Diversion of Hierarchical Phrases as Reordering Templates

Mikio Yamamoto, Jyunya Norimatsu, Mitsuru Koshikawa, Takahiro Fukutomi,
Taku Nishio, Kugatsu Sadamitsu, Takehito Utsuro
University of Tsukuba
1-1-1 Tennodai, Tsukuba, 305-8573, Japan

Masao Utiyama
NICT

Shunji Umetani
Osaka University

Tomomi Matsui
Chuo University

Abstract

In the hierarchical phrase-based translation model (Chiang 2007), translation rules handle both context-sensitive translation and reordering of phrases at the same time. This simultaneity is strengths and weaknesses of the model. Although it enables the rules to be applied to the accurate and correct context, it deteriorates the applicability of the rules. In other words, the rules work very well in domains of training data, but they lost robustness in out of the domains. In this paper, we will try to improve the applicability of the original model by adding extra reordering templates which are separated out from hierarchical phrase translation rules. An original hierarchical phrase rule with two non-terminals is regarded as either monotone or swap reordering template according to if the two non-terminals in the source side have monotone or swap relation to the target side in the original rule. We will describe experiments in which the original model compares with our extensions in BLEU as a metric of translation quality using shared data at the NTCIR-7 patent translation task.

Keywords: *Patent Information Processing, Statistical Machine Translation, Phrase Reordering.*

1 Introduction

The hierarchical phrase-based model (Chiang 2007) is a simple and powerful framework to integrate context-sensitive translation and reordering of phrases at the same time. Such an integration of syntactic information is expected to be a key idea to go beyond the borders of simple phrase-based models, and many groups continue to improve the hierarchical model with a variety of approaches. For example, Menezes and Quirk (2007) pointed out Chiang's style of integration can narrow the applicability of hierarchical rules and they proposed the 'dependency treelet translation' model, which provides the wide applicability of the rules. Their model separates context-sensitive translation rules and reordering rules using external syntactic knowledge of source and target languages such as parts of speech of words and syntactic structures of sentences. Through this separation, their model realizes high generality of the applicability to out of the domain.

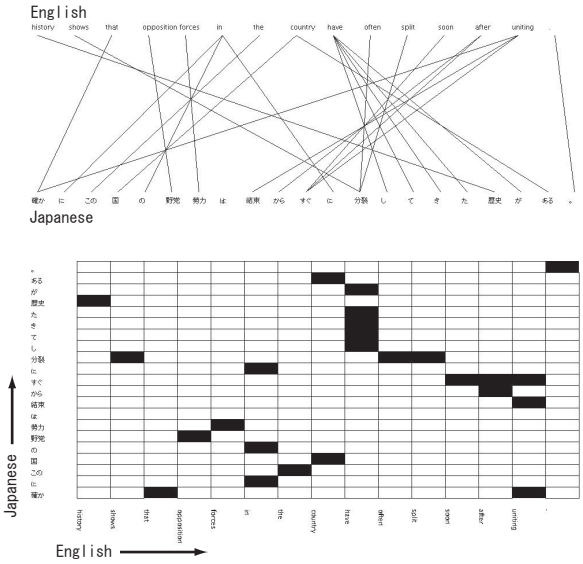
We investigated another method to separate out translation and reordering rules. Especially reordering rules are important on translation between Japanese and English, because many phrases in a English sentence move to longer distant positions in a corresponding Japanese sentence, relatively rather than the other language pairs such as a pair of Chinese and English. In our model, extra reordering templates are separated out from original hierarchical phrase rules without external syntactic knowledge. Each template supports either monotone or swap order of a pair of adjacent phrases (hypotheses) according to if the two non-terminals have monotone or swap relation in the original rule.

In the next section we observe phrase reordering in Japanese-English translation and results of a simple experiment to see an effect of reordering constraints in decoding. In Section 3, we describe our simple method to extract reordering templates from original hierarchical rules, which are named Hierarchical Phrases As Reordering Templates, HPART for short. In Section 4, we compare the original hierarchical phrase-based model with our extended models including HPART in BLEU as a metric of translation quality using shared data at the NTCIR-7 patent translation task.

2 Phrase Reordering in Japanese-English translation

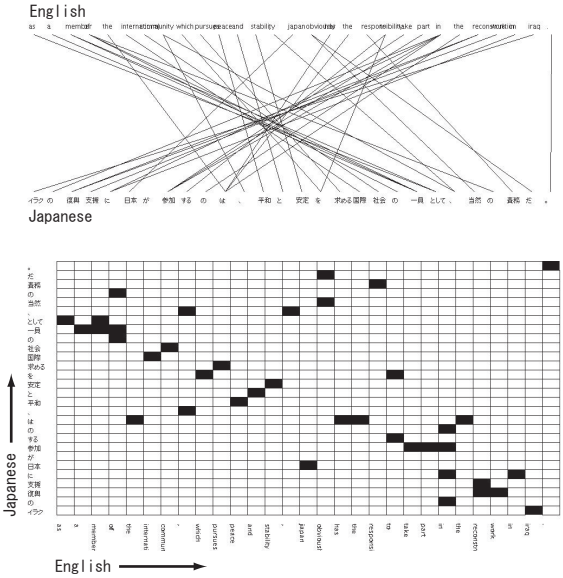
In this section, we describe our incentive to focus on phrase reordering by observations and simple experiments in Japanese-English translation.

Nagata et al. (2006) reported that non-local phrase reordering patterns such as monotone-gap and reverse-gap are more frequent in Japanese-English (35%) rather than those in Chinese-English (14%). And also, we surprised that reverse reordering make



English: History shows that opposition forces in the country have often split soon after uniting.
Japanese: 確かにこの国の野党勢力は結束からすぐに分裂してきた歴史がある。

Figure 1. A typical word alignment in J-E translation: a highly random alignment.



English: As a member of the international community, which pursues peace and stability, Japan obviously has the responsibility to take part in the reconstruction work in Iraq.
Japanese: イラクの復興支援に日本が参加するのは、平和と安定を求める国際社会の 一員として、当然の責務だ。

Figure 2. A typical word alignment in J-E translation: a reverse (or swapping) alignment.

up 28% in Japanese-English, on the other hand, 7% in Chinese-English. Figure 1 and 2 shows typical word alignments between Japanese and English sentences automatically computed by GIZA++ (Och and Ney 2003). Many words in a Japanese sentence move to long distant positions in a corresponding English sentence.

Using our decoder (Hiero copy) for the hierarchical phrase-based model, we conducted a simple experiment to see an effect of global reordering constraints. The original Hiero decoder adopts a monotone rule as global phrase reordering [1]. However, as is evident from Figure 1 and 2 in Japanese-English translation, we need global reordering rules allowing long distance phrases movement with swap. We run the decoder with two kinds of global reordering constraints; the original global monotone constraint and that with the ITG constraint (Wu 1996) which allow two phrase hypotheses to be swapped on the arbitrary node in the binary tree [2].

Table 1 shows BLEU values of two kinds of global reordering constraints. The experiment was conducted in the same condition described in Section 4. The table shows that there exist many good phrases swapped globally or locally which cannot be generated as translation candidates by the original hierarchical phrase-based model in Japanese-English translation. In this paper, we will investigate not global reordering rules but rather local reordering rules slipped out of the original model. Reordering rules are distinct from reordering constraints. Although reordering constraints don't give reordering hypotheses probabilistic scores, reordering rules do. In the case of using only a constraint for reordering, since only the language model for target side provides information to select a phrase order from possibilities satisfying the constraint, it's not necessarily the case that a selected phrase order properly reflects the meanings in source language.

In the next section, we describe our simple method to extract reordering regularities from hierarchical phrase rules in order to utilize information the original model intrinsically has.

[1] S→ <SX, SX>

[2] ITG constraints are realized by the following two rules:
S → <S_0 S_1, S_0 S_1>
S → <S_0 S_1, S_1 S_0>
The second rule allows two adjacent nodes to be swapped.

Table 1. BLEU in JE translation

Constraints		Monotone	ITG
BLEU	span=10	22.8 %	24.2 %
	span=15	23.0 %	23.6 %

3 HPART: Hierarchical Phrases As Reordering Templates

3.1 Basic idea

Consider translation for noun phrases including the Japanese word 'に関する' such as the pattern 'NOUN + に関する + NOUN.' The word 'に関する', which roughly means 'relating to' in English, can be translated into a variety of English expressions such as 'of', 'for', 'on', comma and so on, but the positions of the anteroposterior noun phrases are consistently translated into English in reverse order (Figure 3). The hierarchical phrase model formalizes one of the translation by the following rule, for example.

X → <X_0 に関する X_1, X_1 of X_0 >

This is a rewrite rule of a synchronous CFG (Chiang 2007). The left sequence in the bracket is a rewrite rule of normal CFG in source language, the right sequence is for target language. 'X_0' and 'X_1' are non-terminal symbols, and the same named symbols in each side have to keep meanings equivalent. For example, a Japanese phrase rewritten from 'X_0' next to 'に関する' should be translated into an English phrase at the position of 'X_0' in target side, that is, at the right position of the word 'of'.

As mentioned above, since the word 'に関する' can be translated into a variety of English expressions, we need as many rules as the number of expressions such as:

X → <X_0 に関する X_1, X_1 for X_0 >
X → <X_0 に関する X_1, X_1 on X_0 >
X → <X_0 に関する X_1, X_1 ',' X_0 >
...

In the ideal case, a training program can lean all rules from enough parallel data.

However, ideal learning of rules is prevented by noise of parallel data such as non-literal translations and poor sentence alignments, and also many errors of automatic word alignments (Blunsom et al., 2008). Instead, many partial phrase or hierarchical phrase translation rules are extracted from that noisy parallel data, such as the followings.

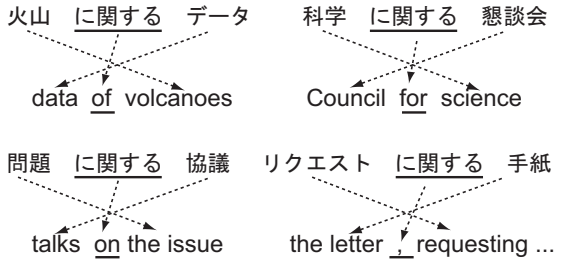


Figure 3. Variety and consistency in Japanese-English translation.

X^a → <に関する X_0, X_0 of >
X^b → <科学 に関する, for science >
X^c → <に関する, on >
...

Each of the above rules not only losts reordering information for swapping anteroposterior phrases around the word 'に関する', but also cannot generate the swap rule even if they collaborate, because the range of covering by the above three rules are duplicated in a word sequence of source sentences. Figure 4 shows the duplicated application of the above three rules to the Japanese fragment '科学 (science) に関する (of) ...' Since phrase-based models don't allow duplicated application of rules to the same word sequence in general, the above three rules are exclusive and cannot been combined. Our central idea in this paper is to allow these duplications in order to combine relatively general reordering regularities and partial translation rules.

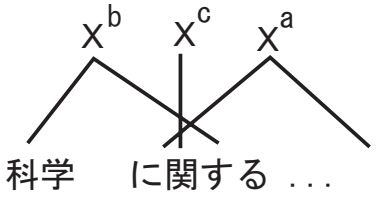


Figure 4. Examples of duplicated application of some rules.

3.2 Extra reordering templates

We exploit reordering regularities in the hierarchical phrase rules including two non-terminal symbols in the each side, and ignore translation information, that is, don't use lexical word sequences in the target side of rules. Consider the following rule again.

X → <X_0 に関する X_1, X_1 of X_0 >

This rule implies some reordering information: anteroposterior phrases on 'に関する' in source language (Japanese) should be swapped in the target language (English). At the same time, we can ignore the fact that the word 'of' is a translated word corresponding to the word 'に関する'.

All hierarchical phrase rules including two non-terminal symbols can be regarded as extra reordering templates, and they are classified into two types: monotone and swap. In templates of the type monotone, the order of two nonteminals in each language side is the same, and the reverse order in templates of the type swap. For example, the above templates of 'に関する' is classified into the type swap. Both types of extra reordering templates are defined as the following general ITG rules.

X → <X_2 X_3, X_2 X_3 >(monotone)
X → <X_2 X_3, X_3 X_2 >(swap)

However, this generalization is limited to only the case that two adjacent source phrase hypotheses (that is, 'X_2 X_3' in source side of the ITG rules) match with the pattern of the source side of the original hierarchical phrase rule [3]. Figure 5 shows two examples of application of the extra reordering rule converted from the above rule.

We added two scores related to extra reordering templates as features of the log-linear model to give a target sentence a probability; the number of uses of templates (templates penalty) and the total scores of the original rules that formed the templates used for translation.

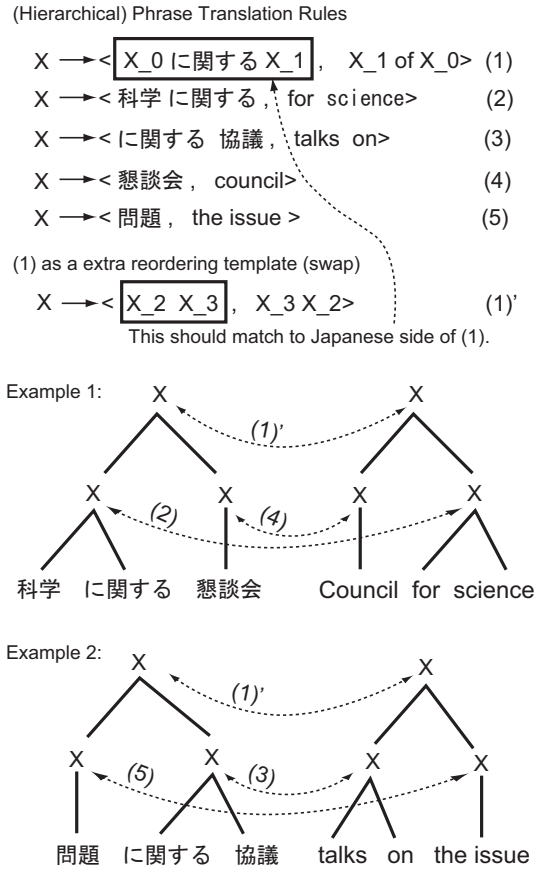


Figure 5. Examples of extra reordering rules and the application

[3] You might have concerned how a decoder can decide where the source word sequence divides at for each X_2 and X_3, but a standard decoder using CKY algorithm generates phrase hypotheses in the bottom-up manner, so the decoder knows dividing points as applying an extra reordering rule.

Table 2. BLEU with the single reference in JE translation for the formal run

Global reordering constraint: Monotone

	HP		HP+HPARTs
span=10	22.84 %	$\ll$	24.04 %
span=15	23.00 %	$\ll$	23.99 %

Global reordering constraint: ITG

	HP		HP+HPARTs
span=10	24.20 %	$\not<$	24.24 %
span=15	23.57 %	$\not<$	23.59(19.93) %

'$\ll$' means significantly difference with significance level of 1%.
'HP' stands for 'Hierarchical Phrase rules.'

4 Experiments

We estimated a hierarchical phrase table for J-E translation from PSD-1 of training data at the NTCIR-7 (Fujii et al. 2008) using Chiang's heuristics (Chiang 2007) and a phrase table from PSD-1 using the script within Moses package (Koehn et al. 2007). 5-gram language model was constructed from text of English side of PSD-1 using SRILM (Stolcke 2007) with Interpolated modified Kneser-Ney smoothing (Chen and Goodman 1998). The test-set (899 sentences) in PSD-1 training data was used as development data for MERT (Och 2003) using Utiyama's implementation (Uchiyama 2006) written in Ruby.

BLEU (Papinei et al. 2002) is computed by the official tool for the NTCIR-7 patent translation task: 'bleu_kit' (Norimatsu 2008). Our submitted results, marked by 'MIBEL' in the overview paper (Fujii et al. 2008), for the formal run were the output of our decoder based on the hierarchical phrase-based model with HPARTs. However, after we submitted the results, we found out the order of parameters in the hierarchical phrase table doesn't match with the order of weights table, so the official evaluation on the NTCIR-7 patent translation task may undervalue our system's performance. Table 2 shows the BLEU values with the single reference in Japanese-English translation (1381 sentences) of the intrinsic evaluation for the formal run. 'Span' in the table means the length (the number of words) limitation as a phrase in the source side for decoding. The value in the parentheses is the official score of MIBEL team.

In all cases, the models with HPARTs improves over the original models. However, only the cases of the monotone global reordering constraint are statistically significant ($p < 0.01$). In the case of the ITG constraint, advancement by HPARTs is little. I think that flexibility of global reordering by the ITG may negate an effect of HPARTs.

5 Conclusion

We described a method for diversion of hierarchical phrase translation rules as reordering templates to utilize information about reordering the hierarchical phrase-based models intrinsically have. In the case that the global reordering constraint is monotone, HPARTs are significantly effective for translation quality.

We think our method is too simple to avoid generating many over-generalized reordering templates. To improve this weak points of HPARTs, we plan to utilize the other translation rules sharing a part of the word sequence in the phrase or the target side information of the original rules.

References

[Blunsom et al., 2008] P.Blunsom, T.Cohn and M.Osborne. 2008. A discriminative latent variable model for statistical machine translation. In Proc. of ACL-08, pages 200–208.

[Chen and Goodman 1998] S.F.Chen and J.Goodman. 1998. An empirical study of smoothing techniques for language modeling. Tech. Report TR-10-98, Harvard Univ. Center for Research in Computing Technology.

[Chiang et al. 2005] D. Chiang et al. 2005. The Hiero machine translation system: extensions, evaluation, and analysis. In Proc. of HLT/EMNLP-05, pages 779–786.

[Chiang 2007] David Chiang. 2007. Hierarchical phrase-based translation. Computational Linguistics 33(2), pages 201–228.

[Fujii et al. 2008] A.Fujii, M.Yamamoto, M.Utiyama and T.Utsuro. 2008. Overview of the patent translation task at the NTCIR-7 workshop. In Proc. of NTCIR-7.

[Koehn et al. 2007] P. Koehn et al. 2007. Moses: open source toolkit for statistical machine translation. In Proc. of ACL-07, demonstration session.

[Menzes and Quirk 2007] Arul Menezes and Chris Quirk. 2007. Using dependency order templates to improve generality in translation. In Proc of 2nd workshop on SMT, pages 1–8.

[Nagata et al. 2007] M.Nagata, K.Saito, K.Yamamoto and K.Ohashi. 2006. A clustered global phrase reordering model for statistical machine translation. In Proc. of ACL-06, pages 713–720.

[Norimatsu 2008] Jyunya Norimatsu. 2008. NTCIR scoring tools for patent translation task. http://www.mibel.cs.tsukuba.ac.jp/ñorimatsu/bleu_kit/

[Och 2003] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In Proc. of ACL-03.

[Och and Ney 2003] Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), pages 19–51.

[Papineni et al. 2002] K.Papineni, S.Roukos, T.Ward and W.Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proc. of ACL-02, pages 311-318.

[Stolcke 2002] A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In Proc. of ICSLP-02, Vol.2, pages 901–904.

[Utiyama 2006] Masao Utiyama. 2006. Mert — minimum error rate training package for machine translation. http://www2.nict.go.jp/x/x161/members/mutiyama

[Wu 1996] Dekai Wu. 1996. A polynomial-time algorithm for statistical machine translation. In Proc. of ACL-96, pages 152–158.