Left-to-Right Target Generation for Hierarchical Phrase-based Translation

Taro Watanabe **Hajime Tsukada** **Hideki Isozaki**
2-4, Hikaridai, Seika-cho, Soraku-gun,
Kyoto, JAPAN 619-0237
{taro,tsukada,isozaki}@cslab.kecl.ntt.co.jp

Abstract

We present a hierarchical phrase-based statistical machine translation in which a target sentence is efficiently generated in left-to-right order. The model is a class of synchronous-CFG with a Greibach Normal Form-like structure for the projected production rule: The paired target-side of a production rule takes a phrase prefixed form. The decoder for the target-normalized form is based on an Early-style top down parser on the source side. The target-normalized form coupled with our top down parser implies a left-to-right generation of translations which enables us a straightforward integration with ngram language models. Our model was experimented on a Japanese-to-English newswire translation task, and showed statistically significant performance improvements against a phrase-based translation system.

1 Introduction

In a classical statistical machine translation, a foreign language sentence $f_1^J = f_1, f_2, ... f_J$ is translated into another language, i.e. English, $e_1^I = e_1, e_2, ..., e_I$ by seeking a maximum likely solution of:

$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I} Pr(e_1^I|f_1^J) \quad (1)$$

$$= \operatorname*{argmax}_{e_1^I} Pr(f_1^J|e_1^I)Pr(e_1^I) \quad (2)$$

The source channel approach in Equation 2 independently decomposes translation knowledge into a translation model and a language model, respectively (Brown et al., 1993). The former represents the correspondence between two languages and the latter contributes to the fluency of English.

In the state of the art statistical machine translation, the posterior probability $Pr(e_1^I|f_1^J)$ is directly maximized using a log-linear combination of feature functions (Och and Ney, 2002):

$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I} \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{e'^{I'}_1} \exp\left(\sum_{m=1}^M \lambda_m h_m(e'^{I'}_1, f_1^J)\right)} \quad (3)$$

where $h_m(e_1^I, f_1^J)$ is a feature function, such as a ngram language model or a translation model. When decoding, the denominator is dropped since it depends only on f_1^J. Feature function scaling factors λ_m are optimized based on a maximum likely approach (Och and Ney, 2002) or on a direct error minimization approach (Och, 2003). This modeling allows the integration of various feature functions depending on the scenario of how a translation is constituted.

A phrase-based translation model is one of the modern approaches which exploits a phrase, a contiguous sequence of words, as a unit of translation (Koehn et al., 2003; Zens and Ney, 2003; Tillman, 2004). The idea is based on a word-based source channel modeling of Brown et al. (1993): It assumes that e_1^I is segmented into a sequence of K phrases $\bar{e}_1^K$. Each phrase $\bar{e}_k$ is transformed into $\bar{f}_k$. The translated phrases are reordered to form f_1^J. One of the benefits of the modeling is that the phrase translation unit preserves localized word reordering. However, it cannot hypothesize a long-distance reordering required for linguistically divergent language pairs. For instance, when translating Japanese to English, a Japanese SOV structure has to be reordered to match with an En-

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 777–784,
Sydney, July 2006.

glish SVO structure. Such a sentence-wise movement cannot be realized within the phrase-based modeling.

Chiang (2005) introduced a hierarchical phrase-based translation model that combined the strength of the phrase-based approach and a synchronous-CFG formalism (Aho and Ullman, 1969): A rewrite system initiated from a start symbol which synchronously rewrites paired non-terminals. Their translation model is a binarized synchronous-CFG, or a rank-2 of synchronous-CFG, in which the right-hand side of a production rule contains at most two non-terminals. The form can be regarded as a phrase translation pair with at most two holes instantiated with other phrases. The hierarchically combined phrases provide a sort of reordering constraints that is not directly modeled by a phrase-based model.

Rules are induced from a bilingual corpus without linguistic clues first by extracting phrase translation pairs, and then by generalizing extracted phrases with holes (Chiang, 2005). Even in a phrase-based model, the number of phrases extracted from a bilingual corpus is quadratic to the length of bilingual sentences. The grammar size for the hierarchical phrase-based model will be further exploded, since there exists numerous combination of inserting holes to each rule. The spuriously increasing grammar size will be problematic for decoding without certain heuristics, such as a length based thresholding.

The integration with a ngram language model further increases the cost of decoding especially when incorporating a higher order ngram, such as 5-gram. In the hierarchical phrase-based model (Chiang, 2005), and an inversion transduction grammar (ITG) (Wu, 1997), the problem is resolved by restricting to a binarized form where at most two non-terminals are allowed in the right-hand side. However, Huang et al. (2005) reported that the computational complexity for decoding amounted to $O(J^{3+3(n-1)})$ with n-gram even using a hook technique. The complexity lies in memorizing the ngram's context for each constituent. The order of ngram would be a dominant factor for higher order ngrams.

As an alternative to a binarized form, we present a target-normalized hierarchical phrase-based translation model. The model is a class of a hierarchical phrase-based model, but constrained so that the English part of the right-hand side is restricted to a Greibach Normal Form (GNF)-like structure: A contiguous sequence of terminals, or a phrase, is followed by a string of non-terminals. The target-normalized form reduces the number of rules extracted from a bilingual corpus, but still preserves the strength of the phrase-based approach. An integration with ngram language model is straightforward, since the model generates a translation in left-to-right order. Our decoder is based on an Earley-style top down parsing on the foreign language side. The projected English-side is generated in left-to-right order synchronized with the derivation of the foreign language side. The decoder's implementation is taken after a decoder for an existing phrase-based model with a simple modification to account for production rules. Experimental results on a Japanese-to-English newswire translation task showed significant improvement against a phrase-based modeling.

2 Translation Model

A weighted synchronous-CFG is a rewrite system consisting of production rules whose right-hand side is paired (Aho and Ullman, 1969):

$$X \leftarrow \langle \gamma, \alpha, \sim \rangle \tag{4}$$

where X is a non-terminal, γ and α are strings of terminals and non-terminals. For notational simplicity, we assume that γ and α correspond to the foreign language side and the English side, respectively. $\sim$ is a one-to-one correspondence for the non-terminals appeared in γ and α. Starting from an initial non-terminal, each rule rewrites non-terminals in γ and α that are associated with $\sim$.

Chiang (2005) proposed a hierarchical phrase-based translation model, a binary synchronous-CFG, which restricted the form of production rules as follows:

- Only two types of non-terminals allowed: S and X.
- Both of the strings γ and α must contain at least one terminal item.
- Rules may have at most two non-terminals but non-terminals cannot be adjacent for the foreign language side γ.

The production rules are induced from a bilingual corpus with the help of word alignments. To alleviate a data sparseness problem, glue rules are

added that prefer combining hierarchical phrases in a serial manner:

$$S \rightarrow \left\langle S_{\boxed{1}} X_{\boxed{2}}, S_{\boxed{1}} X_{\boxed{2}} \right\rangle \quad (5)$$

$$S \rightarrow \left\langle X_{\boxed{1}}, X_{\boxed{1}} \right\rangle \quad (6)$$

where boxed indices indicate non-terminal's linkages represented in $\sim$.

Our model is based on Chiang (2005)'s framework, but further restricts the form of production rules so that the aligned right-hand side α follows a GNF-like structure:

$$X \leftarrow \left\langle \gamma, \bar{b}\beta, \sim \right\rangle \quad (7)$$

where $\bar{b}$ is a string of terminals, or a phrase, and *beta* is a (possibly empty) string of non-terminals. The foreign language at right-hand side γ still takes an arbitrary string of terminals and non-terminals. The use of a phrase $\bar{b}$ as a prefix keeps the strength of the phrase-base framework. A contiguous English side coupled with a (possibly) discontiguous foreign language side preserves a phrase-bounded local word reordering. At the same time, the target-normalized framework still combines phrases hierarchically in a restricted manner.

The target-normalized form can be regarded as a type of rule in which certain non-terminals are always instantiated with phrase translation pairs. Thus, we will be able to reduce the number of rules induced from a bilingual corpus, which, in turn, help reducing the decoding complexity.

The contiguous phrase-prefixed form generates English in left-to-right order. Therefore, a decoder can easily hypothesize a derivation tree integrated with a ngram language model even with higher order.

Note that we do not imply arbitrary synchronous-CFGs are transformed into the target normalized form. The form simply restricts the grammar extracted from a bilingual corpus explained in the next section.

2.1 Rule Extraction

We present an algorithm to extract production rules from a bilingual corpus. The procedure is based on those for the hierarchical phrase-based translation model (Chiang, 2005).

First, a bilingual corpus is annotated with word alignments using the method of Koehn et al. (2003). Many-to-many word alignments are induced by running a one-to-many word alignment model, such as GIZA++ (Och and Ney, 2003), in both directions and by combining the results based on a heuristic (Koehn et al., 2003).

Second, phrase translation pairs are extracted from the word alignment corpus (Koehn et al., 2003). The method exhaustively extracts phrase pairs (f_j^{j+m}, e_i^{i+n}) from a sentence pair (f_1^J, e_1^I) that do not violate the word alignment constraints a:

$$\exists (i', j') \in a : j' \in [j, j+m], i' \in [i, i+n]$$
$$\nexists (i', j') \in a : j' \in [j, j+m], i' \notin [i, i+n]$$
$$\nexists (i', j') \in a : j' \notin [j, j+m], i' \in [i, i+n]$$

Third, based on the extracted phrases, production rules are accumulated by computing the "holes" for contiguous phrases (Chiang, 2005):

1. A phrase pair $(\bar{f}, \bar{e})$ constitutes a rule
$$X \rightarrow \left\langle \bar{f}, \bar{e} \right\rangle$$

2. A rule $X \rightarrow \langle \gamma, \alpha \rangle$ and a phrase pair $(\bar{f}, \bar{e})$ s.t. $\gamma = \gamma' \bar{f} \gamma''$ and $\alpha = \bar{e}' \bar{e} \beta$ constitutes a rule
$$X \rightarrow \left\langle \gamma' X_{\boxed{k}} \gamma'', \bar{e}' X_{\boxed{k}} \beta \right\rangle$$

Following Chiang (2005), we applied constraints when inducing rules with non-terminals:

- At least one foreign word must be aligned to an English word.
- Adjacent non-terminals are not allowed for the foreign language side.

2.2 Phrase-based Rules

The rule extraction procedure described in Section 2.1 is a corpus-based, therefore will be easily suffered from a data sparseness problem. The hierarchical phrase-based model avoided this problem by introducing the glue rules 5 and 6 that combined hierarchical phrases sequentially (Chiang, 2005).

We use a different method of generalizing production rules. When production rules without non-terminals are extracted in step 1 of Section 2.1,

$$X \rightarrow \left\langle \bar{f}, \bar{e} \right\rangle \quad (8)$$

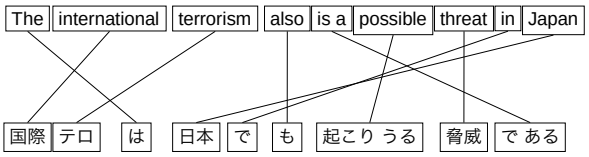
then, we also add production rules as follows:

$$X \rightarrow \left\langle \bar{f} X_{\boxed{1}}, \bar{e} X_{\boxed{1}} \right\rangle \quad (9)$$

$$X \rightarrow \left\langle X_{\boxed{1}} \bar{f}, \bar{e} X_{\boxed{1}} \right\rangle \quad (10)$$

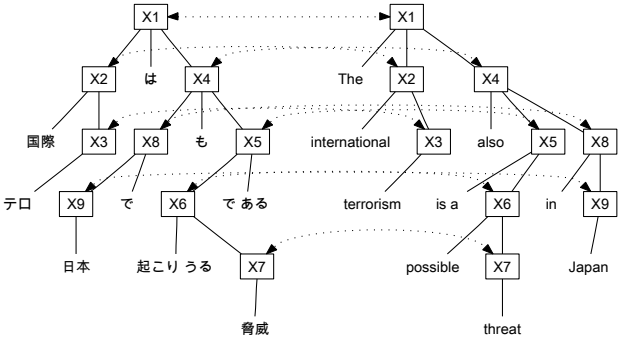
$$X \rightarrow \left\langle X_{\boxed{1}} \bar{f} X_{\boxed{2}}, \bar{e} X_{\boxed{1}} X_{\boxed{2}} \right\rangle \quad (11)$$

$$X \rightarrow \left\langle X_{\boxed{2}} \bar{f} X_{\boxed{1}}, \bar{e} X_{\boxed{1}} X_{\boxed{2}} \right\rangle \quad (12)$$



Reference translation: "International terrorism is a threat even to Japan"

(a) Translation by a phrase-based model.



(b) A derivation tree representation for Figure 1(a).Indices in non-terminal X represent the order to perform rewriting.

Figure 1: An example of Japanese-to-English translation by a phrase-based model.

We call them phrase-based rules, since four types of rules are generalized directly from phrase translation pairs.

The class of rules roughly corresponds to the reordering constraints used in a phrase-based model during decoding. Rules 8 and 9 are sufficient to realize a monotone decoding in which phrase translation pairs are simply combined sequentially. With rules 10 and 11, the non-terminal $X_{\boxed{1}}$ behaves as a place holder where certain number of foreign words are skipped. Therefore, those rules realize a window size constraint used in many phrase-based models (Koehn et al., 2003). The rule 12 further gives an extra freedom for the phrase pair reordering. The rules 8 through 12 can be interpreted as ITG-constraints where phrase translation pairs are hierarchically combined either in a monotonic way or in an inverted manner (Zens and Ney, 2003; Wu, 1997). Thus, by controlling what types of phrase-based rules employed in a grammar, we will be able to simulate a phrase-based translation model with various constraints. This reduction is rather natural in that a finite state transducer, or a phrase-based model, is a subclass of a synchronous-CFG.

Figure 1(a) shows an example Japanese-to-English translation by a phrase-based model described in Section 5. Using the phrase-based rules, the translation results is represented as a derivation tree in Figure 1(b).

3 Decoding

Our decoder is an Earley-style top down parser on the foreign language side with a beam search strategy. Given an input sentence f_1^J, the decoder seeks for the best English according to Equation 3 using the feature functions described in Section 4. The English output sentence is generated in left-to-right order in accordance with the derivation of the foreign language side synchronized with the cardinality of already translated foreign word positions.

The decoding process is very similar to those described in (Koehn et al., 2003): It starts from an initial empty hypothesis. From an existing hypothesis, new hypothesis is generated by consuming a production rule that covers untranslated foreign word positions. The score for the newly generated hypothesis is updated by combining the scores of feature functions described in Section 4. The English side of the rule is simply concatenated to form a new prefix of English sentence. Hypotheses that consumed m foreign words are stored in a priority queue Q_m.

Hypotheses in Q_m undergo two types of pruning: A histogram pruning preserves at most M hypotheses in Q_m. A threshold pruning discards a hypotheses whose score is below the maximum score of Q_m multiplied with a threshold value τ. Rules are constrained by their foreign word span of a non-terminal. For a rule consisting of more than two non-terminals, we constrained so that at least one non-terminal should span at most κ words.

The decoder is characterized as a weighted synchronous-CFG implemented with a push-down automaton rather a weighted finite state transducer (Aho and Ullman, 1969). Each hypothesis maintains following knowledge:

- A prefix of English sentence. For space efficiency, the prefix is represented as a word graph.
- Partial contexts for each feature function. For instance, to compute a 5-gram language model feature, we keep the consecutive last four words of an English prefix.

- A stack that keeps track of the uncovered foreign word spans. The stack for an initial hypothesis is initialized with span $[1, J]$.

When extending a hypothesis, the associated stack structure is popped. The popped foreign word span $[j^l, j^r]$ is used to locate the rules for uncovered foreign word positions. We assume that the decoder accumulates all the applicable rules from a large database and stores the extracted rules in a chart structure. The decoder identifies what rules to consume when extending a hypothesis using the chart structure. A new hypothesis is created with an updated stack by pushing foreign non-terminal spans: For each rule spanning $[j^l, j^r]$ at foreign-side with non-terminal spans of $[k_1^l, k_1^r], [k_2^l, k_2^r], ...$, the non-terminal spans are pushed in the reverse order of the projected English side. For example, A rule with foreign word non-terminal spans:

$$X \rightarrow \left\langle X_{\boxed{2}} : [k_2^l, k_2^r]\ \bar{f}\ X_{\boxed{1}} : [k_1^l, k_1^r], \bar{e}\ X_{\boxed{1}}\ X_{\boxed{2}} \right\rangle$$

will update a stack by pushing the foreign word spans $[k_2^l, k_2^r]$ and $[k_1^l, k_1^r]$ in order. This ordering assures that, when popped, the English-side will be generated in left-to-right order. A hypothesis with an empty stack implies that the hypothesis has covered all the foreign words.

Figure 2 illustrates the decoding process for the derivation tree in Figure 1(b). Starting from the initial hypothesis of $[1, 11]$, the stack is updated in accordance with non-terminal's spans. The span is popped and the rule with the foreign word pan $[1, 11]$ is looked up from the chart structure. The stack structure for the newly created hypothesis is updated by pushing non-terminal spans $[4, 11]$ and $[1, 2]$.

Our decoder is based on an in-house developed phrase-based decoder which uses a bit vector to represent uncovered foreign word positions for each hypothesis. We basically replaced the bit vector structure to the stack structure: Almost no modification was required for the word graph structure and the beam search strategy implemented for a phrase-based modeling. The use of a stack structure directly models a synchronous-CFG formalism realized as a push-down automation, while the bit vector implementation is conceptualized as a finite state transducer. The cost of decoding with the proposed model is cubic to foreign language sentence length.

Rules	Stack
	[1, 11]
$X : [1, 11] \rightarrow \langle X_{\boxed{1}} : [1, 2]$ は $X_{\boxed{2}} : [4, 11]$, The $X_{\boxed{1}}\ X_{\boxed{2}} \rangle$	[1, 2] [4, 11]
$X : [1, 2] \rightarrow \langle$国際 $X_{\boxed{1}} : [2, 2]$, international $X_{\boxed{1}} \rangle$	[2, 2] [4, 11]
$X : [2, 2] \rightarrow \langle$テロ, terrorism$\rangle$	[4, 11]
$X : [4, 11] \rightarrow \langle X_{\boxed{2}} : [4, 5]$ も $X_{\boxed{1}} : [7, 11]$, also $X_{\boxed{1}}\ X_{\boxed{2}} \rangle$	[7, 11] [4, 5]
$X : [7, 11] \rightarrow \langle X_{\boxed{1}} : [7, 9]$ で ある, is a $X_{\boxed{1}} \rangle$	[7, 9] [4, 5]
$X : [7, 9] \rightarrow \langle$起こり うる $X_{\boxed{1}} : [9, 9]$, possible $X_{\boxed{1}} \rangle$	[9, 9] [4, 5]
$X : [9, 9] \rightarrow \langle$脅威, threat$\rangle$	[4, 5]
$X : [4, 5] \rightarrow \langle X_{\boxed{1}} : [4, 4]$ で, in $X_{\boxed{1}} \rangle$	[4, 4]
$X : [4, 4] \rightarrow \langle$日本, Japan$\rangle$	

Figure 2: An example decoding process of Figure 1(b) with a stack to keep track of foreign word spans.

4 Feature Functions

The decoder for our translation model uses a log-linear combination of feature functions, or sub-models, to seek for the maximum likely translation according to Equation 3. This section describes the models experimented in Section 5, mainly consisting of count-based models, lexicon-based models, a language model, reordering models and length-based models.

4.1 Count-based Models

Main feature functions $h_\phi(f_1^J|e_1^I, \mathcal{D})$ and $h_\phi(e_1^I|f_1^J, \mathcal{D})$ estimate the likelihood of two sentences f_1^J and e_1^I over a derivation tree $\mathcal{D}$. We assume that the production rules in $\mathcal{D}$ are independent of each other:

$$h_\phi(f_1^J|e_1^I, \mathcal{D}) = \log \prod_{\langle\gamma,\alpha\rangle\in\mathcal{D}} \phi(\gamma|\alpha) \qquad (13)$$

$\phi(\gamma|\alpha)$ is estimated through the relative frequency on a given bilingual corpus.

$$\phi(\gamma|\alpha) = \frac{count(\gamma, \alpha)}{\sum_\gamma count(\gamma, \alpha)} \qquad (14)$$

where $count(\cdot)$ represents the cooccurrence frequency of rules γ and α.

The relative count-based probabilities for the phrase-based rules are simply adopted from the original probabilities of phrase translation pairs.

4.2 Lexicon-based Models

We define lexically weighted feature functions $h_w(f_1^J|e_1^I, \mathcal{D})$ and $h_w(e_1^I|f_1^J, \mathcal{D})$ applying the independence assumption of production rules as in

Equation 13.

$$h_w(f_1^J|e_1^I, \mathcal{D}) = \log \prod_{\langle\gamma,\alpha\rangle\in\mathcal{D}} p_w(\gamma|\alpha) \quad (15)$$

The lexical weight $p_w(\gamma|\alpha)$ is computed from word alignments a inside γ and α (Koehn et al., 2003):

$$p_w(\gamma|\alpha, a) = \prod_{i=1}^{|\alpha|} \frac{1}{|\{j|(i,j)\in a\}|} \sum_{\forall(i,j)\in a} t(\gamma_j|\alpha_i) \quad (16)$$

where $t(\cdot)$ is a lexicon model trained from the word alignment annotated bilingual corpus discussed in Section 2.1. The alignment a also includes non-terminal correspondence with $t(X_{\boxed{k}}|X_{\boxed{k}}) = 1$. If we observed multiple alignment instances for γ and α, then, we take the maximum of the weights.

$$p_w(\gamma|\alpha) = \max_a p_w(\gamma|\alpha, a) \quad (17)$$

4.3 Language Model

We used mixed-cased n-gram language model. In case of 5-gram language model, the feature function is expressed as follows:

$$h_{lm}(e_1^I) = \log \prod_i p_n(e_i|e_{i-4}e_{i-3}e_{i-2}e_{i-1}) \quad (18)$$

4.4 Reordering Models

In order to limit the reorderings, two feature functions are employed based on the backtracking of rules during the top-down parsing on foreign language side.

$$h_h(e_1^I, f_1^J, \mathcal{D}) = \sum_{\mathcal{D}_i\in back(\mathcal{D})} height(\mathcal{D}_i) \quad (19)$$

$$h_w(e_1^I, f_1^J, \mathcal{D}) = \sum_{\mathcal{D}_i\in back(\mathcal{D})} width(\mathcal{D}_i) \quad (20)$$

where $back(\mathcal{D})$ is a set of subtrees backtracked during the derivation of $\mathcal{D}$, and $height(\mathcal{D}_i)$ and $width(\mathcal{D}_i)$ refer the height and width of subtree $\mathcal{D}_i$, respectively. In Figure 1(b), for instance, a rule of $X_{\boxed{1}}$ with non-terminals $X_{\boxed{2}}$ and $X_{\boxed{4}}$, two rules $X_{\boxed{2}}$ and $X_{\boxed{3}}$ spanning two terminal symbols should be backtracked to proceed to $X_{\boxed{4}}$. The rationale is that positive scaling factors prefer a deeper structure whereby negative scaling factors prefer a monotonized structure.

4.5 Length-based Models

Three trivial length-based feature functions were used in our experiment.

$$h_l(e_1^I) = I \quad (21)$$

$$h_r(\mathcal{D}) = rule(\mathcal{D}) \quad (22)$$

$$h_p(\mathcal{D}) = phrase(\mathcal{D}) \quad (23)$$

Table 1: Japanese/English news corpus

		Japanese	English
train	sentence	175,384	
	dictionary	+ 1,329,519	
	words	8,373,478	7,222,726
	vocabulary	297,646	397,592
dev.	sentence	1,500	
	words	47,081	39,117
	OOV	45	149
test	sentence	1,500	
	words	47,033	38,707
	OOV	51	127

Table 2: Phrases/rules extracted from the Japanese/English bilingual corpus. Figures do not include phrase-based rules.

	# rules/phrases
Phrase	5,433,091
Normalized-2	6,225,630
Normalized-3	6,233,294
Hierarchical	12,824,387

where $rule(\mathcal{D})$ and $phrase(\mathcal{D})$ are the number of production rules extracted in Section 2.1 and phrase-based rules generalized in Section 2.2, respectively. The English length feature function controls the length of output sentence. Two feature functions based on rule's counts are hypothesized to control whether to incorporate a production rule or a phrase-based rule into $\mathcal{D}$.

5 Experiments

The bilingual corpus used for our experiments was obtained from an automatically sentence aligned Japanese/English Yomiuri newspaper corpus consisting of 180K sentence pairs (refer to Table 1) (Utiyama and Isahara, 2003). From one-to-one aligned sentences, 1,500 sentence pairs were sampled for a development set and a test set[1]. Since the bilingual corpus is rather small, especially for the newspaper translation domain, Japanese/English dictionaries consisting of 1.3M entries were added into a training set to alleviate an OOV problem[2].

Word alignments were annotated by a HMM translation model (Och and Ney, 2003). After

[1] Japanese sentences were segmented by MeCab available from `http://mecab.sourceforge.jp`.

[2] The dictionary entries were compiled from JEDICT/JNAMEDICT and an in-house developed dictionary.

the annotation via Viterbi alignments with refinements, phrases translation pairs and production rules were extracted (refer to Table 2). We performed the rule extraction using the hierarchical phrase-based constraint (Hierarchical) and our proposed target-normalized form with 2 and 3 non-terminals (Normalized-2 and Normalized-3). Phrase translation pairs were also extracted for comparison (Phrase). We did not threshold the extracted phrases or rules by their length. Table 2 shows that Normalized-2 extracted slightly larger number of rules than those for phrase-based model. Including three non-terminals did not increase the grammar size. The hierarchical phrase-based translation model extracts twice as large as our target-normalized formalism. The target-normalized form is restrictive in that non-terminals should be consecutive for the English-side. This property prohibits spuriously extracted production rules.

Mixed-casing 3-gram/5-gram language models were estimated from LDC English GigaWord 2 together with the 100K English articles of Yomiuri newspaper that were used neither for development nor test sets [3].

We run the decoder for the target-normalized hierarchical phrase-based model consisting of at most two non-terminals, since adding rules with three non-terminals did not increase the grammar size. ITG-constraint simulated phrase-based rules were also included into our grammar. The foreign word span size was thresholded so that at least one non-terminal should span at most 7 words.

Our phrase-based model employed all feature functions for the hierarchical phrase-based system with additional feature functions:

- A distortion model that penalizes the reordering of phrases by the number of words skipped $|j - (j' + m') - 1|$, where j is the foreign word position for a phrase f_j^{j+m} translated immediately after a phrase for $f_{j'}^{j'+m'}$ (Koehn et al., 2003).
- Lexicalized reordering models constrain the reordering of phrases whether to favor monotone, swap or discontinuous positions (Tillman, 2004).

The phrase-based decoder's reordering was constrained by ITG-constraints with a window size of 7.

[3] We used SRI ngram language modeling toolkit with limited vocabulary size.

Table 3: Results for the Japanese-to-English newswire translation task.

		BLEU [%]	NIST
Phrase	3-gram	7.14	3.21
	5-gram	7.33	3.19
Normalized-2	3-gram	10.00	4.11
	5-gram	10.26	4.20

The translation results are summarized in Table 3. Two systems were contrasted by 3-gram and 5-gram language models. Results were evaluated by ngram precision based metrics, BLEU and NIST, on the casing preserved single reference test set. Feature function scaling factors for each system were optimized on BLEU score under the development set using a downhill simplex method. The differences of translation qualities are statistically significant at the 95% confidence level (Koehn, 2004). Although the figures presented in Table 3 are rather low, we found that Normalized-2 resulted in statistically significant improvement over Phrase. Figure 3 shows some translation results from the test set.

6 Conclusion

The target-normalized hierarchical phrase-based model is based on a more general hierarchical phrase-based model (Chiang, 2005). The hierarchically combined phrases can be regarded as an instance of phrase-based model with a place holder to constraint reordering. Such reordering was realized either by an additional constraint for decoding, such as window constraints, IBM constraints or ITG-constraints (Zens and Ney, 2003), or by lexicalized reordering feature functions (Tillman, 2004). In the hierarchical phrase-based model, such reordering is explicitly represented in each rule.

As experimented in Section 5, the use of the target-normalized form reduced the grammar size, but still outperformed a phrase-based system. Furthermore, the target-normalized form coupled with our top down parsing on the foreign language side allows an easier integration with ngram language model. A decoder can be implemented based on a phrase-based model by employing a stack structure to keep track of untranslated foreign word spans.

The target-normalized form can be interpreted

Reference:	Japan needs to learn a lesson from history to ensure that it not repeat its mistakes .
Phrase:	At the same time , it never mistakes that it is necessary to learn lessons from the history of criminal .
Normalized-2:	It is necessary to learn lessons from history so as not to repeat similar mistakes in the future .
Reference:	The ministries will dispatch design and construction experts to China to train local engineers and to research technology that is appropriate to China's economic situation .
Phrase:	Japan sent specialists to train local technicians to the project , in addition to the situation in China and its design methods by exception of study .
Normalized-2:	Japan will send experts to study the situation in China , and train Chinese engineers , construction design and construction methods of the recipient from .
Reference:	The Health and Welfare Ministry has decided to invoke the Disaster Relief Law in extending relief measures to the village and the city of Niigata .
Phrase:	The Health and Welfare Ministry in that the Japanese people in the village are made law .
Normalized-2:	The Health and Welfare Ministry decided to apply the Disaster Relief Law to the village in Niigata .

Figure 3: Sample translations from two systems: Phrase and Normalized-2

as a set of rules that reorders the foreign language to match with English language sequentially. Collins et al. (2005) presented a method with hand-coded rules. Our method directly learns such serialization rules from a bilingual corpus without linguistic clues.

The translation quality presented in Section 5 are rather low due to the limited size of the bilingual corpus, and also because of the linguistic difference of two languages. As our future work, we are in the process of experimenting our model for other languages with rich resources, such as Chinese and Arabic, as well as similar language pairs, such as French and English. Additional feature functions will be also investigated that were proved successful for phrase-based models together with feature functions useful for a tree-based modeling.

Acknowledgement

We would like to thank to our colleagues, especially to Hideto Kazawa and Jun Suzuki, for useful discussions on the hierarchical phrase-based translation.

References

Alfred V. Aho and Jeffrey D. Ullman. 1969. Syntax directed translations and the pushdown assembler. *J. Comput. Syst. Sci.*, 3(1):37–56.

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.

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

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proc. of ACL 2005*, pages 531–540, Ann Arbor, Michigan, June.

Liang Huang, Hao Zhang, and Daniel Gildea. 2005. Machine translation as lexicalized parsing with hooks. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 65–73, Vancouver, British Columbia, October.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of NAACL 2003*, pages 48–54, Edmonton, Canada.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of EMNLP 2004*, pages 388–395, Barcelona, Spain, July.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of ACL 2002*, pages 295–302.

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 in statistical machine translation. In *Proc. of ACL 2003*, pages 160–167.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In *HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts, USA, May 2 - May 7.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *Proc. of ACL 2003*, pages 72–79.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.*, 23(3):377–403.

Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *Proc. of ACL 2003*, pages 144–151.