Distortion Model Based on Word Sequence Labeling for Statistical Machine Translation

ISAO GOTO, National Institute of Information and Communications Technology and Kyoto University
MASAO UTIYAMA, EIICHIRO SUMITA, and AKIHIRO TAMURA, National Institute of Information and Communications Technology
SADAO KUROHASHI, Kyoto University

This article proposes a new distortion model for phrase-based statistical machine translation. In decoding, a distortion model estimates the source word position to be translated next (subsequent position; SP) given the last translated source word position (current position; CP). We propose a distortion model that can simultaneously consider the word at the CP, the word at an SP candidate, the context of the CP and an SP candidate, relative word order among the SP candidates, and the words between the CP and an SP candidate. These considered elements are called *rich context*. Our model considers rich context by discriminating label sequences that specify spans from the CP to each SP candidate. It enables our model to learn the effect of relative word order among SP candidates as well as to learn the effect of distances from the training data. In contrast to the learning strategy of existing methods, our learning strategy is that the model learns preference relations among SP candidates in each sentence of the training data. This leaning strategy enables consideration of all of the rich context simultaneously. In our experiments, our model had higher BLUE and RIBES scores for Japanese-English, Chinese-English, and German-English translation compared to the lexical reordering models.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Machine translation*

General Terms: Theory, Algorithms, Design, Experimentation

Additional Key Words and Phrases: Distortion model, machine translation, reordering

ACM Reference Format:
Goto, I., Utiyama, M., Sumita, E., Tamura, A., and Kurohashi, S. 2014. Distortion model based on word sequence labeling for statistical machine translation. *ACM Trans. Asian Lang. Inform. Process.* 13, 1, Article 2 (February 2014), 21 pages.
DOI: http://dx.doi.org/10.1145/2537128

1. INTRODUCTION

Estimating appropriate word order in a target language is one of the most difficult problems for statistical machine translation (SMT). This is particularly true when translating between languages with widely different word orders.

To address this problem, there has been a lot of research done into word reordering: lexical reordering model [Tillman 2004], which is one of the distortion models, reordering constraints [Zens et al. 2004], pre-ordering [Xia and McCord 2004],

I. Goto is currently affiliated with NHK Science & Technology Research Laboratories and Kyoto University. Authors' addresses: I. Goto, NHK Science & Technology Research Laboratories, 1-10-11 Kinuta, Setagaya-ku, Tokyo 157-8510, Japan; email: goto.i-es@nhk.or.jp; M. Utiyama, E. Sumita, and A. Tamura, Multilingual Translation Laboratory, National Institute of Information and Communications Technology, 3-5 Hikaridai, Keihanna Science City, Kyoto 619-0289, Japan; emails: {mutiyama, eiichiro.sumita, akihiro.tamura}@nict.go.jp; S. Kurohashi, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan; email: kuro@i.kyoto-u.ac.jp.

1530-0226/2014/02-ART2
DOI: http://dx.doi.org/10.1145/2537128

hierarchical phrase-based SMT [Chiang 2007], and syntax-based SMT [Yamada and Knight 2001].

In general, source language syntax is useful for handling long distance word reordering. However, obtaining syntax requires a syntactic parser, which is not available for many languages. Phrase-based SMT [Koehn et al. 2007] is a widely used SMT method that does not use a parser.

Phrase-based SMT mainly[1] estimates word reordering using distortion models.[2] Therefore, distortion models are one of the most important components for phrase-based SMT. There are methods other than distortion models for improving word reordering for phrase-based SMT, such as pre-ordering or reordering constraints. However, these methods also use distortion models when translating by phrase-based SMT. Therefore, distortion models do not compete against these methods and are commonly used with them. If a distortion model improves, it will improve the translation quality of phrase-based SMT and will benefit the methods using distortion models.

In decoding by phrase-based SMT, a distortion model estimates the source word position to be translated next (SP) given the last translated source word position (CP). In order to estimate the SP given the CP, many elements need to be considered: the word at the CP, the word at an SP candidate (SPC), the words surrounding the CP and an SPC (context), the relative word order among the SPCs, and the words between the CP and an SPC. In this article, these elements are called *rich context*. The major challenge of distortion modeling is consideration of all of the rich context.

Previous distortion models could not consider all of the rich context simultaneously. This is because the learning strategy for existing methods was that the models learned probabilities in all of the training data. This meant that the models did not learn preference relations among SPCs in each sentence of the training data. Consequently, it is hard to consider all of the rich context simultaneously using this learning strategy. The MSD lexical reordering model [Tillman 2004] and a discriminative distortion model [Green et al. 2010] could not simultaneously consider both the word specified at the CP and the word specified at an SPC, or consider relative word order. There is a distortion model that used the word at the CP and the word at an SPC [Al-Onaizan and Papineni 2006], but this model did not use context, relative word order, or words between the CP and an SPC. All of these elements are important, and the reasons for their importance will be detailed in Section 2.

In this article[3], we propose a new distortion model consisting of one probabilistic model and which does not require a parser for phrase-based SMT. In contrast to the learning strategy of existing methods, our learning strategy is that the model learns preference relations among SPCs in each sentence of the training data. This leaning strategy enables consideration of all of the rich context simultaneously. Our proposed model, *the sequence model*, can simultaneously consider all of the rich context by identifying the label sequence that specifies the span from the CP to the SP. It enables our model to learn the effect of relative word order among the SPCs as well as learn the effect of distances from the training data. Experiments confirmed the effectiveness of

[1]A language model also supports estimation.

[2]In this article, reordering models for phrase-based SMT, which are intended to estimate the source word position to be translated next in decoding, are called distortion models. This estimation is used to produce a hypothesis in the target language word order sequentially from left to right.

[3]This article is based on a presentation given at the ACL 2013 conference [Goto et al. 2013]. Additional material includes experiments on Chinese-English translation using an NIST dataset and on German-English translation using the Europarl corpus; evaluations and analyses based on RIBES; and investigation of the effects of context, the effects of part of speech, the relation between data sizes and the translation quality, and the relation between distortion limits and translation quality without the effects of differences in the SMT weighting parameters.

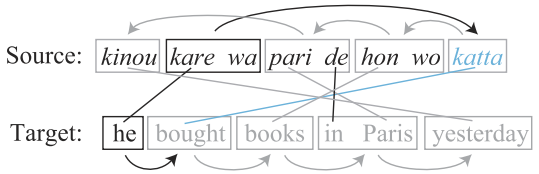

Fig. 1. An example of left-to-right translation for Japanese-English. Boxes represent phrases and arrows indicate the translation order of the phrases.

our method for Japanese-English, Chinese-English, and German-English translation using NTCIR-9 Patent Machine Translation Task data [Goto et al. 2011], NIST 2008 Open MT task data, and WMT 2008 Europarl data [Callison-Burch et al. 2008].

2. DISTORTION MODELS FOR PHRASE-BASED SMT

A Moses-style phrase-based SMT [Koehn et al. 2007] generates target hypotheses sequentially from left to right. Therefore, the role of the distortion model is to estimate the source phrase position to be translated next whose target side phrase will be located immediately to the right of the already generated hypotheses. An example is shown in Figure 1. In Figure 1, we assume that only the *kare wa* (English: "he") has been translated. The target word to be generated next will be "bought", and the source word to be selected next will be its corresponding Japanese word *katta*. Thus, a distortion model should estimate phrases including *katta* as a source phrase position to be translated next.

To explain the distortion model task in more detail, we need to redefine two terms more precisely, the *current position* (CP) and *subsequent position* (SP) in the source sentence. CP is the source sentence position corresponding to the rightmost aligned target word in the generated target word sequence. SP is the source sentence position corresponding to the leftmost aligned target word in the target phrase to be generated next. The task of the distortion model is to estimate the SP[4] from SP candidates (SPCs) for each CP.[5]

It is difficult to estimate the SP. Figure 2 shows examples of sentences that are similar yet have different SPs, with the superscript numbers indicating the word position in the source sentence.

In Figure 2(a), the SP is 8. However, in 2(b), the word (*kare*) at the CP is the same as 2(a), but the SP is different (the SP is 10). From these example sentences, we see that distance is not the essential factor in deciding an SP. We can also see that the word at the CP alone is not enough to estimate the SP. Thus, it is not only the word at the CP, but also the word at an SP candidate (SPC) that should be considered simultaneously.

In Figures 2(c) and 2(d), the word (*kare*) at the CP is the same and *karita* (borrowed) and *katta* (bought) are at the SPCs. *Karita* is the word at the SP for 2(c), while *katta*, not *karita*, is the word at the SP for 2(d). One of the reasons for this difference is the relative word order between words. Thus, we can see that considering relative

[4]SP is not always one position, because there may be multiple correct hypotheses.

[5]This definition is slightly different from that of existing methods, such as Moses [Koehn et al. 2007] and Green et al. [2010]. In existing methods, CP is the rightmost position of the last translated source phrase and SP is the leftmost position of the source phrase to be translated next. Note that existing methods do not consider word-level correspondences.

Fig. 2. Examples of CP and SP for Japanese-English translation. The upper sentence is the source sentence and the sentence underneath is a target hypothesis for each example. The SP is in bold, and the CP is in bold italics. The point of an arrow with an × mark indicates a wrong SP candidate.

word order, not just looking at what the word at the SP is, is important for estimating the SP.[6]

In Figures 2(d) and 2(e), *kare* (he) is at the CP for both, and the word order between *katta* and *karita* are the same. However, the word at the SP for 2(d) and the word at the SP for 2(e) are different, which shows us that selecting a nearby word is not always correct. The difference is caused by the words surrounding the SPCs (context), the CP context, and the words between the CP and the SPC. Thus, these should all be considered when estimating the SP.

In order to estimate the SP, the following should be considered simultaneously: the word at the CP, the word at an SPC, the relative word order among the SPCs, the

[6]We checked the probability of a relatively close word position being the SP by using the NTCIR-9 JE data [Goto et al. 2011]. We made lists of words at the SP for each word at the CP in the training data. When a sentence contains two or more words that are included in the list for each word at the CP, and their orientations are the same as that of the SP, we extracted those word pairs from these words. For example, when Figures 2(c) and 2(d) are the training data, the list of words at the SP for *kare* at the CP consists of *karita* and *katta*. We extract $karita^6$ and $katta^{10}$ as the word pair from Figure 2(c), and extract $katta^6$ and $karita^{10}$ as the word pair from Figure 2(d). For Figure 2(c), the word position relatively close to the CP in the extracted pair is 6 ($karita^6$). The probability of a word position relatively close to the CP in the extracted pairs being the SP was 81.2%.

words surrounding the CP and an SPC (context), and the words between the CP and an SPC. In other words, rich context should be considered simultaneously.

Returning back to the distribution models, there are distortion models that do not require a parser for phrase-based SMT. The linear distortion cost model used in Moses [Koehn et al. 2007], whose costs are linearly proportional to the reordering distance, always gives a high cost to long distance reordering, even if the reordering is correct. The MSD lexical reordering model [Tillman 2004; Koehn et al. 2005; Galley and Manning 2008] only calculates probabilities for the three types of phrase reorderings (monotone, swap, and discontinuous), and does not consider relative word order or words between the CP and an SPC. Thus, these models are not sufficient for long-distance word reordering.

Xiong et al. [2006] proposed distortion models that used context to predict the orientations {left, right} of the SP for their CYK-style decoder. Zens and Ney [2006] proposed distortion models that used context to predict four classes {left, right}×{continuous, discontinuous}. Green et al. [2010] extended the distortion models to use finer classes. Green et al.'s [2010] model (the outbound model) estimates how far the SP should be from the CP using the word at the CP and its context.[7] Feng et al. [2013] also predicted those finer classes using a CRF model. These models do not simultaneously consider both the word specified at the CP and the word specified at an SPC, nor do they consider relative word order.

Al-Onaizan and Papineni [2006] proposed a distortion model that used the word at the CP and the word at an SPC. However, their model did not use context, relative word order, or words between the CP and an SPC.

There is a method that adjusts the linear distortion cost using the word at the CP and its context [Ni et al. 2009]. This model does not simultaneously consider both the word specified at the CP and the word specified at an SPC.

In contrast, our distortion model, the sequence model, addresses the aforementioned issues and utilizes all of the rich context.

3. PROPOSED METHOD

In this section, we first define our distortion model and explain our learning strategy. Then, we describe two models: *the pair model* and *the sequence model*. The pair model is our base model and the sequence model is our main proposed model.

3.1. Distortion Model and Learning Strategy

Our distortion model is defined as the model calculating the distortion probability. In this article, *distortion probability* is defined as

$$P(X = j|i, S), \tag{1}$$

which is the probability of j being the SP, where i is a CP, j is an SPC, S is a source sentence, and X is the random variable of the SP.

[7]They also proposed another model (the inbound model) that estimates reverse direction distance. Each SPC is regarded as an SP, and the inbound model estimates how far the corresponding CP should be from the SP using the word at the SP and its context.

We train this model as a discriminative model that discriminates the SP from SPCs. Let J be a set of word positions in S other than i. We train the distortion model subject to

$$\sum_{j \in J} P(X = j|i, S) = 1.$$

The model parameters are learned to maximize the distortion probability of the SP among all of the SPCs J in each source sentence. This learning strategy is a type of preference relation learning [Evgeniou and Pontil 2002]. In this learning, the distortion probability of the actual SP will be relatively higher than those of all the other SPCs J.

This learning strategy is different from that of Al-Onaizan and Papineni [2006] and Green et al. [2010]. Green et al. [2010], for example, trained their outbound model subject to $\sum_{c \in C} P(Y = c|i, S) = 1$, where C is a set of nine distortion classes[8] and Y is the random variable of the correct distortion class that the correct distortion is classified into. Distortion is defined as $j - i - 1$. Namely, the model probabilities that they learned were the probabilities of distortion classes in all of the training data, not the relative preferences among the SPCs in each source sentence.

3.2. Pair Model

The pair model, which is our base model, utilizes the word at the CP, the word at an SPC, and the context of the CP and the SPC simultaneously to estimate the SP. This can be done using our distortion model definition and the learning strategy described in the previous section.

In this work, we use the maximum entropy method [Berger et al. 1996] as a discriminative machine learning method. The reason for this is that a model based on the maximum entropy method can calculate probabilities. However, if we use scores as an approximation of the distortion probabilities, various discriminative machine learning methods can be applied to build the distortion model.

Let s be a source word and $s_1^n = s_1 s_2 ... s_n$ be a source sentence. We add a beginning of sentence (BOS) marker to the head of the source sentence and an end of sentence (EOS) marker to the end, so the source sentence S is expressed as s_0^{n+1} ($s_0 =$ BOS, $s_{n+1} =$ EOS). Our distortion model calculates the distortion probability for an SPC $j \in \{j | 1 \leq j \leq n+1 \ \wedge \ j \neq i\}$ for each CP $i \in \{i | 0 \leq i \leq n\}$

$$P(X = j|i, S) = \frac{1}{Z_i} \exp\left(\mathbf{w}^{\mathrm{T}} \mathbf{f}(i, j, S, o, d)\right), \tag{2}$$

where

$$o = \begin{cases} 0 & (i < j) \\ 1 & (i > j) \end{cases},$$

$$d = \begin{cases} 0 & (|j - i| = 1) \\ 1 & (2 \leq |j - i| \leq 5) \\ 2 & (6 \leq |j - i|) \end{cases},$$

$$Z_i = \sum_{j \in \{j | 1 \leq j \leq n+1 \ \wedge \ j \neq i\}} \exp\left(\mathbf{w}^{\mathrm{T}} \mathbf{f}(i, j, S, o, d)\right),$$

[8] $(-\infty, -8]$, $[-7, -5]$, $[-4, -3]$, -2, 0, 1, $[2, 3]$, $[4, 6]$, and $[7, \infty)$. In Green et al. [2010], -1 was used as one of distortion classes. However, -1 represents the CP in our definition, and CP is not an SPC. Thus, we shifted all of the distortion classes for negative distortions by -1.

Table I. Feature Templates

Template
$\langle o\rangle$, $\langle o, s_{i+p}\rangle^1$, $\langle o, s_{j+p}\rangle^1$, $\langle o, t_i\rangle$, $\langle o, t_j\rangle$, $\langle o, d\rangle$, $\langle o, s_{i+p}, s_{j+q}\rangle^2$, $\langle o, t_i, t_j\rangle$, $\langle o, t_{i-1}, t_i, t_j\rangle$, $\langle o, t_i, t_{i+1}, t_j\rangle$, $\langle o, t_i, t_{j-1}, t_j\rangle$, $\langle o, t_i, t_j, t_{j+1}\rangle$, $\langle o, s_i, t_i, t_j\rangle$, $\langle o, s_j, t_i, t_j\rangle$

Note: t is the Part of Speech for s.
[1] $p \in \{p| -2 \le p \le 2\}$
[2] $(p,q) \in \{(p,q)| -2 \le p \le 2 \ \wedge \ -2 \le q \le 2 \ \wedge \ (|p| \le 1 \ \vee \ |q| \le 1)\}$

Table II. The "C, I, and S" Label Set

Label	Description
C	The current position (CP).
I	A position between the CP and an SPC.
S	A subsequent position candidate (SPC).

$\mathbf{w}$ is a weight parameter vector, and each element of $\mathbf{f}(\cdot)$ is a binary feature function which returns 1 when its feature is matched and if else, returns 0. Z_i is a normalization factor, o is an orientation of i to j, and d is a distance class.

Table I shows the feature templates used to produce features. A feature is defined as an instance of a feature template. Using example (a) from Figure 2 will show some instances of each variable, where $i = 2$ and $j = 8$: $o = 1$, $s_i = kare$, $s_{i+1} = wa$, $s_j = katta$, $t_i =$ NOUN, and $d = 2$. t is the part of speech for s. In this case, a feature of $\langle o, s_i, s_j\rangle$ is $\langle o = 1, s_i = kare, s_j = katta\rangle$ and a feature of $\langle o, s_{i+1}, s_j\rangle$ is $\langle o = 1, s_{i+1} = wa, s_j = katta\rangle$.

In Equation (2), i, j, and S are used by the feature functions. Thus, Equation (2) can utilize features consisting of both s_i, which is the word specified at i, and s_j, which is the word specified at j, or both the context of i and the context of j simultaneously. Distance is considered using the distance class d. Distortion is represented by distance and orientation. The pair model considers distortion using six joint classes of d and o.

3.3. Sequence Model

The pair model does not consider relative word order among the SPCs nor all the words between the CP and an SPC. Our main proposed model, *the sequence model*, which is described in this section, considers rich context, including relative word order among the SPCs and including all the words between the CP and an SPC.

In Figures 2(c) and 2(d), *karita* (borrowed) and *katta* (bought) both occur in the source sentences. The pair model considers the effect of distances using only the distance class d. If these positions are in the same distance class, the pair model cannot consider the differences in distances. In this case, these are conflict instances during training and it is difficult to distinguish the SP for translation. However, this problem can be solved if the model can consider the relative word order.

The sequence model considers the relative word order. It does this by discriminating the label sequence corresponding to the SP from the label sequences corresponding to each SPC in each sentence. Since each label sequence corresponds to one SPC, if we can identify the label sequence that corresponds to the SP, then we can obtain the SP. The label sequences specify the spans from the CP to each SPC using three kinds of labels that indicate the type of word positions in the spans. The three kinds of labels, "C, I, and S," are shown in Table II. Figure 3 shows examples of the label sequences for Figure 2(c). The label sequences are represented by boxes and the elements of the sequences are labels. The SPC is used as the label sequence ID for each label sequence.

The label sequence can handle relative word order. Looking at Figure 3, the label sequence ID of 10 knows that *karita* exists to the left of the SPC of 10. This is because

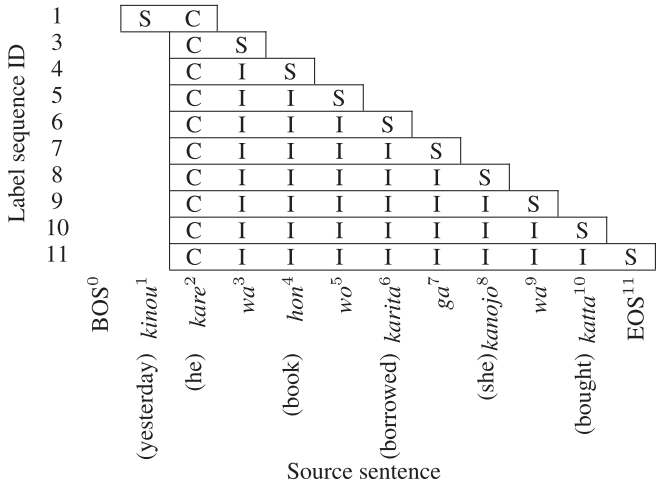

Fig. 3. Example of label sequences that specify spans from the CP to each SPC for the case of Figure 2(c). The labels (C, I, and S) in the boxes are the label sequences.

karita^6 carries a label I, while *katta*^{10} carries a label S, and a position with label I is defined as relatively closer to the CP than a position with label S. By utilizing the label sequence and corresponding words, the model can reflect the effect of *karita* existing between the CP and the SPC of 10 on the probability.

Karita (borrowed) and *katta* (bought) in Figures 2(c) and 2(d) are not conflict instances in training for the sequence model, whereas they are conflict instances in training for the pair model. The reason is because it is necessary to make the probability of the SPC of 10 smaller than that of the SPC of 6. The pair model tries to make the weight parameters for features with respect to *katta* smaller than those for features with respect to *karita* for 2(c), but it also tries to make the weight parameters for features with respect to *karita* smaller than those for features with respect to *katta* for 2(d). Since they have the same features, this causes a conflict. In contrast, the sequence model can give negative weight parameters for the features with respect to the word at the position of 6 with label I, instead of making the weight parameters for the features with respect to the word at the position of 10 with label S smaller than those of 6 with label S.

We use a sequence discrimination technique based on CRF [Lafferty et al. 2001] to identify the label sequence that corresponds to the SP.[9] There are two differences between our task and the CRF task. One difference is that CRF identifies label sequences that consist of labels from all of the label candidates, whereas we constrain the label sequences to sequences where the label at the CP is C, the label at an SPC is S, and the labels between the CP and the SPC are I. The other difference is that CRF is designed for discriminating label sequences corresponding to the same object sequence, whereas we do not assign labels to words outside the spans from the CP to each SPC. However, when we assume that another label such as E has been assigned to the words outside the spans and there are no features involving label E, CRF with our label constraints

[9]The critical difference between CRFs and maximum entropy Markov models is that a maximum entropy Markov model uses per-state exponential models for the conditional probabilities of next states given the current state, while a CRF has a single exponential model for the joint probability of the entire sequence of labels given the observation sequence [Lafferty et al. 2001].

can be applied to our task. In this article, the method designed to discriminate label sequences corresponding to the different word sequence lengths is called *partial CRF*.

The sequence model based on partial CRF is derived by extending the pair model. We introduce the label l and add two extensions to the pair model to identify the label sequences corresponding to the SP. One of the extensions uses labels and the other uses sequence. For the extension using labels, we suppose that label sequences specify the spans from the CP to each SPC using the labels in Table II. We conjoin all the feature templates in Table I with an additional feature template $\langle l_i, l_j \rangle$ to include the labels into features, where l_i is the label corresponding to the position of i. For example, a feature template of $\langle o, s_{i+1}, s_j, l_i, l_j \rangle$ is derived by conjoining $\langle o, s_{i+1}, s_j \rangle$ in Table I with $\langle l_i, l_j \rangle$. The other extension uses sequence. In the pair model, the position pair of (i, j) is used to derive features. In contrast, to discriminate label sequences in the sequence model, the position pairs of (i, k), $k \in \{k | i < k \leq j \ \vee \ j \leq k < i\}$ and (k, j), $k \in \{k | i \leq k < j \ \vee \ j < k \leq i\}$ are used to derive features. Note that in the feature templates in Table I, i and j are used to specify two positions. When features are used for the sequence model, a value of k is used as one of the two positions. For example, for the position pairs of (i, k), the value of s_k is used as the value of s_j and the value of l_k is used as the value of l_j in the feature template of $\langle o, s_{i+1}, s_j, l_i, l_j \rangle$ to obtain a feature for each k. This is conducted by interpreting the parameters of $\mathbf{f}(\cdot)$ as $\mathbf{f}(i, j, S, o, d, l_i, l_j)$ when the feature templates are used to derive features in the following Equations (3) and (4).

The distortion probability for an SPC j being the SP given a CP i and a source sentence S is calculated as

$$P(X = j | i, S) = \frac{1}{Z_i} \exp\Bigg(\sum_{k \in M \cup \{j\}} \mathbf{w}^{\mathrm{T}} \mathbf{f}(i, k, S, o, d, l_i, l_k) + \sum_{k \in M \cup \{i\}} \mathbf{w}^{\mathrm{T}} \mathbf{f}(k, j, S, o, d, l_k, l_j) \Bigg), \tag{3}$$

where

$$M = \begin{cases} \{m | i < m < j\} & (i < j), \\ \{m | j < m < i\} & (i > j), \end{cases}$$

and

$$Z_i = \sum_{j \in \{j | 1 \leq j \leq n+1 \ \wedge \ j \neq i\}} \exp\Bigg(\sum_{k \in M \cup \{j\}} \mathbf{w}^{\mathrm{T}} \mathbf{f}(i, k, S, o, d, l_i, l_k) + \sum_{k \in M \cup \{i\}} \mathbf{w}^{\mathrm{T}} \mathbf{f}(k, j, S, o, d, l_k, l_j) \Bigg). \tag{4}$$

Since j is used as the label sequence ID, discriminating $X = j$ from $X \neq j$ also means discriminating the label sequence ID of the SP from the label sequence IDs of the non-SPs.

The first term in $\exp(\cdot)$ in Equation (3) considers all of the word pairs located at i and other positions in the sequence, and also their context. The second term in $\exp(\cdot)$ in Equation (3) considers all of the word pairs located at j and other positions in the sequence, and also their context.

By designing our model to discriminate among different length label sequences, our model can naturally handle the effect of distances. Many features are derived from a long label sequence because it will contain many labels between the CP and the SPC. On the other hand, fewer features are derived from a short label sequence because a

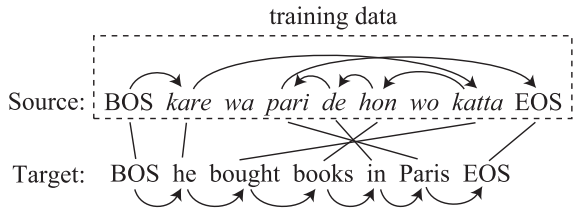

Fig. 4. Examples of supervised training data. The lines represent word alignments between source words and target words. The English side arrows point to the nearest word aligned on the right.

short label sequence will contain fewer labels between the CP and the SPC. The bias from these differences provides important clues for learning the effect of distances.[10]

3.4. Training Data for Discriminative Distortion Model

In order to train our discriminative distortion model, supervised training data built from a parallel corpus and word alignments between corresponding source words and target words is necessary. Figure 4 shows examples of this training data. We create the training data by selecting the target words aligned to the source words sequentially from left to right (target side arrows), then deciding on the order of the source words in the target word order (source side arrows). The source sentence and the source side arrows are the training data.

4. EXPERIMENT

In order to confirm the effects of our distortion model, we conducted a series of Japanese to English (JE), Chinese to English (CE), and German to English (GE) translation experiments.[11]

4.1. Data

We used the patent data from the NTCIR-9 Patent Machine Translation Task [Goto et al. 2011] for JE and CE translation. There were 2,000 sentences for the test data and 2,000 sentences for the development data. The reference data is single reference. The translation model was trained using sentences of 40 words or less from the training data. So approximately 2.05 million sentence pairs consisting of approximately 54 million Japanese tokens whose lexicon size was 134k and 50 million English tokens whose lexicon size was 213k were used for JE. Approximately 0.49 million sentence pairs consisting of 14.9 million Chinese tokens whose lexicon size was 169k and 16.3 million English tokens whose lexicon size was 240k were used for CE.

We also used the newswire data from the NIST 2008 Open MT task[12] for CE translation. There were 1,357 sentences for the test data. The reference data is multi-reference (4 references). We used the NIST 2006 test set consisting of 1,664 test sentences as the development data. The translation model was trained using sentences of

[10]Note that the sequence model does not only consider larger context than the pair model, but that it also considers labels. The pair model does not discriminate labels, whereas the sequence model uses label S and label I for the positions except for the CP, depending on each situation. For example, in Figure 3, at position 6, label S is used in the label sequence ID of 6, but label I is used in the label sequence IDs of 7 to 11. Namely, even if they are at the same position, the labels in the label sequences are different. The sequence model discriminates the label differences.

[11]We conducted JE, CE, and GE translation as examples of language pairs with different word orders and of languages where there is a great need for translation into English.

[12]To reduce the computational cost, we did not use the comparable corpus (LDC2007T09), the UN corpus (LDC2004E12), or hansard and law domains in the Hong Kong corpus (LDC2004T08).

40 words or less from the training data. So approximately 2.19 million sentence pairs[13] consisting of 18.4 million Chinese tokens whose lexicon size was 907k and 20.7 million English tokens whose lexicon size was 932k were used.

We used the Europarl data from the WMT 2008 [Callison-Burch et al. 2008] translation task for GE translation. There were 2,000 sentences for the test data. The reference data is single reference. We used the WMT 2007 test set consisting of 2,000 test sentences as the development data. The translation model was trained using sentences of 40 words or less from the training data. So approximately 1.00 million sentence pairs consisting of 20.4 million German tokens whose lexicon size was 226k and 21.4 million English tokens whose lexicon size was 87k were used.

4.2. Common Settings

MeCab[14] was used for the Japanese morphological analysis. We adjusted the tokenization of the alphanumeric characters in Japanese to be the same as for the English. The Stanford segmenter[15] and tagger[16] were used for Chinese segmentation and POS tagging and for German POS tagging. GIZA++ and grow-diag-final-and heuristics were used to obtain word alignments. In order to reduce word alignment errors, we removed articles {a, an, the} in English, particles {*ga*, *wo*, *wa*} in Japanese, and articles {der, die, das, des, dem, den, ein, eine, eines, einer, einem, einen} in German before performing word alignments because these function words do not correspond to any words in the other languages (JE and CE) or articles do not always correspond like content words or prepositional words (GE). After word alignment, we restored the removed words and shifted the word alignment positions to the original word positions. We used 5-gram language models with modified Kneser-Ney discounting [Chen and Goodman 1998] using SRILM [Stolcke et al. 2011]. The language models were trained using the English side of each set of bilingual training data.

We used an in-house standard phrase-based SMT system compatible with the Moses decoder [Koehn et al. 2007]. The phrase table and the lexical distortion model were built using the Moses tool kit. The SMT weighting parameters were tuned by MERT [Och 2003] using the development data. The tuning was based on the BLEU score [Papineni et al. 2002]. To stabilize the MERT results, we tuned the parameters three times by MERT using the first half of the development data and we selected the SMT weighting parameter set that performed the best on the second half of the development data based on the BLEU scores from the three SMT weighting parameter sets.

We compared systems that used a common SMT feature set from standard SMT features and different distortion model features. The common SMT feature set consists of four translation model features, phrase penalty, word penalty, and a language model feature. The compared different distortion model features are as follows.

— The linear distortion cost model feature (LINEAR)
— The linear distortion cost model feature and the six MSD bidirectional lexical distortion model [Koehn et al. 2005] features (LINEAR+LEX)
— The outbound and inbound distortion model features discriminating nine distortion classes [Green et al. 2010] (9-CLASS)

[13] 1.27 million sentence pairs were from a lexicon (LDC2002L27) and a Named Entity list (LDC2005T34).
[14] http://mecab.sourceforge.net/
[15] http://nlp.stanford.edu/software/segmenter.shtml
[16] http://nlp.stanford.edu/software/tagger.shtml

— The proposed pair model feature (PAIR)
— The proposed sequence model feature (SEQUENCE)

4.3. Training for the Proposed Models

Our distortion model was trained as follows: We used 0.2 million sentence pairs and their word alignments from the data used to build the translation model as the training data for our distortion models. The features that were selected and used were the ones that had been counted[17], using the feature templates in Table I, at least four times for all of the (i, j) position pairs in the training sentences. We conjoined the features with three types of label pairs $\langle l_i = \mathrm{C}, l_j = \mathrm{I}\rangle$, $\langle l_i = \mathrm{I}, l_j = \mathrm{S}\rangle$, or $\langle l_i = \mathrm{C}, l_j = \mathrm{S}\rangle$ to produce features for SEQUENCE. The L-BFGS method [Liu and Nocedal 1989] was used to estimate the weight parameters of maximum entropy models. The Gaussian prior [Chen and Rosenfeld 1999] was used for smoothing.[18]

4.4. Training for the Compared Models

For 9-CLASS, we used the same training data as for our distortion models. We used the following feature templates to produce features for the outbound model: $\langle s_{i-2}\rangle$, $\langle s_{i-1}\rangle$, $\langle s_i\rangle$, $\langle s_{i+1}\rangle$, $\langle s_{i+2}\rangle$, $\langle t_i\rangle$, $\langle t_{i-1}, t_i\rangle$, $\langle t_i, t_{i+1}\rangle$, and $\langle s_i, t_i\rangle$, where t_i is the part of speech for s_i. These feature templates correspond to the components of the feature templates of our distortion models. In addition to these features, we used a feature consisting of the relative source sentence position as the feature used by Green et al. [2010]. The relative source sentence position is discretized into five bins, one for each quintile of the sentence. For the inbound model[19], i of the feature templates was changed to j. Features occurring four or more times in the training sentences were used. The maximum entropy method with Gaussian prior smoothing was used to estimate the model parameters.

The MSD bidirectional lexical distortion model was built using all of the data used to build the translation model.

4.5. Results and Discussion

We evaluated translation quality based on the case-insensitive automatic evaluation score BLEU-4 [Papineni et al. 2002] and RIBES v1.01 [Isozaki et al. 2010a]. RIBES is an automatic evaluation measure based on word order correlation coefficients between reference sentences and translation outputs. We used distortion limits of 10, 20, 30, and unlimited (∞), which limited the number of words for word reordering to a maximum number for JE and CE. We used distortion limits of 6, 10, and 20 for GE. Our main results are presented in Tables III to VI. The values given are case-insensitive scores. Bold numbers indicate no significant difference from the best result in each language pair and in each evaluation measure using the bootstrap resampling test at a significance level $\alpha = 0.01$ [Koehn 2004].

[17]When we counted features for selection, we counted features that were from all of the feature templates in Table I when j was the SP, but we only counted features that were from the feature templates of $\langle s_i, s_j\rangle$, $\langle t_i, t_j\rangle$, $\langle s_i, t_i, t_j\rangle$, and $\langle s_j, t_i, t_j\rangle$ in Table I when j was not the SP, in order to avoid increasing the number of features.

[18]Let $\mathcal{L}_{\mathbf{w}}$ be the log likelihood of the training data, $\arg\max_{\mathbf{w}}(\mathcal{L}_{\mathbf{w}} - \frac{1}{2\sigma^2}\mathbf{w}^T\mathbf{w})$ is used to estimate $\mathbf{w}$. $\sigma^2 = 0.01$ was used for all of the experiments.

[19]The inbound model is explained in footnote 7.

Table III. Japanese-English Translation Evaluation Results for NTCIR-9 Data

	BLEU				RIBES			
Distortion limit	10	20	30	∞	10	20	30	∞
LINEAR	27.98	27.74	27.75	27.30	67.10	67.00	65.89	63.53
LINEAR+LEX	30.25	30.37	30.17	29.98	68.62	68.33	67.31	64.56
9-CLASS	30.74	30.98	30.92	30.75	70.43	69.11	67.97	65.60
PAIR	31.62	32.36	31.96	32.03	70.71	72.04	70.14	68.19
SEQUENCE	32.02	32.96	**33.29**	32.81	71.14	**72.78**	**72.86**	70.55

Table IV. Chinese-English Translation Evaluation Results for NTCIR-9 Data

	BLEU				RIBES			
Distortion limit	10	20	30	∞	10	20	30	∞
LINEAR	29.18	28.74	28.31	28.33	75.24	73.46	72.27	71.27
LINEAR+LEX	30.81	30.24	30.16	30.13	75.68	73.54	71.58	70.20
9-CLASS	31.80	31.56	31.31	30.84	77.05	74.43	72.92	71.30
PAIR	32.51	32.30	32.25	32.32	77.75	76.14	74.75	73.93
SEQUENCE	**33.41**	**33.44**	**33.35**	**33.41**	**78.57**	77.67	77.15	76.64

Table V. Chinese-English Translation Evaluation Results for NIST 2008 Data

	BLEU				RIBES			
Distortion limit	10	20	30	∞	10	20	30	∞
LINEAR	22.50	21.98	21.92	22.09	74.41	71.85	70.92	69.61
LINEAR+LEX	23.29	22.53	23.14	22.85	75.00	72.24	70.67	71.17
9-CLASS	23.30	23.16	22.89	22.98	75.28	73.47	70.26	69.51
PAIR	24.25	23.53	23.87	23.63	**75.88**	73.43	71.20	69.97
SEQUENCE	**24.67**	**24.47**	24.18	**24.34**	**75.92**	73.75	72.42	72.40

The proposed SEQUENCE outperformed the baselines for Japanese to English, Chinese to English, and German to English translation for both BLEU and RIBES.[20] This demonstrates the effectiveness of the proposed SEQUENCE.[21] The proposed method is thought to be better than the compared methods for local word ordering since BLEU is sensitive to local word order. The proposed method is also thought to be better than the compared methods for global word ordering since RIBES is sensitive to global word order. The BLEU and RIBES scores of the proposed SEQUENCE were higher than those of the proposed PAIR. This confirms its effectiveness in considering relative word order and words between the CP and an SPC. The proposed PAIR outperformed 9-CLASS

[20]In order to verify the performance of our decoder, we also conducted several experiments for baselines of LINEAR and LINEAR+LEX using the Moses phrase-based decoder. The scores for Moses are follows. LINEAR achived a BLEU score of 27.78 and a RIBES score of 67.08 for JE at distortion limit of 10. LINEAR+LEX achieved a BLEU score of 30.62 and a RIBES score of 69.03 for JE at distortion limit of 20. LINEAR achieved a BLEU score of 22.64 and a RIBES score of 74.73 for CE (NIST 2008) at distortion limit of 10. LINEAR+LEX achieved a BLEU score of 22.85 and a RIBES score of 75.58 for CE (NIST 2008) at distortion limit of 10. These scores and the scores for our decoder were similar.

[21]There are differences in the improvements of the scores from the baselines between the NTCIR-9 results and the NIST 2008 results for CE translation. However, note that when the rates of gains from the baselines are compared, the differences were smaller than the differences of the absolute scores. We think that one of the reasons for the differences is that patent translation is more literal than news translation. If translations are literal, then predicting the subsequent position is easier than with non-literal translations, because there are smaller variations in the translations. This results in a consistency in the subsequent positions in the training data and between the training data and the test set.

Table VI. German-English Translation Evaluation Results for WMT 2008 Europarl Data

	BLEU			RIBES		
Distortion limit	6	10	20	6	10	20
LINEAR	26.89	26.59	25.92	78.26	77.83	75.54
LINEAR+LEX	27.09	26.13	26.26	78.38	77.23	75.56
9-CLASS	27.38	27.51	26.97	**78.88**	78.41	76.04
PAIR	**27.87**	27.76	26.89	**78.88**	78.64	75.32
SEQUENCE	**27.88**	**28.04**	27.60	**79.06**	**78.78**	76.74

Table VII. Evaluation Results for Hierarchical Phrase-Based SMT

		BLEU	RIBES
HIER	Japanese-English (NTCIR-9)	30.47	70.43
	Chinese-English (NTCIR-9)	32.66	78.25
	Chinese-English (NIST 2008)	23.62	75.86
	German-English (WMT 2008)	27.93	78.78

for both BLEU and RIBES in most cases[22], confirming that considering both the word specified at the CP and the word specified at the SPC simultaneously was more effective than that of 9-CLASS.

For translating between languages with widely different word orders such as Japanese and English, a small distortion limit is undesirable because there are cases where correct translations cannot be produced with a small distortion limit, since the distortion limit prunes the search space that does not fit within the constraint. Therefore, a large distortion limit is required to translate correctly. For JE translation, our SEQUENCE achieved significantly better results at distortion limits of 20 and 30 than that at a distortion limit of 10 for both BLEU and RIBES, while the baseline systems of LINEAR, LINEAR+LEX, and 9-CLASS did not achieve this. This indicates that SEQUENCE could treat long distance reordering candidates more appropriately than the compared methods.

We also tested hierarchical phrase-based SMT [Chiang 2007] (HIER) using the Moses implementation [Hoang et al. 2009]. The common data was used to train HIER. We used unlimited max-chart-span for the system setting. Results are given in Table VII. Our SEQUENCE outperformed HIER for JE and achieved better than or comparable to HIER for CE and GE. Since phrase-based SMT generally has a faster decoding speed than hierarchical phrase-based SMT, there is merit in achieving better or comparable scores.

To investigate how well SEQUENCE learns the effect of distance, we checked the average distortion probabilities for large distortions of $j - i - 1$. Figure 5 shows three types of probabilities for distortions from 3 to 20 for Japanese-English translation. One type is the average distortion probabilities in the Japanese test sentences for each distortion for SEQUENCE, and another is this for PAIR. The third (CORPUS) is the probabilities for the actual distortions in the training data that were obtained from the word alignments used to build the translation model. The probability for a distortion for CORPUS was calculated by the number of the distortion divided by the total number of distortions in the training data.

Figure 5 shows that when a distance class feature used in the model was the same (e.g., distortions from 5 to 20 had the same distance class feature), PAIR produced

[22]There were two cases in which PAIR was worse than 9-CLASS, in Table V at a distortion limit of 20 and in Table VI at a distortion limit of 20. We think that these were caused by the differences in the SMT weight parameters tuned by MERT.

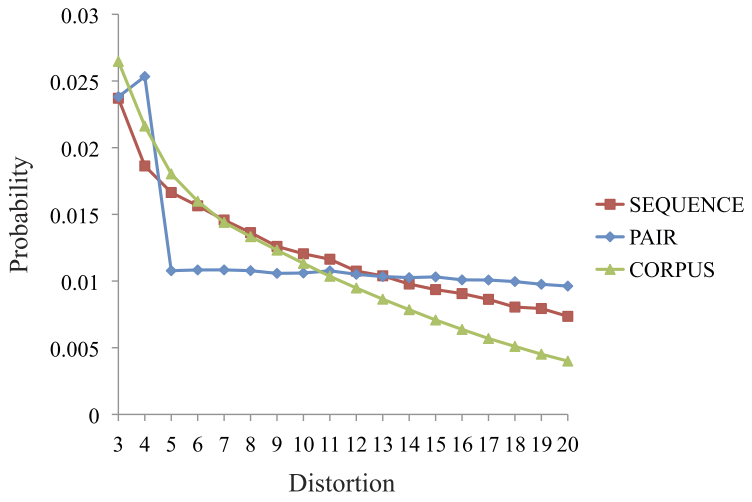

Fig. 5. Average probabilities for large distortions in Japanese-English translation.

Table VIII. Japanese-English Evaluation Results without and with the Words Surrounding the SPCs and the CP (context)

	BLEU	RIBES
PAIR without surrounding words	30.01	69.02
PAIR (with surrounding words)	32.36	72.04
SEQUENCE without surrounding words	31.72	70.71
SEQUENCE (with surrounding words)	33.29	72.86

Note: The best distortion limit of 20 for PAIR and the best distortion limit of 30 for SEQUENCE in Table III were used. The "without" results used the same SMT weighting parameters as those of the "with" results to avoid the effects of differences in SMT weighting parameters.

average distortion probabilities that were almost the same. In contrast, the average distortion probabilities for SEQUENCE decreased when the lengths of the distortions increased even if the distance class feature was the same, and this behavior was the same as that of CORPUS. This confirms that the proposed SEQUENCE could learn the effect of distances appropriately from the training data.[23]

To investigate the effect of using the words surrounding the SPCs and the CP (context), we conducted experiments without using the words surrounding the SPCs and the CP for PAIR and SEQUENCE. The models without using the surrounding words were trained using only the features that did not contain context. Table VIII shows

[23]We also checked the average distortion probabilities for the 9-CLASS outbound model in the Japanese test sentences for Japanese-English translation. We averaged the average probabilities for distortions in a distortion span of [4, 6] and also averaged those in a distortion span of [7, 20], where the distortions in each span are in the same distortion class. The average probability for [4, 6] was 0.058 and that for [7, 20] was 0.165. From CORPUS, the average probabilities in the training data for each distortion in [4, 6] were higher than those for each distortion in [7, 20]. However, the converse was true for the comparison between the two average probabilities for the outbound model. This is because the sum of probabilities for distortions from 7 and above was larger than the sum of probabilities for distortions from 4 to 6 in the training data. This comparison indicates that the 9-CLASS outbound model could not appropriately learn the effects of large distances for JE translation.

Table IX. Japanese-English Evaluation Results without and with Part of Speech (POS) Tags

	BLEU	RIBES
PAIR without POS	31.41	70.62
PAIR (with POS)	32.36	72.04
SEQUENCE without POS	32.79	72.21
SEQUENCE (with POS)	33.29	72.86

Note: The best distortion limit of 20 for PAIR in Table III and the best distortion limit of 30 for SEQUENCE were used. The "without" results used the same SMT weighting parameters as those of the "with" results to avoid the effects of differences in SMT weighting parameters.

the results for Japanese-English translation.[24] Both the BLEU and RIBES scores for SEQUENCE without using the words surrounding the SPCs and the CP (context) were lower than those for SEQUENCE using the words surrounding SPCs and the CP (context). There was a 1.5 point difference in the BLEU scores for SEQUENCE. This result confirms that using the words surrounding the SPCs and the CP (context) was very effective.

To investigate the effect of using part of speech tags, we conducted experiments without using part of speech tags for PAIR and SEQUENCE. The models without using part of speech tags were trained using only the features that did not contain part of speech tags. The results of this experiment for Japanese-English translation are shown in Table IX. Both the BLEU and RIBES scores for SEQUENCE without using part of speech tags were slightly lower than those using part of speech tags. There was a 0.5 point difference in the BLEU scores for SEQUENCE. This result confirms that using part of speech tags was slightly effective for SEQUENCE.

To investigate the training data sparsity tolerance, we reduced the training data for the sequence model to 100,000, 50,000, and 20,000 sentences for Japanese-English translation.[25] Figure 6 show the results for PAIR and SEQUENCE. The best distortion limit of 20 for PAIR and the best distortion limit of 30 for SEQUENCE in Table III were used. To avoid effects from differences in the SMT weighting parameters, the same SMT weighting parameters used in Table III were used for each method. SEQUENCE using only 20,000 training sentences achieved a BLEU score of 32.22 and a RIBES score of 71.33. Although the scores are lower than the scores of SEQUENCE with a distortion limit of 30 in Table III, the scores were still higher than those of LINEAR, LINEAR+LEX, and 9-CLASS for JE in Table III. This indicates that the sequence model also works even when the training data is not large. This is because the sequence model considers not only the word at the CP and the word at an SPC but also rich context, and rich context would be effective even on a smaller set of training data.

[24]Since both the distortion model features with and without the surrounding words represent the same probability shown by Equation (1), the same SMT weighting parameters can be used for these features. This was confirmed using SEQUENCE and PAIR, which are also different distortion model features and represent the same probability shown by Equation (1). The scores for PAIR with a distortion limit of 30 in Table X are higher than those in Table III. SEQUENCE was used to tune the SMT weighting parameters in Table X, whereas PAIR was used to tune the SMT weighting parameters in Table III, which indicates that the same SMT weighting parameters can be used for features representing the same probability. However, the SMT weighting parameters tuned by MERT differed for each tuning, and these differences had an effect on the results. For example, the scores for SEQUENCE with a distortion limit of 20 in Tables III and X differ. This difference was caused by the difference in the SMT weighting parameters. It is therefore important to avoid the effects of differences in SMT weighting parameters for comparison.

[25]We did not conduct experiments using larger training data because there would have been a very high computational cost to build models using the L-BFGS method.

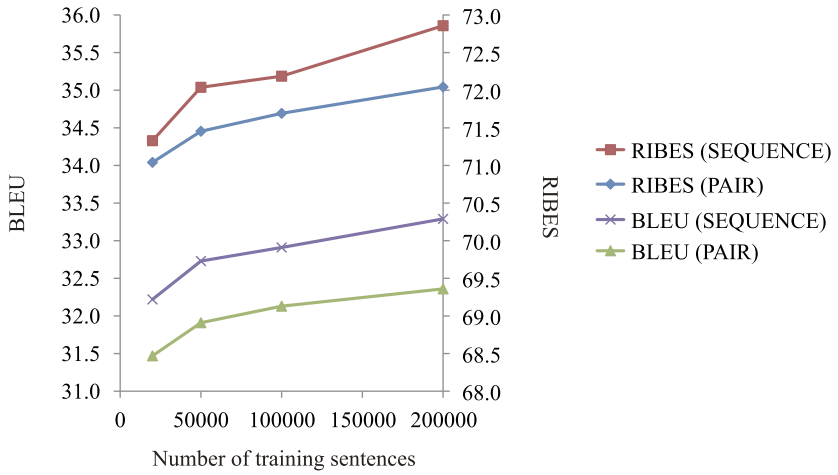

Fig. 6. Relation between the BLEU/RIBES scores and the number of training sentences of the distortion models for Japanese-English translation.

Table X. Japanese-English Translation Evaluation Results Using the Same SMT Weighting Parameters

	BLEU				RIBES			
Distortion limit	10	20	30	∞	10	20	30	∞
PAIR	31.34	32.29	32.17	32.18	71.12	72.00	70.77	69.16
SEQUENCE	32.24	**33.35**	**33.29**	**33.33**	71.86	**73.75**	72.86	71.60

To investigate the effect of distortion limits for PAIR and SEQUENCE for Japanese-English translation more precisely, we conducted experiments using the same SMT weighting parameters to avoid the effects of differences in SMT weighting parameters. For all of the distortion limits of PAIR and SEQUENCE, we used the same SMT weighting parameters that were used for SEQUENCE with a distortion limit of 30 in Table III, which achieved the best scores in Table III. The results of this are given in Table X.

In Table III, the BLEU score for SEQUENCE with an unlimited distortion was lower than that with a distortion limit of 30. However, Table X shows that SEQUENCE with an unlimited distortion achieved almost the same BLEU score as that achieved by SEQUENCE with a distortion limit of 30. This indicates that the difference in BLUE scores for SEQUENCE between a distortion limit of 30 and an unlimited distortion in Table III was mainly caused by the difference in SMT weighting parameters. However, although the RIBES score for SEQUENCE with an unlimited distortion in Table X was higher than that in Table III, the RIBES score for SEQUENCE with an unlimited distortion was still lower than that with a distortion limit of 30 in Table X. The RIBES score for SEQUENCE with a distortion limit of 30 was also lower than that with a distortion limit of 20 in Table X. This indicates that SEQUENCE could not sufficiently handle long distance reordering over 20 or 30 words. For such long distance reordering, incorporation with methods that consider sentence-level consistency, such as ITG constraint [Zens et al. 2004], would be useful.

5. RELATED WORK

In this section, we will discuss related work other than that discussed in Section 2. There is a method that uses SMT sparse features to improve reordering in

phrase-based SMT [Cherry 2013]. However, since the training for this method depends on the SMT weight parameter tuning, the sparse features can only learn from the development data for the SMT weight parameter tuning and cannot utilize a large supply of word aligned training data. Thus, they viewed the sparse features as complementary to existing distortion models. In contrast, our model utilizes a large supply of word aligned training data for training, and it can be built independently of the SMT weight parameter tuning. In addition, SMT sparse features do not calculate the probability of an SPC, whereas our model does. Since Cherry's [2013] sparse features learn from the development data and our model learns from the training data with word alignments, if they are used together, then the SMT system can utilize both the development data and the training data with word alignments to learn reorderings.

There are also reordering models that use a parser: a linguistically annotated ITG [Xiong et al. 2008], a model predicting the orientation of an argument with respect to its verb using a parser [Xiong et al. 2012], and an MSD reordering model using a CCG parser [Mehay and Brew 2012]. However, none of these methods consider reordering distances. Structural information such as syntactic structures and predicate-argument structures are useful for reordering, but orientations do not handle distances. A distortion model considering distances of distortions is also useful for methods predicting orientations using a parser when a phrase-based SMT is used, which means that our distortion model does not compete against methods predicting orientations using a parser, but would assist them if used together.

There are word reordering constraint methods that use ITG for phrase-based SMT [Cherry et al. 2012; Feng et al. 2010; Zens et al. 2004]. These methods consider sentence level consistency with respect to ITG. The ITG constraint does not consider distances of reordering and is used with other distortion models. Our distortion model does not consider sentence level consistency, so our distortion model and ITG constraint methods are thought to be complementary.

There are pre-ordering methods using a supervised parser [Dyer and Resnik 2010; Ge 2010; Genzel 2010; Isozaki et al. 2010b; Wang et al. 2007; Xia and McCord 2004] and methods that do not require a supervised parser [DeNero and Uszkoreit 2011; Neubig et al. 2012; Visweswariah et al. 2011]. These methods are not distortion models, and a distortion model would be useful for their methods when a phrase-based SMT is used for translation.

There are also tree-based SMT methods [Chiang 2007, 2010; Galley et al. 2004; Huang et al. 2006; Liu et al. 2006, 2009; Shen et al. 2008; Yamada and Knight 2001]. In many cases, tree-based SMT methods do not use distortion models that consider reordering distance apart from translation rules, because using distortion scores that consider the distances for decoders which do not generate hypotheses from left to right is not trivial. Our distortion model might contribute to tree-based SMT methods if it could be applied to these methods. Investigating the effects will be for future work.

6. CONCLUSION

This article described our distortion models for phrase-based SMT. Our sequence model consists of only one probabilistic model, but it can consider rich context. In contrast to the learning strategy of existing methods, our learning strategy is that the model learns preference relations among SPCs in each sentence of the training data. This leaning strategy enables consideration of all of the rich context simultaneously. Experiments indicated that our models achieved better performances as measured by both BLEU and RIBES for Japanese-English, Chinese-English, and German-English translation, and that the sequence model could learn the effect of distances appropriately. Since our models do not require a parser, they can be applied to many languages.

Future work includes application to other language pairs, incorporation into ITG constraint methods and other reordering methods, and application to tree-based SMT methods.

REFERENCES

Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 529–536. DOI:http://dx.doi.org/10.3115/1220175.1220242.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist. 22*, 1, 39–71. http://dl.acm.org/citation.cfm?id=234285.234289.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 70–106. http://www.aclweb.org/anthology/W/W08/W08-0309.

Stanley F. Chen and Joshua T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Tech. rep. TR-10-98. Computer Science Group, Harvard University.

Stanley F. Chen and Ronald Rosenfeld. 1999. A Gaussian prior for smoothing maximum entropy models. Tech. rep., School of Computer Science, Carnegie Mellon University.

Colin Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 22–31. http://www.aclweb.org/anthology/N13-1003.

Colin Cherry, Robert C. Moore, and Chris Quirk. 2012. On hierarchical re-ordering and permutation parsing for phrase-based decoding. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 200–209. http://www.aclweb.org/anthology/W12-3125.

David Chiang. 2007. Hierarchical phrase-based translation. *Comput. Linguistics 33*, 2, 201–228.

David Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1443–1452. http://www.aclweb.org/anthology/P10-1146.

John DeNero and Jakob Uszkoreit. 2011. Inducing sentence structure from parallel corpora for reordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 193–203. http://www.aclweb.org/anthology/D11-1018.

Chris Dyer and Philip Resnik. 2010. Context-free reordering, finite-state translation. *Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 858–866. http://www.aclweb.org/anthology/N10-1128.

Theodoros Evgeniou and Massimiliano Pontil. 2002. Learning preference relations from data. In *Proceedings of the 13th Italian Workshop on Neural Nets*. Lecture Notes in Computer Science, vol. 2486, 23–32.

Minwei Feng, Jan-Thorsten Peter, and Hermann Ney. 2013. Advancements in reordering models for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers). Association for Computational Linguistics, 322–332. http://www.aclweb.org/anthology/P13-1032.

Yang Feng, Haitao Mi, Yang Liu, and Qun Liu. 2010. An efficient shift-reduce decoding algorithm for phrased-based machine translation. In *Proceedings of the International Conference on Computational Linguistics*. Coling 2010 Organizing Committee, 285–293. http://www.aclweb.org/anthology/C10-2033.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 848–856. http://www.aclweb.org/anthology/D08-1089.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04)*. D. Marcu, S. Dumais, and S. Roukos Eds., Association for Computational Linguistics, 273–280.

Niyu Ge. 2010. A direct syntax-driven reordering model for phrase-based machine translation. *Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 849–857. http://www.aclweb.org/anthology/N10-1127.

Dmitriy Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling'10)*. Coling 2010 Organizing Committee, 376–384. http://www.aclweb.org/anthology/C10-1043.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of the 9th NTCIR Workshop (NTCIR-9)*. 559–578.

Isao Goto, Masao Utiyama, Eiichiro Sumita, Akihiro Tamura, and Sadao Kurohashi. 2013. Distortion model considering rich context for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers). Association for Computational Linguistics, 155–165. http://www.aclweb.org/anthology/P13-1016.

Spence Green, Michel Galley, and Christopher D. Manning. 2010. Improved models of distortion cost for statistical machine translation. *Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 867–875. http://www.aclweb.org/anthology/N10-1129.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*. 152–159.

Liang Huang, Kevin Knight, and Aravind Joshi. 2006. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*. 66–73.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010a. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 944–952. http://www.aclweb.org/anthology/D10-1092.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010b. Head finalization: A simple reordering rule for SOV languages. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, 244–251. http://www.aclweb.org/anthology/W10-1736.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*. D. Lin and D. Wu Eds., Association for Computational Linguistics, 388–395.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, 177–180. http://www.aclweb.org/anthology/P07-2045.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 282–289. http://dl.acm.org/citation.cfm?id=645530.655813.

D. C. Liu and J. Nocedal. 1989. On the limited memory method for large scale optimization. *Math. Program. B* 45, 3, 503–528.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 609–616. DOI:http://dx.doi.org/10.3115/1220175.1220252.

Yang Liu, Yajuan Lü, and Qun Liu. 2009. Improving tree-to-tree translation with packed forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, 558–566. http://www.aclweb.org/anthology/P/P09/P09-1063.

Dennis Nolan Mehay and Christopher Hardie Brew. 2012. CCG syntactic reordering models for phrase-based machine translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 210–221. http://www.aclweb.org/anthology/W12-3126.

Graham Neubig, Taro Watanabe, and Shinsuke Mori. 2012. Inducing a discriminative parser to optimize machine translation reordering. In *Proceedings of the Joint Conference on Empirical Methods in Natural*

Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 843–853. http://www.aclweb.org/anthology/D12-1077.

Yizhao Ni, Craig Saunders, Sandor Szedmak, and Mahesan Niranjan. 2009. Handling phrase reorderings for machine translation. In *Proceedings of the ACL-IJCNLP Conference Short Papers*. Association for Computational Linguistics, 241–244. http://www.aclweb.org/anthology/P/P09/P09-2061.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 160–167. DOI:http://dx.doi.org/10.3115/1075096.1075117.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 311–318. DOI:http://dx.doi.org/10.3115/1073083.1073135.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 577–585. http://www.aclweb.org/anthology/P/P08/P08-1066.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of the Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics* (Short Papers). D. Marcu, S. Dumais, and S. Roukos Eds., Association for Computational Linguistics, 101–104.

Karthik Visweswariah, Rajakrishnan Rajkumar, Ankur Gandhe, Ananthakrishnan Ramanathan, and Jiri Navratil. 2011. A word reordering model for improved machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 486–496. http://www.aclweb.org/anthology/D11-1045.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, 737–745. http://www.aclweb.org/anthology/D/D07/D07-1077.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*. 508–514.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 521–528. DOI:http://dx.doi.org/10.3115/1220175.1220241.

Deyi Xiong, Min Zhang, Aiti Aw, and Haizhou Li. 2008. Linguistically annotated BTG for statistical machine translation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling'08)*. Coling 2008 Organizing Committee, 1009–1016. http://www.aclweb.org/anthology/C08-1127.

Deyi Xiong, Min Zhang, and Haizhou Li. 2012. Modeling the translation of predicate-argument structure for SMT. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 902–911. http://www.aclweb.org/anthology/P12-1095.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 523–530. DOI:http://dx.doi.org/10.3115/1073012.1073079.

Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 55–63. DOI:http://www.aclweb.org/anthology/W/W06/W06-3108.

Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*. 205–211.

Received May 2013; revised September 2013; accepted October 2013