Long-Distance Hierarchical Structure Transformation Rules Utilizing Function Words

Chenchen Ding, Takashi Inui, Mikio Yamamoto

Department of Computer Science, University of Tsukuba
1-1-1 Tennodai, Tsukuba, 305-8573, Japan
{tei@mibel., inui, myama@}cs.tsukuba.ac.jp

Abstract

In this paper, we propose structure transformation rules for statistical machine translation which are lexicalized by only function words. Although such rules can be extracted from an aligned parallel corpus simply as original phrase pairs, their structure is hierarchical and thus can be used in a hierarchical translation system. In addition, structure transformation rules can take into account long-distance reordering, allowing for more than two phrases to be moved simultaneously. The rule set is used as a core module in our hierarchical model together with two other modules, namely, a basic reordering module and an optional gap phrase module. Our model is considerably more compact and produces slightly higher BLEU scores than the original hierarchical phrase-based model in Japanese-English translation on the parallel corpus of the NTCIR-7 patent translation task.

1. Introduction

The task of reordering words and phrases in a source language to match the structure of a target language is one of the most challenging problems in statistical machine translation (SMT). In particular, when there are considerable differences between the source language and the target language in terms of syntactic structure, such as in the case of Japanese and English, global reordering is necessary but difficult to achieve.

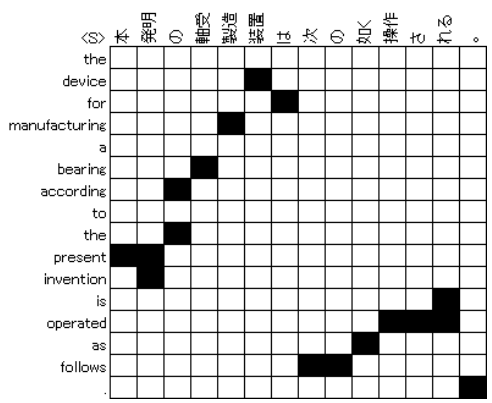

Figure 1: *A Japanese-English word alignment table*

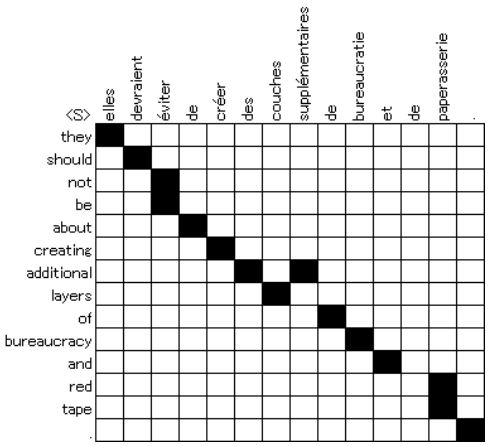

Figure 2: *A French-English word alignment table*

Figs. 1 and 2 show the respective word alignment tables for a Japanese-English sentence pair from the NTCIR-7 parallel corpus and a French-English sentence pair from the Europarl corpus. Both tables are generated automatically by the word aligner GIZA++ [1][2] with grow-diag-final-and symmetrization heuristics [3]. It is clear that the aligned words (black cells) lie approximately along the main diagonal of the table in Fig. 2, which indicates the similarity in word and phrase order between French and English, in contrast to Fig. 1, in which the scattering of aligned words away from the main diagonal reveal drastic differences in word and phrase order between Japanese and English.

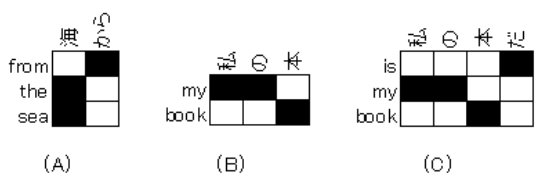

Figure 3: *Several reordering patterns*

In phrase-based models, the monotone/swap orientation of a phrase pair with respect to its adjacent cell is modeled with a lexicalized reordering model [4]. As Fig. 3 (A) shows, when the phrase pair $< から\ ,\ from >$ is extracted from

the training corpus, it can also be seen that swap orientation exists between this phrase pair and its adjacent cell. Thus, this information can be quantified by a lexicalized reordering model and used in decoding. Also, in Fig. 3 (B), the phrase pair $<$ 私 の , my $>$ maintains the monotone orientation with respect to its adjacent cell. This approach will fail to extract reordering information in more complex alignment patterns, such as that in Fig. 3 (C), from which the swap orientation of the phrase pair $<$ だ , is $>$ cannot be extracted unless the rule generation algorithm is aware that the pair $<$ 私 の 本 , $my\ book$ $>$ is integral [5].

As shown in Fig. 1, local monotone/swap reordering is insufficient in Japanese-English translation. In this regard, hierarchical phrase-based models [6] can provide higher performance through SCFG-style hierarchical rules. From the alignment in Fig. 3 (C), Rules (1) and (2) can be extracted, where Rule (1) indicates that the copula (だ) after a noun phrase on the Japanese side appears before the corresponding noun phrase on the English side, and Rule (2) can translate the discontinuous part on the Japanese side directly.

$$X \rightarrow < X_1 \text{ だ} , \ is\ X_1 > \tag{1}$$

$$X \rightarrow < \text{私 の } X_1 \text{ だ} , \ is\ my\ X_1 > \tag{2}$$

Although hierarchical phrase-based models are powerful, they usually become excessively large since the original approach is to extract all possible rules. A common strategy in this case is to apply several heuristics in order to reduce the model size. Therefore, the original hierarchical phrases are constrained to form a 2-SCFG; in other words, there can be at most two nonterminals in a hierarchical phrase pair. However, this is not always sufficient for long-distance reordering.

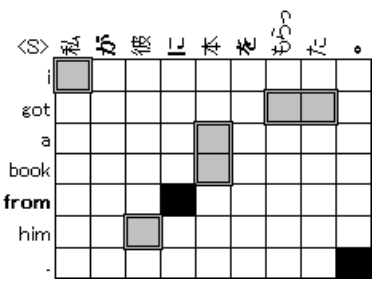

Figure 4: *Structure transformation with multi-nonterminal*

As illustrated by the example shown in Fig. 4, in Japanese, the postpositions が, に and を in a sentence or a clause are markers of the subject, indirect object and direct object, respectively, and this three-tuple structural information cannot be captured by a single 2-SCFG rule.

A shortcoming of the original hierarchical phrase-based model is that by performing indiscriminate extraction, useful rules are always mixed with ineffective rules, and when using heuristics to reduce the model size, the ineffective rules are discarded together with a portion of the useful rules.

In this paper, we propose a method for extracting structure transformation rules. The extracted rule set can capture the exact structural information, as in Rule (3) from the example in Fig. 4, while remaining rather compact.

$$\begin{aligned} X \rightarrow <& X_1 \text{ が } X_2 \text{ に } X_3 \text{ を } X_4 , \\ & X_1\ X_4\ X_3\ from\ X_2 > \end{aligned} \tag{3}$$

In the following section, we describe the procedure for extraction and estimation of structure transformation rules in detail, and in Section 3 we describe the other two modules used in our hierarchical model. Furthermore, in Section 4 we compare the performance of our model with that of the original phrase-based model and the hierarchical phrase-based model by using BLEU [7] as a metric of translation quality on the Japanese-English parallel corpus of the NTCIR-7 patent translation task [8]. Finally, in Section 5 we discuss the properties of our model on the basis of the experimentation results, and in Section 6 we present the direction of future work.

2. Structure Transformation Rules

2.1. Rule Extraction

In attempting to extract long-distance reordering rules, indiscriminate extraction, namely, the extraction of rules lexicalized by all the words in the vocabulary of the foreign language, produces results that are redundant and unable to reveal the true sentence structure. Therefore, we focus on a word set $\mathcal{I}$ of the foreign language vocabulary which can suggest the structure of a sentence (e.g., function words), and extract rules lexicalized only by these words from $\mathcal{I}$, thus obtaining a more precise sentence structure.

Another compelling question is how to determine $\mathcal{I}$ automatically. However, this issue lies outside the scope of this paper, and we limit ourselves to the function word set of Japanese as obtained with the tagging system of the Japanese Morphological Analyzer ChaSen [9]. Therefore, $\mathcal{I}$ is composed of postpositions, conjunctions, auxiliary verbs and punctuation marks occurring in Japanese. We attempt to avoid the utilization of full grammatical information in order to generalize our approach. Such information is taken into account only in the process of determining $\mathcal{I}$, and in all other cases it is no longer needed for the arguments presented in this paper.

The rules are extracted from an aligned parallel corpus, which can be obtained automatically with a word aligner such as GIZA++ and several symmetrization heuristics. If a set $\mathcal{S}$ is defined as all the possible index pairs corresponding to a sentence pair e_1^I,f_1^J, such that

$$\mathcal{S} = \{(i,j) \mid 1 \leq i \leq I, 1 \leq j \leq J\} \tag{4}$$

then a subset $\mathcal{A}$ of $\mathcal{S}$ can be defined by word alignment of e_1^I and f_1^J.

$$\mathcal{A} = \{(i,j) \mid (i,j) \in \mathcal{S}, and\ e_i\ is\ aligned\ to\ f_j\} \tag{5}$$

The original hierarchical phrases can be extracted from e_1^I and f_1^J with $\mathcal{A}$.

In our approach, once $\mathcal{I}$ has been determined, $\mathcal{A}$ can be partitioned into two subsets $\mathcal{R}$ and $\mathcal{O}$ as follows.

$$\mathcal{R} = \{(i,j) \mid (i,j) \in \mathcal{A}, and\ f_j \in \mathcal{I}\} \tag{6}$$

$$\mathcal{O} = \{(i,j) \mid (i,j) \in \mathcal{A}, and\ f_j \notin \mathcal{I}\} \tag{7}$$

It is clear that $\mathcal{R} \cup \mathcal{O} = \mathcal{A}$ and $\mathcal{R} \cap \mathcal{O} = \varnothing$. Considering the example in Fig. 1, in Fig. 5 $\mathcal{R}$ corresponds to black cells and $\mathcal{O}$ corresponds to gray cells, with $\mathcal{I}$ composed of the function words described above.

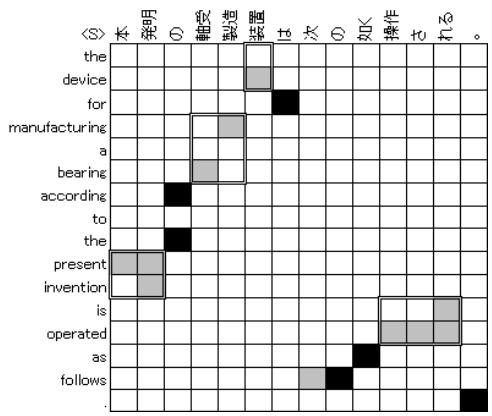

Figure 5: *The set of $\mathcal{R}$, $\mathcal{O}$ and $\mathcal{M}$*

We define $\mathcal{P}$ as the set of all phrase pairs which can be extracted from a sentence pair e_1^I,f_1^J with $\mathcal{A}$, after which a set $\mathcal{P}_o$ is extracted as a subset of $\mathcal{P}$ composed of those phrase pairs which do not cover $(i,j) \in \mathcal{R}$.

$$\begin{aligned}\mathcal{P}_o = \{p \mid p \in \mathcal{P},\ and\ \forall (i,j)\ covered\ by\ p : \\ (i,j) \in \mathcal{A} \rightarrow (i,j) \in \mathcal{O}\}\end{aligned} \tag{8}$$

Next, a subset $\mathcal{M}$ of $\mathcal{P}_o$ is derived as follows.

$$\begin{aligned}\mathcal{M} = \{p \mid p \in \mathcal{P}_o,\ and\ \exists (i,j)\ covered\ by\ p : \\ (i,j)\ is\ covered\ by\ q \in \mathcal{P}_o \rightarrow q = p\}\end{aligned} \tag{9}$$

As a result, $\mathcal{M}$ contains only the maximum phrase pairs from $\mathcal{P}_o$. In Fig. 5, the phrase pairs of $\mathcal{M}$ are enclosed in double-lined boxes. It should be noted that $\mathcal{P}_o$ (and thus $\mathcal{M}$) is not always capable of covering all $(i,j) \in \mathcal{O}$. As shown in Fig. 5, the Japanese word 次 is a noun translated together with の as the English word *follows* (literally *next*), and therefore the cell $< 次,\ follows >$ cannot be covered by any $p \in \mathcal{P}_o$.

Finally, each $p \in \mathcal{M}$ is compacted to a cell and denoted by an aligned nonterminal pair $< X,\ X >$. Therefore, the table in Fig. 5 is transformed into the table in Fig. 6, where each double-lined box in Fig. 5 is compacted to a single black cell and corresponding word sequences from the original sentences on both sides are replaced by the nonterminal mark X. After the compacting process, a new sentence pair

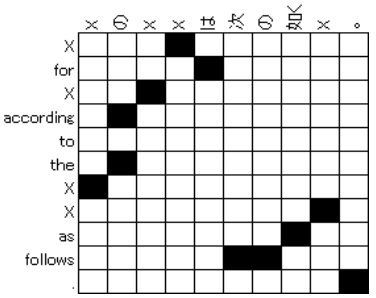

Figure 6: *Compacting the original alignment table*

$\tilde{e}_1^{\tilde{I}}$, $\tilde{f}_1^{\tilde{J}}$ and a new $\tilde{\mathcal{A}}$ can be obtained, after which phrase pairs with nonterminals (i.e., hierarchical phrase pairs), can be extracted in the same manner as ordinary phrase pairs.

Nevertheless, several heuristics are still required, where

- At least one nonterminal must exist on both sides of extracted phrase pairs.
- Aligned words (excluding the nonterminal mark X) must exist in phrase pairs.
- There is a maximum number of words on one or both sides.
- Adjacent nonterminals are not allowed on the foreign language side.

As there is no limit to the number of nonterminals in a rule, the only constraint is the limit on the number of words combined with the constraint that there are no adjacent nonterminals on the foreign language side.

2.2. Parameter Estimation

Regarding the parameters of the rules [10], similarly to the original phrase-based model and the hierarchical phrase-based model,

- the phrase translation probabilities for both directions
- the lexical translation probabilities for both directions
- a constant rule penalty $\exp(1)$
- a constant word penalty $\exp(1)$

are used in our rule set and the other two modules, which will be described in the following section.

In contrast to the original hierarchical phrase-based model, our approach does not make use of an initial phrase pair, and therefore the extracted phrase pairs are simply counted and the phrase translation probabilities are calculated on the basis of the relative frequency in the same manner as in the phrase-based model. Furthermore, if there is more than one pattern, the lexical translation probabilities of a phrase pair adopt the probabilities from the most frequently

occurring alignment pattern. However, if there is still more than one pattern with the same highest frequency of occurrence, the highest probabilities of the two sides are considered separately.

3. A Multi-Module Model

In our hierarchical translation model, the structure transformation rule set is the core module. However, since the structure transformation rules alone cannot construct a complete translation model, two additional simple modules are also incorporated into the proposed model.

3.1. Phrase Reordering Module

The phrase pair set of the original phrase-based model is used in our model in the following hierarchical phrase style

- $X \rightarrow <$ ア , $A >$,

where ア is a Japanese phrase and A is an English phrase.

For each original phrase pair, the hierarchical reordering rules of four patterns (not all of which are necessarily available) are further extracted as follows:

- $X \rightarrow <$ ア X_1 , $A\ X_1 >$
- $X \rightarrow <$ ア X_1 , $X_1\ A >$
- $X \rightarrow <$ X_1 ア , $A\ X_1 >$
- $X \rightarrow <$ X_1 ア , $X_1\ A >$

This module can be regarded as a transformation of the lexicalized reordering model into hierarchical phrase style. Watanabe et al. [11] proposed similar rules as part of their model, where all monotone/swap rules share the same parameters with the original phrase pairs. Therefore, the rules are simply phrase pairs with appended nonterminals and do not contain quantified information about the reordering. In contrast, our model extracts the monotone/swap hierarchical rules of the original phrase pairs and estimates the phrase translation probabilities separately. The procedure starts from a phrase pair and searches the adjacent cells or phrase pairs. Three approaches are considered for the extraction of the reordering model (a word-based [4], a phrase-based [5] and a hierarchical phrase-based approach [5]) in order to obtain the reordering information.

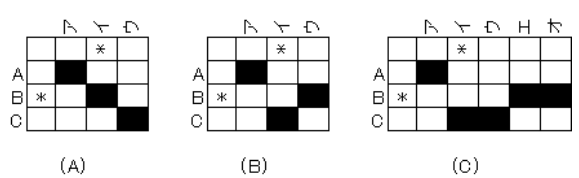

Figure 7: *Reordering rule extraction*

In Fig. 7 (A), the right-down cell of the phrase pair $<$ ア , $A >$ is black, while the right-up and left-down cells (marked with *) are white, therefore Rule (10) can be extracted.

$$X \rightarrow <\text{ア}\ X_1\ ,\ A\ X_1 > \quad (10)$$

If only the adjacent cells are considered, as in the word-based extraction described above, the procedure would fail to extract Rule (10) in Fig. 7 (B), where the right-down cell of the phrase pair $<$ ア , $A >$ is white, even though there is a phrase pair at the right-down of $<$ ア , $A >$. The phrase-based extraction procedure can eliminate this issue by searching the adjacent phrase pairs instead of the adjacent cells. Hierarchical phrase-based extraction is essentially the same as phrase-based extraction, with the exception that during the search, the size of the expected adjacent phrase pairs is unlimited. Usually, the limitation on the size of adjacent phrase pair during the search is the same as that in the phrase pair extraction. For example, in Fig. 7 (C), if the length of the extracted phrases on the foreign language side is set to less than 3 words, then both phrase-based extraction and word-based extraction would fail, even though there is a relatively large phrase pair at the right-down of the phrase pair $<$ ア , $A >$. Lastly, it becomes unnecessary to check cells marked with * in phrase-based and hierarchical phrase-based extraction due to the consistency of the adjacent phrase pair.

Although any of the three extraction approaches can be applied, we do not use two or all of them simultaneously.

The parameters of this module and the procedure for their estimation are the same as those described in Section 2.2.

3.2. Gap Phrase Module

Considering that the structure transformation rules might lack the ability to account for local discontinuous phrase pairs, we also use a gap phrase module as an optional module in our model. This module is composed of rules about discontinuous parts on one or both sides.

The following five patterns exist for rules with a single nonterminal:

- $X \rightarrow <$ ア X_1 イ , $A\ X_1 >$
- $X \rightarrow <$ ア X_1 イ , $X_1\ A >$
- $X \rightarrow <$ ア X_1 , $A\ X_1\ B >$
- $X \rightarrow <$ X_1 ア , $A\ X_1\ B >$
- $X \rightarrow <$ ア X_1 イ , $A\ X_1\ B >$

We use these rules containing a single nonterminal to construct the module. Since the extraction of the rules follows the same procedure as for the original hierarchical phrases, initial phrases are necessary.

The count of the extracted phrase pairs becomes a problem in this case. The fractional count from a uniform distribution was obtained for all extracted phrases sharing the same initial phrase in parallel with a simple incremental integer count. Since both counts resulted in equal performance

we adopted the simple integer count in our experiment. The parameters of this module and the procedure for their calculation are as same as those described in Section 2.2.

4. Experiment

4.1. Decoding with the Multi-Module Model

Since the multi-module model consists of N-SCFG rules, all the rules are organized in the hierarchical phrase style, and a CKY++ decoder as implemented in Moses [12] is used.

When there are no referable rules during the decoding, we use the following glue rules, which are also used in the original hierarchical phrase-based model.

- $S \rightarrow < S\ X\ ,\ S\ X >$
- $S \rightarrow < X\ ,\ X >$

A separate penalty $\exp\ (1)$ is assigned to these rules, and all other rules in the proposed model, even from different modules, share the same rule penalty $\exp(1)$ [1].

4.2. Corpus and Model Settings

Due to the large size of the original hierarchical phrase-based model, only part of the parallel corpus of NTCIR-7 was used as training data amounting to 100 000 sentence pairs selected at random from among the 1.8 million sentence pairs in the corpus. 1842 sentences pairs were used for MERT [13] as a development set, and 1381 sentence pairs were used as the test set in accordance with the sets of the NTCIR-7 translation task.

GIZA++ was used to obtain word alignments with the default settings of Moses together with the grow-diag-final-and symmetrization heuristics, which can achieve the highest performance in Japanese-English translation. The rules for the proposed model and the original models were extracted from the same aligned corpus.

Regarding the extraction heuristics, Moses was used to extract the original phrase translation model and hierarchical phrase translation model. In the phrase model, the max-phrase-length option was set to 5 with the msd-bidirectional-fe reordering, and in the hierarchical phrase model we set MaxSpan to 15, MaxSymbolsSource to 5, and both MinHoleSource and MinHoleTarget to 1. In our model, the structure transformation rule set used an initial word set composed of the function words in Japanese as described in Section 2.1, with at most five words on the Japanese side and no limitation on the number of words on the English side. For the reordering module, the original phrases were extracted with the maximum length for the Japanese side set to 5 and no constraints on the English side below the upper bound of 15. Also, the heuristics of the gap phrase model were the same as those used in the original hierarchical phrase model, with the exception that the number of nonterminals in the original model was set to be between 0 and 2, while our model contained only the set of rules with a single nonterminal with discontinuous parts and an integer count of the phrase pairs.

Finally, a 5-gram language model was built with SRILM [14] by applying the interpolated modified Kneser-Ney method [15] using the English side of all 1.8 million sentence pairs in the corpus.

4.3. Results

Table 1: *BLEU scores for the test set of the original phrase-based model (Ph) with different reordering rules (reo.w(ord-based)/reo.ph(rase-based)/reo.h(ierarchical)) and the hierarchical phrase-based model (Hier)*

$dl/span$	20	30
$Ph(\_reo.w)$	27.34	$\mathbf{27.38}$
$Ph\_reo.ph$	$27.41 \not> Ph$	$27.61 \not> Ph$
$Ph\_reo.h$	$27.00 \not< Ph$	$\mathbf{27.76} > Ph$
$Hier$	$\mathbf{28.61} \gg Ph$	$28.49 \gg Ph$

Table 2: *BLEU scores for the test set of the proposed model composed of R(eo), the reordering module with different extraction rules (*.w/*.ph/*.h), and G(ap), the gap phrase module, and S(tr), the structure transformation rule set*

$dl/span$	20	30
$Reo.w$	$26.96 \not< Ph, \ll H$	$26.98 \not< Ph, \ll H$
$Reo.ph$	$27.04 \not< Ph, \ll H$	$27.33 \not< Ph, \ll H$
$Reo.h$	$27.36 \not> Ph, \ll H$	$\mathbf{27.37} \not< Ph, \ll H$
$GapR.ph$	$27.44 \not> Ph, \ll H$	$27.67 \not> Ph, \ll H$
$GapR.h$	$28.05 \gg Ph, < H$	$\mathbf{28.08} \gg Ph, < H$
$StrR.ph$	$27.75 \not> Ph, \ll H$	$28.01 \gg Ph, < H$
$StrR.h$	$\mathbf{28.47} \gg Ph, \not< H$	$28.22 \gg Ph, \not< H$
$SGR.ph$	$28.54 \gg Ph, \not< H$	$\mathbf{28.91} \gg Ph, > H$
$SGR.h$	$28.79 \gg Ph, \not> H$	$28.73 \gg Ph, \not> H$

In the experiment, both the original models and our model were tested on relatively large reordering settings in decoding, with a distortion limit of 20/30 for the phrase-based model and a span of 20/30 for the hierarchical models. All models were tuned with MERT by using the distortion limit for phrase-based models and the span for hierarchical models separately. The translation performance was evaluated with BLEU [2]. Furthermore, we used bootstrap resam-

[1] In our experiment, assigning different rule penalties to rules of different modules does not lead to any notable improvement in the translation performance, while it increases the risk of morbidity weights due to unstable minimum error rate training (MERT).

[2] According to [8], The highest single-reference BLEU (SRB) score achieved by any SMT system in J-E intrinsic evaluation in NTCIR-7 is 27.20, achieved by the NTT Group, and the SRB score of Moses, which is produced by the organizers, was 27.14 in 2008.

Table 3: *a translation example*

$Source$	また 、 同時に 強 誘電 体 キャパシタ １ ２ から BL ２ へ 電荷 が 移動する 。
$Reference$	at the same time , electric charges transfer from the ferroelectric capacitor 12 to bl2 .
$Hier$	at the same time , bl2 to charge is moved from the ferroelectric capacitor 12 .
$SGR.h$	at the same time , the charge moves from the ferroelectric capacitor 12 to bl2 .

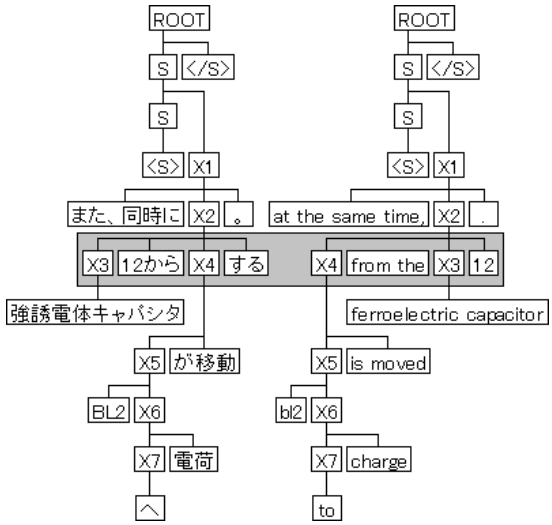

Figure 8: *Derivation of the original 2-SCFG in Hier*

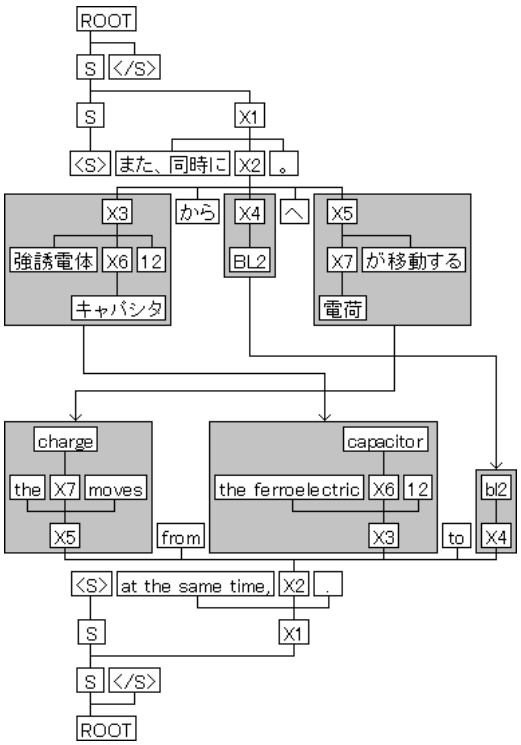

Figure 9: *Structure transformation in SGR*

pling [16] to examine the statistical significance of the obtained results. In Table 1 and Table 2, $\gg$ ($\ll$) indicates that the model on the left side outperforms (is outperformed by) the model on the right at the $p \leq 0.01$ p-level, $>$ ($<$) indicates the same for the $p \leq 0.05$ p-level, and $\not> $ ($\not<$) indicates no statistical significance. The comparison was performed by considering the results for models with the same distortion limit or span.

In Table 1, the results for the phrase translation model with word-based estimated reordering ($reo.w$) are regarded as a baseline and compared with the results for the same phrase translation model with phrase-based ($reo.ph$) and hierarchical phrase-based ($reo.h$) reordering. It can be inferred that although both models provide a certain level of improvement, the difference from word-based reordering is not always statistically significant. In the last row of Table 1, the results for the original hierarchical phrase-based model clearly show that it outperformed the original phrase-based model in our Japanese-English translation experiment.

Table 2 shows the results for our model with some or all proposed modules. The results for Reo indicate that the performance of the basic (reordering) module is the same as that of the original phrase-based model when using phrase-based ($.ph$) and hierarchical ($.h$) reordering, although its performance with word-based ($.w$) reordering is comparatively low. Furthermore, the results for $GapR$ and $StrR$ show an improvement (greater for $StrR$); in other words, both modules can achieve a statistically significant improvement in performance in comparison with the original phrase based-model. Here, only the $StrR$ module reaches approximately the same level of performance as the original hierarchical phrase-based model without statistical significance. Although the performance of the model with all modules(SGR) is higher than that of the original hierarchical phrase-based model, the difference is not always statistically significant, which is likely due to unstable MERT.

Table 3 presents an example of translation performed with our model and the original hierarchical phrase-based model, and Figs. 8 and 9 show the derivations. In the original 2-SCFG derivation in Fig. 8, there is a rule in the gray box which swaps the two parts around the Japanese word から (literally *from*), whereas the Japanese word へ is translated separately as an ordinary word, even though in this case it is used together with から to express a complete idea (literally ... *from* ... *to* ...). As shown in Fig. 9, in our model, the three-tuple Japanese pattern X_3 から X_4 へ X_5 has been transformed as an entity into the pattern X_5 *from* X_3 *to* X_4 on the English side. Grammatically, the part of X_5 is a verb

phrase which usually precedes the preposition phrase in English but always follows the postposition phrase in Japanese. With the structure transformation rules, several phrases can be moved simultaneously, which allows for long-distance reordering to be realized in translation.

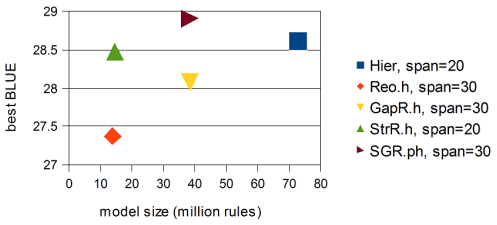

Figure 10: *Highest performance of the original model and our hierarchical model, and their size in millions of rules*

Fig. 10 clearly shows the large size of the original hierarchical phrase-based model in comparison with our model with different modules. The number of rules in the structure transformation rule set is sufficiently small to allow the set to be omitted, considering the size of the other modules. However, this would affect the result since there is considerable improvement in performance between the model with this rule set and that without it.

5. Discussion

Different approaches have been developed to handle the long-distance reordering problem. Nearly all of these approaches use grammatical information, either only in training or both in training and decoding.

In [17], the POS tags of the foreign language side were used in modeling the reordering information as well as in decoding, which is also employed in a lattice, where the sequence and the context of POS tags were taken into account together with words. Furthermore, in the approach proposed in [18], the monotone/swap rules are extracted by the foreign language parse tree and therefore a large amount of linguistic knowledge is necessary for the foreign language side. Regarding hierarchical phrase-models, in [19], the dominative relationship of pairs of function words between source and target languages were studied, and thus a derivation with a more appropriate structure was able to achieve a higher score. The function words used in [19] were simply words occurring with high frequency, which might introduce sensitivity towards the type of corpus used in the training, and the relationship between more than two words was not modeled. The approach in [19] can also be applied to obtain word alignments [20]. In contrast, our approach involves a hierarchical rule set which can capture long-distance reordering patterns with simple and corpus-independent POS knowledge in the initial step. In fact, both training and decoding are performed at the lexical level and POS tags are not used anywhere in the process.

Other more complex models have also been devised. In English-Japanese translation in [21], the input sentences were divided into clauses, which were translated separately, and the partial translation results were integrated to obtain the complete translated sentence. Furthermore, in [22], precise morphological analysis was applied in English-Arabic translation, leading to improved performance. These models are designed for particular translation tasks, where specific language features are considered. Although not completely excluded, grammatical knowledge is used moderately in our model in order to ensure the robustness and simplicity. Therefore, our approach is not specific to Japanese and can be generalized to any other translation task in which global reordering is necessary.

There are also several approaches to compaction of large translation models. As proposed in [23], a significance test can be used to prune the rule table, while a method based on a log-likelihood ratio is proposed in [24]. Furthermore, in [25], a second-pass extraction is used for selecting suitable phrases. The fact that our model is smaller is only a natural result of the compact but powerful structure transformation rule set, and it is not the product of intentional design. In fact, the reordering module and the gap phrase module are extracted indiscriminately and could potentially be pruned even further.

6. Conclusions and Future Work

In this paper, we proposed structure transformation rules that are lexicalized only on a small word set in order to allow for the extraction of long-distance reordering hierarchical rules with more than two nonterminals. In a Japanese-English translation experiment, it was found that this rule set can vastly improve the performance of our translation system while maintaining such a small size that it can even be omitted from the total rule set.

In future work, we plan to apply our model on a larger corpus, such as NTCIR-8 (3 million sentence pairs). Furthermore, we intend to explore methods for obtaining the initial word set automatically, as well as to compare the abilities of the method with manual and automatic extraction of the initial word set. We also aim to develop a method for extracting the structure transformation rules directly from a parallel corpus under some optimization criteria.

7. References

[1] Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., and Mercer, R. L., The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, Vol.19, No.2, pp. 263-311, 1993.

[2] Och, F. J., and Ney, H., A systematic comparison of various statistical alignment models *Computational Linguistics*, Vol.29, No.1, pp. 19-51, 2003.

[3] Koehn, P., Och, F. J., and Marcu, D., Statistical phrase

based translation. In *Proceedings of HLT-NAACL*, pp. 48-54, Edmonton, May-June, 2003.

[4] Tillmann, C., A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL*, 2004.

[5] Galley, M., and Manning, C. D., A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 848-856, Honolulu, October,r 2008.

[6] Chiang, D., Hierarchical phrase-based translation. *Computational Linguistics*, Vol.33, No.2, pp. 201-228, 2007.

[7] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311-318, Philadelphia, July, 2002.

[8] Fujii, A., Utiyama, M., Yamamoto, M., and Utsuro, T., Overview of the patent translation task at the NTCIR-7 Workshop. In *Proceedings of NTCIR-7 Workshop Meeting* Tokyo, Japan, December 16-19, 2008.

[9] Matsumoto, Y., Takaoka, K., and Asahara, M., ChaSen morphological analyzer version 2.4.0 user's manual. 2007.

[10] Och, F. J., and Ney, H., Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 295-302, Philadelphia, July, 2002.

[11] Watanabe, T., Tsukada, H., and Isozaki, H., Left-to-right target generation for hierarchical phrase-based translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 777-784, Sydney, Australia, 2006.

[12] Koehn, P., Moses statistical machine translation system user manual and code guide. September, 2011.

[13] Och, F. J., Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 160-167, July, 2003.

[14] Stolcke, A., SRILM - an extensible language modeling toolkit In *Proc. Intl. Conf. Spoken Language Processing* Denver, Colorado, September, 2002.

[15] Chen, S. F., and Goodman, J., An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

[16] Koehn, P., Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388-395, Barcelona, Spain, 2004.

[17] Rottmann, K., and Vogel, S., Word reordering in statistical machine translation with a POS-based distortion model. *The 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, Skövde, Sweden, September 7-9, 2007.

[18] Li, C.-H., Zhang, D., Li, M., Zhou, M., Li, M., and Guan, Y., A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 720-727, Prague, Czech Republic, June, 2007.

[19] Setiawan, H., Kan, M.-Y., Li, H., and Resnik, P., Topological ordering of function words in hierarchical phrase-based translation. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pp. 324-332, Suntec, Singapore, August 2-7, 2009.

[20] Setiawan, H., Dyer, C., and Resnik, P., Discriminative word alignment with a function word reordering model. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 534-544, MIT, Massachusetts, USA, October 9-11, 2010.

[21] Sudoh, K., Duh, K., Tsukada, H., Hirao, T., and Nagata, M., Divide and translate: improving long distance reordering in statistical machine translation. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pp. 418-427, Uppsala, Sweden, July 15-16, 2010.

[22] Badr, I., Zbib, R., and Glass, J., Syntactic phrase reordering for English-to-Arabic statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 86-93, Athens, Greece, March 30 - April 3, 2009.

[23] Johnson, H., Martin, J., Foster, G., and Kuhn, R., Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 967-975, Prague, June, 2007.

[24] Wu, H., and Wang, H., Comparative study of word alignment heuristics and phrase-based SMT. In *Proceedings of the MT Summit XI*, 2007.

[25] Zettlemoyer, L., and Moore, R., Selective phrase pair extraction for improved statistical machine translation. In *Proceedings of NAACL HLT 2007, Companion Volume*, pp. 209-212, Rochester, New York, April, 2007.