links onto the corner of the non-terminal. Afterwards it checks whether the areas on the opposite sides of the non-terminal position are non-aligned in the source and target intervals of this box. The non-terminal is in discontinuous orientation if the box is not a consistent block. If the box is a consistent block, the non-terminal is in monotone orientation if its source-side position is larger than the maximum of the source-side interval of the box, and in swap orientation if its source-side position is smaller than the minimum of the source-side interval of the box.

Figure 4 illustrates how the procedure operates. In left-to-right direction, an initial box is spanned from the last previous aligned target position to the lower (monotone) or upper (swap) left corner of the non-terminal. In the example, starting from $\langle f_3, e_5 \rangle$ (Fig. 4a), this initial box is spanned to the lower left corner by iterating from f_3 to f_4 and expanding its target interval to the minimum aligned target position within these two rows of the alignment matrix. The initial box covers $\langle f_3 f_4, e_3 e_4 e_5 \rangle$ (Fig. 4b). The procedure then repeatedly checks whether the box needs to be expanded—alternating to the bottom (monotone) or top (swap) and to the left—until no alignment links below or to the left of the box break the consistency. Two box expansion are conducted in the example: the first one expands the initial box below, resulting in a larger box which covers $\langle f_1 f_2 f_3 f_4, e_3 e_4 e_5 \rangle$ (Fig. 4c); the second



(a) Left boundary non-terminal that can be placed in left-to-right monotone or discontinuous orientation when the phrase is embedded into another one.

(b) Left boundary non-terminal that can be placed in left-to-right discontinuous or swap orientation when the phrase is embedded into another one.

(c) Left boundary non-terminal that can be placed in left-to-right monotone, swap, or discontinuous orientation when the phrase is embedded into another one.

(d) Left boundary non-terminal that can only be placed in left-to-right discontinuous orientation when the phrase is embedded into another one.

Figure 5: Left boundary non-terminal symbols. Orientations the non-terminal can eventually turn out to get placed in differ depending on existing alignment links in the rest of the phrase. Delayed left-to-right scoring is not required in cases as in Fig. 5d. Fractional costs for the possible orientations are temporarily applied in the other cases and recursively corrected as soon as an orientation is constituted in an upper hypernode.

one expands this new box to the left, resulting in a final box which covers $\langle f_1 f_2 f_3 f_4, e_1 e_2 e_3 e_4 e_5 \rangle$ (Fig. 4d) and does not need to be expanded towards the lower left corner any more. Afterwards the procedure examines whether the final box is a consistent block by inspecting whether the areas on the opposite side of the non-terminal position are non-aligned in the intervals of the box (areas with waved lines in the Fig. 4d). These areas do not contain alignment links in the example: the orientation class of the non-terminal is *monotone* as it has a consistent left-to-right monotone predecessor block. (Suppose an alignment link $\langle f_5, e_2 \rangle$ would break the consistency: the orientation class would then be *discontinuous* as the final box would not be a consistent block.)

Orientations of non-terminals could basically be precomputed and stored in the translation table. We however compute them on demand during decoding. The computational overhead did not seem to be too severe in our experiments.

6.2 Scoring Orientations

Once the orientation is determined, the proper orientation cost of the embedded block needs to be retrieved. We access the topmost rule application in the sub-derivation which corresponds to the embedded block for the respective non-terminal and read the orientation model costs for this rule. The special case of delayed scoring for boundary non-terminals as described in the subsequent section is recursively processed if necessary. The retrieved orientation costs of the embedded blocks of all non-terminals are finally added to the log-linear model combination for the current derivation.

6.3 Boundary Non-Terminals

Cases where a non-terminal orientation cannot be established at the moment when the hierarchical rule is considered arise when a non-terminal symbol is in a *boundary position* on target side. We define a non-terminal to be in (left or right) boundary position *iff* no symbols are aligned between the phrase-internal target-side index of the non-terminal and the (left or right) phrase boundary. Left boundary positions of non-terminals are critical for left-to-right orientation scoring, right boundary positions for right-to-left orientation scoring. We denote non-terminals in boundary position as *boundary non-terminals*.

The procedure as described in Section 6.1 is not applicable to boundary non-terminals because a last previous aligned target position does not exist. If it is impossible to determine the final non-terminal orientation in the hypothesis from information which is inherent to the phrase, we are forced to delay the orientation scoring of the embedded block. Our solution in these cases is to heuristically add fractional costs of all orientations the non-terminal can still eventually turn out to get placed in (cf. Figure 5). We do so because not adding an orientation cost to the derivation would give it an unjustified advantage over other ones. As soon as an orientation is constituted in an up-

per hypernode, any heuristic and actual orientation costs can be collected by means of a recursive call. Note that monotone or swap orientations in upper hypernodes can top-down transition into discontinuous orientations for boundary non-terminals, depending on existing phrase-internal alignment links in the context of the respective boundary non-terminal. In the derivation at the upper hypernode, the heuristic costs are subtracted and the correct actual costs added. Delayed scoring can lead to search errors; in order to keep them confined, the delayed scoring needs to be done separately for all derivations, not just for the first-best subderivations along the incoming hyperedges.

7 Experiments

We evaluate the effect of phrase orientation scoring in hierarchical translation on the Chinese→English 2008 NIST task[2] and on the French→German language pair using the standard WMT[3] newstest sets for development and testing.

7.1 Experimental Setup

We work with a Chinese–English parallel training corpus of 3.0 M sentence pairs (77.5 M Chinese / 81.0 M English running words). To train the German→French baseline system, we use 2.0 M sentence pairs (53.1 M French / 45.8 M German running words) that are partly taken from the Europarl corpus (Koehn, 2005) and have partly been collected within the Quaero project.[4]

Word alignments are created by aligning the data in both directions with GIZA++[5] and symmetrizing the two trained alignments (Och and Ney, 2003). When extracting phrases, we apply several restrictions, in particular a maximum length of ten on source and target side for lexical phrases, a length limit of five on source and ten on target side for hierarchical phrases (including non-terminal symbols), and no more than two non-terminals per phrase.

A standard set of models is used in the baselines, comprising phrase translation probabilities and lexical translation probabilities in both directions, word and phrase penalty, binary features marking hierarchical rules, glue rule, and rules with non-terminals at the boundaries, three simple count-based binary features, phrase length ratios, and a language model. The language models are 4-grams with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998) which have been trained with the SRILM toolkit (Stolcke, 2002).

Model weights are optimized against BLEU (Papineni et al., 2002) with MERT (Och, 2003) on 100-best lists. For Chinese→English we employ MT06 as development set, MT08 is used as unseen test set. For German→French we employ newstest2009 as development set, newstest2008, newstest2010, and newstest2011 are used as unseen test sets. During decoding, a maximum length constraint of ten is applied to all non-terminals except the initial symbol S. Translation quality is measured in truecase with BLEU and TER (Snover et al., 2006). The results on MT08 are checked for statistical significance over the baseline. Confidence intervals have been computed using bootstrapping for BLEU and Cochran's approximate ratio variance for TER (Leusch and Ney, 2009).

7.2 Chinese→English Experimental Results

Table 1 comprises all results of our empirical evaluation on the Chinese→English task.

We first compare the performance of the phrase orientation model in left-to-right direction only with the performance of the phrase orientation model in left-to-right and right-to-left direction (*bidirectional*). In all experiments, monotone, swap, and discontinuous orientation costs are treated as being from different feature functions in the log-linear model combination: we assign a separate scaling factor to each of the orientations. We have three more scaling factors than in the baseline for left-to-right direction only, and six more scaling factors for bidirectional phrase orientation scoring. As can be seen from the results table, the left-to-right model already yields a gain of 1.1 %BLEU over the baseline on the unseen test set (MT08). The bidirectional model performs just slightly better (+1.2 %BLEU over the baseline). With both models, the TER is reduced significantly as well (-1.1 / -1.3 compared to the baseline). We adopted the discriminative lexicalized reordering model (*discrim. RO*) that has been suggested by Huck et al. (2012a) for comparison purposes. The phrase orientation model provides clearly better translation quality in our experiments.

[2] http://www.itl.nist.gov/iad/mig/tests/mt/2008/

[3] http://www.statmt.org/wmt13/translation-task.html

[4] http://www.quaero.org

[5] http://code.google.com/p/giza-pp/

NIST Chinese→English	MT06 (Dev) BLEU [%]	MT06 (Dev) TER [%]	MT08 (Test) BLEU [%]	MT08 (Test) TER [%]
HPBT Baseline	32.6	61.2	25.2	66.6
+ discrim. RO	33.0	61.3	25.8	66.0
+ phrase orientation (left-to-right)	33.3	60.7	**26.3**	**65.5**
+ phrase orientation (bidirectional)	33.2	60.6	**26.4**	**65.3**
+ swap rule	32.8	61.7	25.8	66.6
+ discrim. RO	33.1	61.2	**26.0**	66.1
+ phrase orientation (bidirectional)	33.3	60.7	**26.5**	**65.3**
+ binary swap feature	33.2	61.0	25.9	66.2
+ discrim. RO	33.2	61.3	**26.2**	66.1
+ phrase orientation (bidirectional)	33.6	60.5	**26.6**	**65.1**
+ soft syntactic labels	33.4	60.8	**26.1**	66.4
+ phrase orientation (bidirectional)	33.7	60.1	**26.8**	**65.1**
+ phrase-level s2t+t2s DWL + triplets	34.3	60.1	**27.7**	**65.0**
+ discrim. RO	34.8	59.8	**27.7**	**64.7**
+ phrase orientation (bidirectional)	35.3	59.0	**28.4**	**63.7**

Table 1: Experimental results for the NIST Chinese→English translation task (truecase). On the test set, bold font indicates results that are significantly better than the baseline ($p < .05$).

As a next experiment, we bring in more reordering capabilities by augmenting the hierarchical grammar with a single *swap rule*

$$X \rightarrow \langle X^{\sim 0} X^{\sim 1}, X^{\sim 1} X^{\sim 0} \rangle \quad (6)$$

supplementary to the initial rule and glue rule. The swap rule allows adjacent phrases to be transposed. The setup with swap rule and bidirectional phrase orientation model is about as good as the setup with just the bidirectional phrase orientation model and no swap rule. If we furthermore mark the swap rule with a binary feature (*binary swap feature*), we end up at an improvement of +1.4 %BLEU over the baseline. The phrase orientation model again provides higher translation quality than the discriminative reordering model.

In a third experiment, we investigate whether the phrase orientation model also has a positive influence when integrated into a syntax-augmented hierarchical system. We configured a hierarchical setup with *soft syntactic labels* (Stein et al., 2010), a syntactic enhancement in the manner of preference grammars (Venugopal et al., 2009). On MT08, the syntax-augmented system performs 0.9 %BLEU above the baseline setup. We achieve an additional improvement of +0.7 %BLEU and -1.3 TER by including the bidirectional phrase orientation model. Interestingly, the translation quality of the setup with soft syntactic labels (but without phrase orientation model) is worse than of the setup with phrase orientation model (but without soft syntactic labels) on MT08. The combination of both extensions provides the best result, though.

In a last experiment, we finally took a very strong setup which improves over the baseline by 2.5 %BLEU through the integration of phrase-level discriminative word lexicon (*DWL*) models and *triplet* lexicon models in source-to-target (s2t) and target-to-source (t2s) direction. The models have been presented by Hasan et al. (2008), Bangalore et al. (2007), and Mauser et al. (2009). We apply them in a similar manner as proposed by Huck et al. (2011). In this strong setup, the discriminative reordering model gives gains on the development set which barely carry over to the test set. Adding the bidirectional phrase orientation model, in contrast, results in a nice gain of +0.7 %BLEU and a reduction of 1.3 points in TER on the test set, even on top of the DWL and triplet lexicon models.

7.3 French→German Experimental Results

Table 2 comprises the results of our empirical evaluation on the French→German task.

The left-to-right phrase orientation model boosts the translation quality by up to 0.3 %BLEU. The reduction in TER is in a similar order of magnitude. The bidirectional model performs a bit better again, with an advancement of up to 0.4 %BLEU and a maximal reduction in TER of 0.6 points.

French→German	newstest2008 BLEU [%]	newstest2008 TER [%]	newstest2009 BLEU [%]	newstest2009 TER [%]	newstest2010 BLEU [%]	newstest2010 TER [%]	newstest2011 BLEU [%]	newstest2011 TER [%]
HPBT Baseline	15.2	71.7	15.0	71.7	15.7	69.5	14.2	72.2
+ phrase orientation (left-to-right)	15.1	71.4	15.3	71.4	15.9	69.2	14.5	71.8
+ phrase orientation (bidirectional)	15.4	71.1	15.4	71.3	15.9	69.1	14.6	71.6

Table 2: Experimental results for the French→German translation task (truecase). newstest2009 is used as development set.

8 Conclusion

In this paper, we introduced a phrase orientation model for hierarchical machine translation. The training of a lexicalized reordering model which assigns probabilities for *monotone*, *swap*, and *discontinuous* orientation of phrases was generalized from standard continuous phrases to hierarchical phrases. We explained how phrase orientation scoring can be implemented in hierarchical decoding and conducted a number of experiments on a Chinese→English and a French→German translation task. The results indicate that phrase orientation modeling is a very suitable enhancement of the hierarchical paradigm.

Our implementation will be released as part of Jane (Vilar et al., 2010; Vilar et al., 2012; Huck et al., 2012b), the RWTH Aachen University open source statistical machine translation toolkit.[6]

Acknowledgments

This work was partly achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation. This material is also partly based upon work supported by the DARPA BOLT project under Contract No. HR0011-12-C-0015. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DARPA. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

References

Srinivas Bangalore, Patrick Haffner, and Stephan Kanthak. 2007. Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 152–159, Prague, Czech Republic, June.

Jean-Cédric Chappelier and Martin Rajman. 1998. A Generalized CYK Algorithm for Parsing Stochastic CFG. In *Proc. of the First Workshop on Tabulation in Parsing and Deduction*, pages 133–137, Paris, France, April.

Stanley F. Chen and Joshua Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, USA, August.

Colin Cherry, Robert C. Moore, and Chris Quirk. 2012. On Hierarchical Re-ordering and Permutation Parsing for Phrase-based Decoding. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 200–209, Montréal, Canada, June.

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 263–270, Ann Arbor, MI, USA, June.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228, June.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 847–855, Honolulu, HI, USA, October.

Saša Hasan, Juri Ganitkevitch, Hermann Ney, and Jesús Andrés-Ferrer. 2008. Triplet Lexicon Models for Statistical Machine Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 372–381, Honolulu, HI, USA, October.

Katsuhiko Hayashi, Hajime Tsukada, Katsuhito Sudoh, Kevin Duh, and Seiichi Yamamoto. 2010. Hierarchical Phrase-based Machine Translation with Word-based Reordering Model. In *Proc. of the Int. Conf. on Computational Linguistics (COLING)*, pages 439–446, Beijing, China, August.

[6] http://www.hltpr.rwth-aachen.de/jane/

Zhongjun He, Yao Meng, and Hao Yu. 2010a. Extending the Hierarchical Phrase Based Model with Maximum Entropy Based BTG. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, Denver, CO, USA, October/November.

Zhongjun He, Yao Meng, and Hao Yu. 2010b. Maximum Entropy Based Phrase Reordering for Hierarchical Phrase-based Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 555–563, Cambridge, MA, USA, October.

Matthias Huck, Saab Mansour, Simon Wiesler, and Hermann Ney. 2011. Lexicon Models for Hierarchical Phrase-Based Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 191–198, San Francisco, CA, USA, December.

Matthias Huck, Stephan Peitz, Markus Freitag, and Hermann Ney. 2012a. Discriminative Reordering Extensions for Hierarchical Phrase-Based Machine Translation. In *Proc. of the Annual Conf. of the European Assoc. for Machine Translation (EAMT)*, pages 313–320, Trento, Italy, May.

Matthias Huck, Jan-Thorsten Peter, Markus Freitag, Stephan Peitz, and Hermann Ney. 2012b. Hierarchical Phrase-Based Translation with Jane 2. *The Prague Bulletin of Mathematical Linguistics*, 98:37–50, October.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-Off for M-gram Language Modeling. In *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 181–184, Detroit, MI, USA, May.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 127–133, Edmonton, Canada, May/June.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Demo and Poster Sessions, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of the MT Summit X*, Phuket, Thailand, September.

Gregor Leusch and Hermann Ney. 2009. Edit distances with block movements and error rate confidence estimates. *Machine Translation*, 23(2):129–140, December.

Arne Mauser, Saša Hasan, and Hermann Ney. 2009. Extending Statistical Machine Translation with Discriminative and Trigger-Based Lexicon Models. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 210–218, Singapore, August.

Franz Josef Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, USA, July.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, USA, July.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, pages 223–231, Cambridge, MA, USA, August.

Daniel Stein, Stephan Peitz, David Vilar, and Hermann Ney. 2010. A Cocktail of Deep Syntactic Features for Hierarchical Machine Translation. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, Denver, CO, USA, October/November.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO, USA, September.

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, Boston, MA, USA.

Roy Tromble and Jason Eisner. 2009. Learning Linear Ordering Problems for Better Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 1007–1016, Singapore, August.

Ashish Venugopal, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. 2009. Preference Grammars:

Softening Syntactic Constraints to Improve Statistical Machine Translation. In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 236–244, Boulder, CO, USA, June.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open Source Hierarchical Translation, Extended with Reordering and Lexicon Models. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 262–270, Uppsala, Sweden, July.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2012. Jane: an advanced freely available hierarchical machine translation toolkit. *Machine Translation*, 26(3):197–216, September.

Richard Zens and Hermann Ney. 2008. Improvements in Dynamic Programming Beam Search for Phrase-Based Statistical Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 195–205, Waikiki, HI, USA, October.

Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. 2004. Reordering Constraints for Phrase-Based Statistical Machine Translation. In *Proc. of the Int. Conf. on Computational Linguistics (COLING)*, pages 205–211, Geneva, Switzerland, August.