PAPER

Imposing Constraints from the Source Tree on ITG Constraints for SMT

Hirofumi YAMAMOTO[†,††a)], *Member*, Hideo OKUMA[††], *Nonmember*, *and* Eiichiro SUMITA[††], *Member*

SUMMARY In the current statistical machine translation (SMT), erroneous word reordering is one of the most serious problems. To resolve this problem, many word-reordering constraint techniques have been proposed. Inversion transduction grammar (ITG) is one of these constraints. In ITG constraints, target-side word order is obtained by rotating nodes of the source-side binary tree. In these node rotations, the source binary tree instance is not considered. Therefore, stronger constraints for word reordering can be obtained by imposing further constraints derived from the source tree on the ITG constraints. For example, for the source word sequence { a b c d }, ITG constraints allow a total of twenty-two target word orderings. However, when the source binary tree instance ((a b) (c d)) is given, our proposed "imposing source tree on ITG" (IST-ITG) constraints allow only eight word orderings. The reduction in the number of word-order permutations by our proposed stronger constraints efficiently suppresses erroneous word orderings. In our experiments with IST-ITG using the NIST MT08 English-to-Chinese translation track's data, the proposed method resulted in a 1.8-points improvement in character BLEU-4 (35.2 to 37.0) and a 6.2% lower CER (74.1 to 67.9%) compared with our baseline condition.

key words: *statistical machine translation, word reordering, distortion model, ITG constraints*

1. Introduction

Statistical methods are widely used for machine translation. One of the popular statistical machine translation paradigms is the phrase-based model (PBSMT) [1]–[3]. In PBSMT, errors in word reordering, especially in global reordering, are one of the most serious problems. Approaches used to resolve this problem are categorized into two types. The first type is linguistically syntax-based. In this approach, source-side [4]–[6] or target-side [7]–[9] tree structures, or both [10], [11] are used for model training. The second type is formal constraints on word permutations. Distance based distortion model [12], lexical word reordering model [13], and inversion transduction grammar (ITG) constraints [14], [15] belong to this type of approach. Our approach is an extension of ITG constraints and is a hybrid of the two approaches.

We propose "imposing source tree on ITG" (IST-ITG) constraints for directly introducing source sentence structure into our set of constraints. In IST-ITG, ITG constraints under the given source sentence tree structure are used as stronger constraints than the original ITG. For example, IST-ITG allows only eight word orderings for a four-word sentence, even though twenty-two word orderings are possible in the original ITG constraints.

In Sect. 2, we give an outline of statistical machine translation and previous word reordering constraints. In Sect. 3, we present the proposed IST-ITG for word-based translation. In Sect. 4, the proposed method is extended to phrase-based translation. In Sect. 5, we present a real-time decoding algorithm for IST-ITG constraints. In Sect. 6, we give details of the experiments and present the results. Finally, in Sect. 7, we offer a summary and some concluding remarks.

2. Statistical Machine Translation

The framework of statistical machine translation formulates the following problem.

$$\hat{e} = \underset{e}{argmax}\ P(e|f) \tag{1}$$

In above equation, translation target sentence $\hat{e}$ which gives maximum probability is calculated under the given source sentence f. This equation can re-written to the next equation using Bayes' law.

$$\hat{e} = \underset{e}{argmax}\ P(f|e)P(e) \tag{2}$$

Here, $P(f|e)$ represents a translation model, $P(e)$ represents a language model. Usually, the translation model is decomposed to a lexical translation model and a word reordering model in which proposed IST-ITG constraints belongs. The lexical translation model is represented as $\prod p(F_j|E_i)$, which is the multiplication of word-to-word translation probabilities from ith source word E_i to jth target word F_j. The word reordering model is represented as $p(a|e, f)$, which is the word alignment probability under the given source and target sentences e and f. The translation model $P(f|e)$ can be re-written to next equation as the multiplication of the lexical translation model and the word reordering model.

$$P(f|e)P(e) = \left(\prod p(F_j|E_i)\right) \times p(a|e, f) \tag{3}$$

As word reordering model, we introduce three previous studies on word reordering constraints: distance based distortion penalty; lexical reordering model; and ITG constraints. Here, we consider one-to-one word-aligned source and target language sentence pairs as the simplest cases.

Manuscript received December 24, 2008.
Manuscript revised April 13, 2009.

[†]The author is with School of Science and Engineering, Dept. Informatics, Kinki University, Higashiosaka-shi, 577–8502 Japan.
[††]The authors are with National Institute of Communications Technology, Kyoto-fu, 619–0288 Japan and with ATR Spoken Language Translation Research Laboratories, Kyoto-fu, 619–0288 Japan.
a) E-mail: yama@info.kindai.ac.jp

DOI: 10.1587/transinf.E92.D.1762

Copyright © 2009 The Institute of Electronics, Information and Communication Engineers

2.1 Distance Based Distortion Penalty

In this constraint, a distortion penalty is given in accordance with the gap between the previously and the currently source words, which is represented as the following equation.

$$p_D = exp\left(-\sum_i d_i\right) \tag{4}$$

where d_i for each i is defined as:

$$d_i = abs(position(f_{i-1}) + 1 - position(f_i)) \tag{5}$$

where source word f_i is translated to ith target word e_i, $position(f)$ represents the position of the word f in the source sentence. Sometimes, a limit is set for d_i for similar language pairs such as French and English. However, for dissimilar language pairs, such as Japanese and English or Chinese and English, limiting d_i is not beneficial. If only this model is used as the word reordering model, word reordering model is approximated by next equation.

$$p(a|e, a) \approx p_D \tag{6}$$

2.2 Lexical Reordering Model

In the lexical reordering model, reordering probabilities are assigned to each word pair $\{f_i, e_i\}$. Reordering positions are categorized into three types, monotone, swap, and discontinuous. The probability is assigned to left and right sides as $p_s(t|f_i, e_i)$, where, s is left (l) or right (r), t is monotone (m), swap (s), or discontinuous (d). Therefore, a total of six probabilities are assigned to each word pair. For the source word sub-sequence f_{i-1}, f_i, probabilities of target sub-sequences are calculated as follows:

- $p(e_{i-1}, e_i) = p_r(m|f_{i-1}, e_{i-1})p_l(m|f_i, e_i)$
- $p(e_i, e_{i-1}) = p_r(s|f_{i-1}, e_{i-1})p_l(s|f_i, e_i)$
- $p(otherwise) = p_r(d|f_{i-1}, e_{i-1})p_l(d|f_i, e_i)$

2.3 ITG Constraints

In one-to-one word alignment, the source word f_i is translated into the target word e_i. The source sentence $[f_1, f_2, \ldots, f_N]$ is translated into a reordered sequence of word $[e_1, e_2, \ldots, e_N]$. The number of reorderings is $N!$. When ITG is introduced, this combination $N!$ can be reduced in accordance with the following constraints.

- All possible binary tree structures are generated from the source word sequence.
- The target sentence is obtained by rotating any node of the binary trees.

When $N = 4$, the ITG constraints can reduce the number of combinations from $4! = 24$ to 22 by rejecting combinations $[e_3, e_1, e_4, e_2]$ and $[e_2, e_4, e_1, e_3]$. For a 4-word sentence, the search space is reduced to 92%(22/24), but for 10-word sentence, the search space is only 6%(206,098/3,628,800) of the original full space. Generally, the number of combinations in n-word sentence S_n is represented by the following formula.

$$nS_n = 3(2n-3)S_{n-1} - (n-3)S_{n-2} \tag{7}$$

Experimental results using ITG constraints are reported by Zen et al.[16]. In this experiment using a Japanese-English conversational corpus, ITG constraints resulted in almost the same performance as the distance based model.

3. Imposing the Source Tree on ITG Constraints

3.1 Imposing Source Tree Constraints

In ITG constraints, the source-side binary tree instance is not considered. Therefore, if the source sentence binary tree is utilized, stronger constraints than the original ITG can be created. By parsing the source sentence (for training data, parsing is not required), a parse tree is obtained. After parsing, a bracketed sentence is obtained by removing the node labels, and this bracketed sentence can be converted to a binary tree. For example, the parse tree, (S1 (S (NP (DT This)) (VP (AUX is) (NP (DT a) (NN pen))))), is obtained from the source sentence "This is a pen". By removing the node labels, a bracketed sentence ((This) ((is) ((a) (pen)))) is obtained. Such a bracketed sentence (equivalent to a binary tree) can be used to produce constraints. If IST-ITG is applied, the number of word orderings in $N = 4$ is reduced to 8, down from 22 with ITG. For example, for the source-side bracketed tree $((f_1\ f_2)(f_3\ f_4))$, eight target sequences $[e_1, e_2, e_3, e_4]$, $[e_2, e_1, e_3, e_4]$, $[e_1, e_2, e_4, e_3]$, $[e_2, e_1, e_4, e_3]$, $[e_3, e_4, e_1, e_2]$, $[e_3, e_4, e_2, e_1]$, $[e_4, e_3, e_1, e_2]$, and $[e_4, e_3, e_2, e_1]$ are accepted. For the source-side bracketed tree $(((f_1\ f_2)f_3)f_4)$, eight sequences $[e_1, e_2, e_3, e_4]$, $[e_2, e_1, e_3, e_4]$, $[e_3, e_1, e_2, e_4]$, $[e_3, e_1, e_2, e_4]$, $[e_4, e_1, e_2, e_3]$, $[e_4, e_2, e_1, e_3]$, $[e_4, e_3, e_1, e_2]$, and $[e_4, e_3, e_2, e_1]$ are accepted. Generally, the number of word orderings is reduced to 2^{N-1}. Table 1 shows the number of word orderings in a target word sequence for each N with ITG, IST-ITG, and no constraints.

Table 1 Number of word orderings in each type of constraint.

N	IST-ITG	ITG	No Constraint
1	1	1	1
2	2	2	2
3	4	6	6
4	8	22	24
5	16	90	120
6	32	394	720
7	64	1806	5040
8	128	8558	40320
9	256	41586	362880
10	512	206098	3628800
11	1024	1037718	39916800
12	2048	5293446	479001600
13	4096	27297738	6227020800
14	8192	142078746	87178291200
15	16384	745387038	1307674368000

3.2 Extension to Non-binary Tree

In the above subsection, a source binary tree was assumed in order to perform IST-ITG. However, parsing results sometimes are not binary trees. In this case, some tree nodes have more than two branches. For a non-binary node, any reordering of branches is allowed. In a non-binary tree $(f_1(f_2\ f_3\ f_4))$, twelve target-side sequences $[e_1, e_2, e_3, e_4]$, $[e_1, e_2, e_4, e_3]$, $[e_1, e_3, e_2, e_4]$, $[e_1, e_3, e_4, e_2]$, $[e_1, e_4, e_2, e_3]$, $[e_1, e_4, e_3, e_2]$, $[e_2, e_3, e_4, e_1]$, $[e_2, e_4, e_3, e_1]$, $[e_3, e_2, e_4, e_1]$, $[e_3, e_4, e_2, e_1]$, $[e_4, e_2, e_3, e_1]$, and $[e_4, e_3, e_2, e_1]$ are allowed. For nodes that have more than three branches, the original ITG constraints are locally applied. Therefore, for a non-binary tree $(f_1(f_2\ f_3\ f_4\ f_5))$, $22 \times 2 = 44$ word orderings are allowed in the target-side and represented by the following formula.

$$\prod_{i=1}^{n}(S_{Bi}) \tag{8}$$

where S_k represents the number of combinations from the original ITG constraints for $N = k$ and Bi represents the number of branches at the ith node.

4. IST-ITG in Phrase-Based SMT

In the above section, we described each constraint in the case of a one-to-one word alignment. In this section, we consider phrase-based models. Usually, "phrase" has two meanings. One is simple a word sequence that is used as "phrase-based", the other is syntactic sub parts of a sentence. Through out this paper, "phrase" is meant as a simple word sequence, not as syntactic sub parts. When a phrase-based model is used, each constraint must be extended. For distance based distortion penarty, Eq (5) is rewritten using phrase Pe_n instead of word f_n as follows:

$$\begin{aligned} d_i = {} & abs(last\_position(Pf_{i-1}) + 1 \\ & -first\_position(Pf_i)) \end{aligned} \tag{9}$$

where, $last\_position(Pf_n)$ represents the position of the last word in nth phrase, $first\_position(Pf_n)$ represents the position of the first word in nth phrase. The lexical reordering model and ITG constraints can be extended by changing the model (or constraint) unit from "word" to "phrase". However, in IST-ITG, "word" must be used for the constraint unit since the parse (bracketed tree) unit is in "words". Furthermore, "phrase" is extracted without considering parsing structure of training sentence. To absorb different units between translation models and IST-ITG constraints, we investigated a new limitation for word ordering as follows.

- Word ordering that destroys a phrase is not allowed.

When this limitation is applied, the translated word ordering is obtained from the bracketed source sentence tree by reordering the nodes in the tree, the same as for one-to-one word-alignment. According to this limitation, the following nodes cannot be reordered. If a sub-tree with root node X includes part of a phrase ph, node X cannot be reordered. Consider the source bracketed source tree ((f_a f_b f_c) ((f_d f_e) (f_f f_g))), in which f_b f_c, and f_d form a phrase f_{ph} as in Fig. 1. Node 1 cannot be reordered since part of the phrase f_b f_c is included in node 1's sub-tree. For the same reason, node 2 and 4 cannot be reordered. Node 3 can be reordered since the sub-tree does not include the phrase (target sequence $[e_a e_{ph} e_e e_g e_f]$ is obtained by rotating node 3). Node 5 also can be reordered since it includes the whole phrase (target sequence $[e_g e_f e_e e_{ph} e_a]$ is obtained by rotating node 5). If node 2 is reordered, phrase ph is split into two parts, and translated in two parts in the target sentence. It is inconsistent with the condition that phrase-to-phrase alignment is one-to-one. As a result, only the target sequences $[e_a e_{ph} e_e e_f e_g]$, $[e_a e_{ph} e_e e_g e_f]$, $[e_g e_f e_e e_{ph} e_a]$, and $[e_f e_g e_e e_{ph} e_a]$ are allowed. Here, e_{ph} represents an equivalent phrase in the translation for f_{ph}.

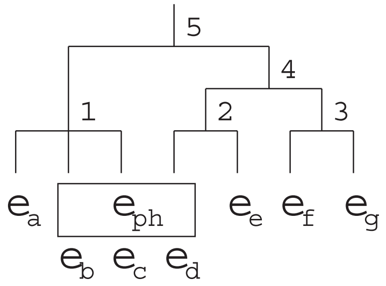

Fig. 1 Example sentence tree with a phrase.

5. Decoding with IST-ITG Constraints

In this section, we describe a one-pass decoding algorithm that uses IST-ITG constraints in the decoder. The translation target sentence is sequentially generated from left (sentence head) to right (sentence tail) according to multi-stack beam search algorithm used in Moses decoder [17]. To introduce the IST-ITG constraints into this algorithm, the target candidate must be checked whether it satisfies the IST-ITG constraints or not whenever a new phrase is selected to extend a target candidate.

To explain this checking algorithm, we categorized source sub-trees into four types **UNTRANSLATED**, **TRANSLATED**, **TRANSLATING**, and **NG** (no good) as follows:

- If a sub-tree consists of only leaf word nodes, and all leaf words are not yet translated, this sub-tree is defined as **UNTRANSLATED**.
- If a sub-tree consists of only **UNTRANSLATED** sub-trees, this sub-tree is also **UNTRANSLATED**.
- If a sub-tree consists of only leaf word nodes, and all leaf words are already translated, this sub-tree is defined as **TRANSLATED**.
- If a sub-tree consists of only **TRANSLATED** sub-trees, this sub-tree is also **TRANSLATED**.
- If a sub-tree consists of only leaf word nodes with both

translated and untranslated words, this sub-tree is defined as **TRANSLATING**.

- If a sub-tree consists of both **TRANSLATED** and **UNTRANSLATED** sub-trees, this sub-tree is **TRANSLATING**.
- If a sub-tree includes only one **TRANSLATING** sub-tree and any number (including zero) of **TRANSLATED** and **UNTRANSLATED** sub-trees, this sub-tree is **TRANSLATING**.
- If a sub-tree includes more than one **TRANSLATING** sub-tree, this sub-tree is **NG**.
- If a sub-tree includes **NG** sub-tree, this sub-tree is also **NG**.

If a translation candidate includes **TRANSLATING** sub-tree t, t must become **TRANSLATED** before anything else can happen. Given sub-tree $((ab)c)$, a is translated, b and c are not yet translated. In this case, b must be translated before c. If c is translated before b, the target word order becomes ACB. This word order does not satisfy the IST-ITG constraints. For the same reason, a candidate that includes an NG sub-tree does not satisfy the IST-ITG constraints. The checking algorithm for IST-ITG constraints is as follows.

1. For old translation candidates, the smallest **TRANSLATING** sub-tree t and its untranslated part u are calculated.
2. When a new target phrase e_{ph} is generated from the source phrase f_{ph}, f_{ph} and untranslated part u calculated in above step are compared. If f_{ph} does not include and is not included in u, the new candidate is rejected. For example, in Fig. 1, only source word f_a is already translated. The smallest **TRANSLATING** sub-tree is 1 and its untranslated part u is $[f_b f_c]$. In this case, phrases $[f_b]$, $[f_c]$, or $[f_b f_c]$ are accepted since these are included in u. Phrases $[f_b f_c f_d]$ or $[f_b f_c f_d f_e]$ are also accepted since these include u.
3. If a new candidate includes **NG** sub-trees, this candidate is rejected.

6. Experiments

6.1 Evaluation Measures

We evaluated the proposed method using four evaluation measures, BLEU [18], NIST [19], WER(word error rate), and PER(position independent word error rate). Before discussing the evaluation, the characteristics of each one are analyzed.

- BLEU: This evaluation measure takes into account middle range word order, but does not take into account global word order. When the translation result is $[w_1, w_2, \ldots, w_{j-1}, X, w_{j+1}, \ldots, w_n]$ for reference translation $[w_1, w_2, \ldots, w_n]$, both WER and BLEU scores will be high. For a translation result $[w_{j+1}, \ldots, w_n, X, w_1, w_2, \ldots, w_{j-1}]$, the BLEU score will be the same as the previous result since BLEU only takes into account 4grams. However, the WER score will be zero since global word positions are taken into account. Therefore, the effectiveness of the proposed method using BLEU is less than that of using WER.
- NIST: This evaluation measure only takes into account n-grams like BLEU. However, importance of higher order n-grams are less than BLEU. Therefore, the effectiveness of the proposed method using NIST will be less than that of using BLEU.
- WER: This evaluation measure takes into account not only local but also global word order, and is the most suitable for evaluating our method.
- PER: With this evaluation measure, we are almost incapable of considering word order. Therefore, our proposed method would seem to offer no improvement in this evaluation measure.

Table 2 E-J patent corpus.

	# of sent.	Total words	# of entries
E/J Train	1.8 M	60 M/64 M	188 K/118 K
E/J Dev	916	30 K/32 K	4,072/3,646
E/J Eval	899	29 K/32 K	3,967/3,682

6.2 English and Japanese Patent Corpus Experiments

First, we conducted experiments on English and Japanese patent translations. Details of the experimental corpus are shown in Table 2. This corpus is created by automatic sentence alignment [20]. The first nine hundred sentence pairs with the best alignment scores were used as the evaluation data (single reference) and the next thousand sentence pairs were used as the development data. This corpus is a subset of the training corpus that will be used in the NTCIR-7 Workshop patent translation track.

6.2.1 English-to-Japanese Translation

The translation direction of the first experiment was English-to-Japanese (E-J). For phrase-based translation model training, we used the GIZA++ toolkit [22]. For language model training, the SRI language model tool kit [23] was used. The language model type was word 5-gram smoothed by Kneser-Ney discounting [24]. For tuning of decoder parameters, we conducted minimum error training [25] with respect to the BLEU score using 916 development sentence pairs. For extraction of source sentence tree structure, we used the Charniak parser [26]. We used Chasen [27] for segmentation of the Japanese. The numbers of entries in the language models were 0.1 M, 2.1 M, 4.3 M, 6.2 M, and 6.9 M for 1, 2, 3, 4, and 5 grams respectively. The number of entries in the phrase-table was 76 M. For decoding, we used an in-house decoder that is a close relative to the Moses decoder. The translation accuracy of this decoder was configured to be the same as Moses. Another conditions are the same as the default conditions of Moses

Table 3 Evaluation results in E-J patent translation.

	BLEU	NIST	WER	PER
Monotone	24.91	6.95	79.97	42.02
No constraint	26.83	7.19	81.10	39.52
Dist	28.35	7.29	78.35	39.25
ITG	27.59	7.26	80.29	39.15
Dist+ITG	28.50	7.30	78.01	39.29
Dist+LR	31.17	7.50	76.30	38.61
IST	30.26	7.41	74.90	38.93
Dist+IST	30.07	7.41	73.38	39.05
Dist+LR+IST	32.20	7.61	71.18	38.15

decoder.

In these experiments, a distance based distortion penalty, a lexical reordering model, the proposed IST-ITC, and combinations of these are compared. The combination of constraints in these experiments is as follows.

1. Monotone: Monotone translation (no reordering).
2. No constraints: There were no constraints for word reordering. Any word order was allowed without penalty.
3. Dist: Distance based distortion penalty without distortion limit.
4. ITG: ITG constraints.
5. Dist+ITG: Both distance based distortion penalty and ITG constraints were used at the same time.
6. Dist+LR: Both distance based distortion penalty and lexical reordering model.
7. IST: Only the proposed IST-ITC constraints.
8. Dist+IST: Both distance based distortion penalty and IST-ITC constraints.
9. Dist+LR+IST: Distance based distortion penalty, Lexical reordering model, and IST-ITG constraints were used at the same time.

Table 3 shows the following experimental results. In comparing the original ITG constraints (ITG) with the proposed IST-ITG (IST) method, the improvement in BLEU was 2.67 points, and in WER was 5.39%. WER had the largest improvement, next was BLEU. This particular improvement order was the same as in the previous subsection. The large improvement of WER helped us confirm the effectiveness of the proposed method for global word ordering. When distance based distortion penalty were used at the same time (Dist+ITG and Dist+IST), the BLEU score improved by 1.57 points and WER improved by 4.63%. When the lexical reordering model was used at the same time (Dist+LR and Dist+LR+IST), BLEU improved by 1.03 points and WER improved by 5.12%. The lexical reordering model fixed phrase position for the monotone and swap categories, but did not fix phrase position for the discontinuous category. IST-ITG fixed phrase position for the discontinuous category, even though it did not assign a probability. Combinations of the lexical reordering model and IST-ITG resulted in a better WER than with both Dist+LR and Dist+IST since both position and probability could be assigned for the discontinuous category.

Next, we conducted a human evaluation for Dist+LR and Dist+LR+IST. The evaluation measures were fluency and adequacy. For these evaluations, we used paired comparison, since patent sentences are sometimes very long and complex to assign ranking values. Evaluation target sentences were the first 200 sentences of the automatic evaluation test set. Evaluation results by single evaluator are shown in Table 4.

Table 4 Human evaluation results in E-J patent translation.

	Dist+LR	Dist+LR+IST	comparable
Fluency	72	80	48
Adequacy	42	94	64

Table 5 Evaluation results in J-E patent translation.

	BLEU	NIST	WER	PER
Monotone	26.29	7.25	76.42	40.85
No constraint	26.20	7.18	81.41	40.76
Dist	27.87	7.34	78.16	39.94
ITG	27.01	7.24	80.43	40.50
Dist+ITG	28.16	7.35	78.04	40.07
Dist+LR	29.93	7.54	77.27	39.12
IST	28.32	7.31	76.62	40.67
Dist+IST	28.14	7.32	74.13	40.40
Dist+LR+IST	29.77	7.50	72.80	39.73

In fluency, the difference is not significant (in 72 sentences, Dist+LR is better than Dist+LR+IST, in 80 sentences Dist+LR+IST is better). In adequacy, the proposed Dist+LR+IST is significantly better than Dist+LR (92 vs. 42). The result of sign test is larger than 4σ (reliability of 2.33σ is 99%.)

6.2.2 Japanese-to-English Translation

Next, we conducted J-E translation experiments using the same corpus. The numbers of entries in the language models were 0.2 M, 3.1 M, 4.1 M, 5.7 M, and 5.9 M for 1, 2, 3, 4, and 5 grams. The number of entries in the phrase-table was 76 M. For parsing of Japanese, we used the dependency structure analyzer CaboCha [28]. From the dependency structure, Japanese bracketed trees were generated. The combination of constraints in these experiments was the same as those of the E-J translation experiments.

Table 5 shows the following experimental results. In comparing the original ITG constraints (ITG) with the proposed IST-ITG (IST), BLEU was improved by 1.31 points, and by 3.81% in WER. The largest improvement was in WER, and BLEU had the next largest. This particular improvement order of these evaluation measures was the same as that of the E-J translation experiments. When Distance based distortion penalty was used at the same time (Dist+ITG and Dist+IST), there was no improvement in BLEU, but WER improved by 3.91%. When the lexical reordering model was used at the same time (Dist+LR and Dist+LR+IST), there was also no improvement in BLEU, but WER improved by 4.47%. This WER improvement is statistically significant. The result of sign test is larger than 6σ. One possible reason for the small (or no) improvement in BLEU is the lower parsing accuracy of Japanese compared with that of the English. Since, Japanese parsing results is obtained through dependency structure, is not obtained directly. However, better the WER figure indicates that using IST-ITC constraints leads to better word order. In the Appendix, differences in the translation results for the first five evaluation sentences between Dist+LR (Baseline:) and Dist+LR+IST (Proposed:) are shown. In the first evaluation sentence, one of the typical improvements is shown. In the baseline result, relation between "rotor 16" and "stator 15" is broken. On the other hand, relation is kept in the proposed result. In the proposed result, relation between "between" and "stator 15" is incorrect, since Japanese parsing result is also incorrect.

6.3 NIST MT08 English-to-Chinese Translation Experiments

Next, we conducted English-to-Chinese (E-C) newspaper translation experiments for different language pairs. The training and evaluation corpora were used in the NIST MT08 evaluation campaign [29] English-to-Chinese translation track. For the translation model training, we used 6.2 M bilingual sentences. For the language model training, we used 20.1 M sentences. A development set with 1,664 sentences was used as evaluation data in the Chinese-to-English translation track in the NIST MT07 evaluation campaign. A single reference was used in the development set. The evaluation set with 1,859 sentences is the same as MT08's evaluation data, with 4 references. Model training and decoding conditions were the same as those in the E-J experiments. In both base line and proposed condition, distance based distortion penalty and lexical reordering model were used at the same time. Therefore, the base line conditions correspond to the Dist+LR condition in the J-E experiments, the proposed conditions correspond to the Dist+LR+IST in the J-E experiments.

Table 6 shows the experimental results. The evaluation unit was both the Chinese character and word as defined by the PKU corpus (the MT08 official evaluation measure was character BLEU-4). As in the E-J experiments, the improvements in WER and CER (character error rate) were large (5.3% in WER, 6.2% in CER), next was BLEU (2.2-points in word, 1.8-points in character BLEU-4, 2.1-points in BLEU-5). We again demonstrated that the proposed method is effective (especially in WER) for multiple language pairs.

Table 6 Evaluation results in NIST MT08 E-C translation.

	Dist+LR	Dist+LR+IST
Word BLEU	21.0	23.2
Word NIST	7.43	7.56
Word Error Rate	75.0	69.7
Character BLEU-4	35.2	37.0
Character BLEU-5	28.2	30.3
Character NIST-4	7.96	8.11
Character Error Rate	74.1	67.9

7. Conclusion

We proposed new word reordering constraints for PBSMT using source tree structure. The proposed IST-ITG constraints are extensions of the ITG constraints. In ITG constraints, the instance of the source-side tree is not taken into account. On the other hand, in IST-ITG constraints, the tree that is obtained by source sentence parsing is imposed on the decoding process. Therefore, IST-ITG constraints are stronger than those of the original ITG. For example, for four-word source sentences, IST-ITG constraints allow eight word orderings in a target sentence compared with twenty-two orderings under the original ITG constraints. IST-ITG constraints can be applied to a common decoder to determine a target sentence from one-pass without re-scoring. In our E-J patent translation experiments, the proposed method resulted in a 2.7-point improvement in BLEU and a 5.7% improvement in WER compared with those of the original ITG constraints. In this paper we have argued the WER is the most appropriate measure to gauge the effectiveness of our approach since it gives importance to the global word order. Our approach gave rise to considerable gains in terms of WER in all of our experiments, indicating that a respectable improvement in global word order was achieved. The improvement could clearly be seen from visual inspection of the output, a few examples of which are presented in the following Appendix.

References

[1] D. Marcu and W. Wong, "A phrase-based, joint probability model for statistical machine translation," Proc. EMNLP-2002, pp.133–139, 2002.

[2] P. Koehn, F.J. Och, and D. Marcu, "Statistical phrase-base translation," Proc. HLT-NAACL, pp.127–133, 2003.

[3] F.J. Och and H. Ney, "The alignment template approach to stattistical machine translation," Computational Linguistics, vol.30, no.4, pp.417–449, 2004.

[4] C. Quirk, A. Menezes, and C. Cerry, "Dependeny treelet translation: Syntactically informed phrasal SMT," Proc. ACL, pp.271–279, 2005.

[5] Y. Liu, Q. Liu, and S. Lin, "Tree-to-string alignment template for statistical machine translation," Proc. ACL2006, pp.609–616, 2006.

[6] L. Huang, K. Knight, and A. Joshi, "Statistical syntax-directed translation with extended domain of locality," Proc. AMTA, 2006.

[7] K. Yamada and K. Knight, "A syntax-based statistical translation model," Proc. ACL, pp.523–530, 2000.

[8] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer, "Scalable inference and training of context-rich syntactic models," Proc. ACL-COLING, 2006.

[9] D. Marcu, W. Wang, A. Echihabi, and K. Knight, "SPMT: Statistical machine translation with syntactified target language phrases," Proc. EMNLP-2006, pp.44–52, 2006.

[10] D. Melamed, "Statistical machine translation by parsing," Proc. ACL, pp.653–660, 2004.

[11] Y. Ding and M. Palmer, "Machine translation using probabilistic synchronous dependency insert grammars," Proc. ACL, pp.541–548, 2005.

[12] A.L. Berger, P.F. Brown, S.A.D. Pietra, V.J.D. Pietra, J.R. Gillett, A.S. Kehler, and R.L. Mercer, "Language translation apparatus and method of using context-based translation models," United States

Patent, patent number 5510981, April 1996.
[13] C. Tillmann, "A unigram orientation model for statistical machine translation," HLT-NAACL, 2004.
[14] D. Wu, "Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora," Proc. IJCAI, pp.1328–1334, Montreal, Aug. 1995.
[15] D. Wu, "Stochastic inversion transuduction grammars and bilingual parsing of parallel corpora," Computational Linguiatics, vol.23, no.3, pp.377–403, 1997.
[16] R. Zens, H. Ney, T. Watanabe, and E. Sumita, "Reordering constraints for phrase-based statistical machine translation," Proc. Coling, pp.205–211, Geneva, Aug. 2004.
[17] Moses. http://www.statmt.org/moses/
[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," Proc. ACL, 2002.
[19] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," Proc. ARPA Workshop on Human Language Technology, 2002.
[20] M. Utiyama and H. Isahara, "Reliable measures for aligning Japanese-English news articles and sentences," ACL-2003, pp.72–79, 2003.
[21] NTCIR-7. http://ntcir.nii.ac.jp/
[22] F.J. Och and H. Ney, "A systematic comparison of various statistical alignment models," Computational Linguistics, vol.29, no.1, pp.19–51, 2003.
[23] A. Stolcke, "SRILM - An extensible language model toolkit," Proc. ICSLP'02, 2002. http://www.speech.sri.com/projects/srilm/
[24] R. Kneser and H. Ney, "Improved backing-off for m-gram language model," Proc. IEEE International Conference of Acoustic, Speech, and Signal processing, vol.1, pp.181–184, 1995.
[25] F.J. Och, "Minimum error rate training for statistical machine trainslation," Proc. ACL, 2003.
[26] E. Charniak, "A maximum-entropy-inspired parser," Proc. NAACL-2000, pp.132–139, 2000.
[27] Chasen. http://chasen-legacy.sourceforge.jp/
[28] Cabocha. http://chasen.org/ taku/software/cabocha/
[29] NIST MT08. http://www.nist.gov/speech/tests/mt/2008/

Appendix A: Samples from the Translation of English Patent into Japanese

Baseline: DIST+LR
Proposed: DIST+LR+IST

A.1 Sentence 1

Source: and, the kinetic energy of the liquid filled between the rotor 16 and stator 15 is converted into thermal energy to thereby produce a brake torque.
Reference: そして、ロータ１６とステータ１５との間に充填した液体の運動エネルギーが熱エネルギーに変換されて制動トルクを発生する。
Baseline: ステータ１５は、ロータ１６との間の運動エネルギーを熱エネルギーに変換する制動トルクを発生させると、液体充填されている。
Proposed: ステータ１５、回転子１６との間に充填液の運動エネルギーと、制動トルクを発生させるには熱エネルギーに変換される。

A.2 Sentence 2

Source: a sealant 7, which serves as a seal for cutting gas 9, also serves as a guide for the moving holder 3.
Reference: ７はシール材であり、後述の加工ガス９のシールと移動ホルダ３のガイドを兼ねたものである。
Baseline: シールの役目も兼ねているため、シール材７をガイドするための加工ガス９、可動保持部３を備えている。
Proposed: 封止剤７を加工ガス９シールの役目を兼ねて、可動ホルダ３をガイドする。

A.3 Sentence 3

Source: suppose that the red signal light of a traffic signal installed at a crossing situated ahead is on, the driver has recognized the red signal light, and the driver 's foot is about to shift from the accelerator pedal to the brake pedal to stop the vehicle.
Reference: 次に、車両前方の交差点の信号機が赤であり、運転者は信号機が赤であることを認知しており、停止しようとして、アクセルからブレーキに踏み換えようとしている場合を例として説明する。
Baseline: いま、交差位置が認識された場合には、運転者が足の前方に設置され、運転者のアクセルペダル、ブレーキペダルで車両を停止してからシフトし赤色信号光は赤色光信号、信号機程度である。
Proposed: 今、赤色信号光を前方に配置された交差点に設置された交通信号には、ドライバが認識された赤色信号光とし、運転者の足は、車両停止状態からブレーキペダルにアクセルペダルからのずれに程度である。

A.4 Sentence 4

Source: in addition, this method is not economical because it requires special steps such as pre-washing of the substrate surface, pre-treatments for providing the substrate with adherability to a coating, a drying step and the like.
Referenc: さらに、被塗装物の表面を予め洗浄し、塗膜が付着しやすいように前処理、乾燥等の工程を必要とし、経済的でない。
Baseline: また、この方法は、基板表面の洗浄等の工程を必要とするため、基板をに塗布、乾燥工程等の前処理、特別経済的でない。
Proposed: また、この方法では、基板表面の洗浄等の特別な工程の前基板を提供するために塗布、乾燥工程等が必要であるため経済的ではない。

A.5 Sentence 5

Sourcee: an oil passage 4 is formed as a hollow portion in the main body 1.
Referenc: ４は送油路を示し、本体１の中空部をもって構

成してある。
Baseline: オイル通路４は、本体１の中空部として形成されている。
Proposed: オイル通路４は、本体１の中空部として形成されている。

Appendix B: Samples from the Translation of Japanese Patent into English

B.1 Sentence 1

Source: そして、ロータ１６とステータ１５との間に充填した液体の運動エネルギーが熱エネルギーに変換されて制動トルクを発生する。
Reference: and, the kinetic energy of the liquid filled between the rotor 16 and stator 15 is converted into thermal energy to thereby produce a brake torque.
Baseline: then, the rotor 16 and the kinetic energy is converted to thermal energy braking torque is generated between the liquid filled in the stator 15.
Proposed: then, the rotor 16 and between the stator 15, the liquid filling the kinetic energy is converted to thermal energy braking torque is generated.

B.2 Sentence 2

Source: ７はシール材であり、後述の加工ガス９のシールと移動ホルダ３のガイドを兼ねたものである。
Reference: a sealant 7, which serves as a seal for cutting gas 9, also serves as a guide for the moving holder 3.
Baseline: the seal and movement of the holder 3 also serves as a guide for the seal member 7 is a work gas 9.
Proposed: 7 denotes a seal material, which also serves as a guide for the working gas 9 described later seal and movement of the holder 3.

B.3 Sentence 3

Source: 次に、車両前方の交差点の信号機が赤であり、運転者は信号機が赤であることを認知しており、停止しようとして、アクセルからブレーキに踏み換えようとしている場合を例として説明する。
Reference: suppose that the red signal light of a traffic signal installed at a crossing situated ahead is on, the driver has recognized the red signal light, and the driver ’s foot is about to shift from the accelerator pedal to the brake pedal to stop the vehicle.
Baseline: next , the tread brake by the driver, the accelerator to be stopped from the traffic of recognizing traffic signals is “red” and the intersection ahead of the vehicle is red, it is described as an example.
Proposed: next, a case will be exemplified below so as to tread brakes from the accelerator to be stopped, and of recognizing traffic signals of red, the driver is “red ” and is traffic light ahead of the vehicle.

B.4 Sentence 4

Source: さらに、被塗装物の表面を予め洗浄し、塗膜が付着しやすいように前処理、乾燥等の工程を必要とし、経済的でない。
Reference: in addition, this method is not economical because it requires special steps such as pre-washing of the substrate surface, pre-treatments for providing the substrate with adherability to a coating, a drying step and the like.
Baseline: further, the coating film is apt to be deposited on the surface of the object to be coated by washing and drying process is required, and the preliminary process advance not economical.
Proposed: further, to clean the surface of the object to be coated beforehand so as to facilitate the adhesion of the coating film preprocessing and drying process is required, and not economical.

B.5 Sentence 5

Source: ４は送油路を示し、本体１の中空部をもって構成してある。
Reference: an oil passage 4 is formed as a hollow portion in the main body 1.
Baseline: 4 is a hollow portion of the body 1 with an oil supply passage is shown.
Proposed: 4 is an oil supply passage, with a hollow portion of the main body 1.

Hirofumi Yamamoto received the M.S. degree in agriculture from the Tokyo University 1981 and the Ph.D. degree in global information and telecommunication from the Waseda University in 2004. Dr. Yamamoto is currently a professor at Kinki University School of Science and Engineering Dept. Informatics, short-term researcher at National Institute of Communications Technology, and cooperate researcher at ATR. His research interests include speech recognition and machine translation. He is a member of the IEEE, the ASJ and the ANLP.

Hideo Okuma received the B.S. degree in mathematics from Tokyo University of Science 1987. Mr. Okuma is currently a researcher at National Institute of Communications Technology. His research interests include machine translation.

Eiichiro Sumita received the M.S. degree in computer science from the University of Electro-Communications in 1982 and the Ph.D. degree in engineering from Kyoto University in 1999. Dr. Sumita is NLP department head of ATR/SLC, research manager of NiCT/KCCRC/SLCG, visiting professor of Kobe University and vice-president of ATR-Langue. His research interests include machine translation and e-Learning. He is a member of the IEEE, the ACL, the IPSJ, the ASJ and the ANLP. He serves Associate Editor of ACM/TSLP.