IEICE TRANS. INF. & SYST., VOL.E95–D, NO.6 JUNE 2012

PAPER

Japanese Argument Reordering Based on Dependency Structure for Statistical Machine Translation

Chooi-Ling GOH[†a)], **Taro WATANABE**[†], ***Nonmembers***, ***and*** **Eiichiro SUMITA**[†], ***Member***

SUMMARY While phrase-based statistical machine translation systems prefer to translate with longer phrases, this may cause errors in a free word order language, such as Japanese, in which the order of the arguments of the predicates is not solely determined by the predicates and the arguments can be placed quite freely in the text. In this paper, we propose to reorder the arguments but not the predicates in Japanese using a dependency structure as a kind of reordering. Instead of a single deterministically given permutation, we generate multiple reordered phrases for each sentence and translate them independently. Then we apply a re-ranking method using a discriminative approach by Ranking Support Vector Machines (SVM) to re-score the multiple reordered phrase translations. In our experiment with the travel domain corpus BTEC, we gain a 1.22% BLEU score improvement when only 1-best is used for re-ranking and 4.12% BLEU score improvement when *n*-best is used for Japanese-English translation.

key words: *predicate-argument structure, reordering, paraphrasing, re-ranking, statistical machine translation*

1. Introduction

Japanese word order is classified as subject-object-verb (SOV). Nonetheless, the word order is strict only in that the verb should be placed at the end of a sentence, and other arguments can be in any order or omitted. The grammatical functions of these arguments can be identified by the particles that are associated with them. Therefore, a sentence with an OSV order is completely correct in Japanese. In certain cases, VS or VO can be acceptable as well, although they are quite rare. These arguments can be complements and adjuncts. The order of the complements and adjuncts is not solely determined by predicates such as verbs and adjectives, although there are some preferences. Normally, an argument that is being emphasized is placed before the others. An argument in Japanese is regarded as a segment in a sentence consisting of at least one content morpheme followed by a sequence of function morphemes.

- バスで 奈良へ 行きます。
  [by bus] [to Nara] *[go]

In the example above, there are two arguments: "バスで" (by bus), "奈良へ" (to Nara); and a predicate: "行きます" (go). The subject "私は" (I) is omitted. Aside from the predicate at the sentence-final, which can be a verb or an adjective, the other arguments can be rearranged quite freely as follows:

- 奈良へ バスで 行きます。
  [to Nara] [by bus] *[go]

Both sentences are grammatically correct although their orders are different. Furthermore, the meanings are similar, both sentences could be translated into English as "I go to Nara by bus." This could be a kind of paraphrasing which is based on the grammatical factor [1] but somehow is not the same. We will call this kind of paraphrase as reordered phrase.

In a phrase-based statistical machine translation (SMT) system, longer phrases are preferred over the shorter ones when translation is being carried out. As a result, the order of the source words will affect the translation output. Traditionally, one would reorder the source side so that it has a similar word order to the target language [2]–[6]. By doing so, the alignment between the source and the target can be done more easily and a better translation model can be built. While the previous research focused on a predetermined order for translation, we try to use multiple orders for translation and compare the results.

In this paper, we propose a method of reordering the Japanese arguments based on dependency structures. We only reorder the arguments, not the predicates, in accordance to Japanese grammar. However, it is not clear whether reordering is required all the time. If the order of a source sentence is appropriate for translation, then reordering is unnecessary. Therefore, we generate multiple reordered phrases and use the re-ranking method to re-score the translation output. A discriminative approach by Ranking Support Vector Machines (SVM) is used for this purpose. The features used include the decoder scores (translation model, language model, distortion model, length penalty), source and target words, word alignments, source language model and source weight (whether or not it is a reordered phrase), and some statistics on the translation phrases used. An experiment on translation of the travel domain using the BTEC corpus from Japanese to English shows an improved BLEU of 1.22%, with 26% of the sentences translated from reordered phrases. If *n*-best translation is used for re-ranking, a 4.12% BLEU score improvement is obtained.

The structure of the remainder of the paper is as follows: Section 2 introduces previous work related to paraphrasing and reordering for machine translation. Section 3 describes the problem tackled and our approach to solving this problem. After that, Sect. 4 shows our experiment re-

Manuscript received September 13, 2011.
Manuscript revised December 21, 2011.

[†]The authors are with the National Institute of Information and Communications Technology, Keihanna Science City, Kyoto-fu, 619–0289 Japan.
a) E-mail: chooiling.goh@gmail.com

DOI: 10.1587/transinf.E95.D.1668

Copyright © 2012 The Institute of Electronics, Information and Communication Engineers

sults and discusses the findings from the results, and finally, Sect. 5 concludes the paper and suggests future work.

2. Previous Work

[7] used paraphrases only when there was an unknown word or phrase. They extracted the unknown word/phrase, found equivalent paraphrases and then added the translation pair into the phrase table. Then, they translated the sentence using the new phrase table. The paraphrases were generated using bilingual parallel corpora where phrases were extracted when a single source phrase was translated into multiple target phrases [8]. While this method may suffer from errors when substituting the paraphrases into the sentence, syntactic constraints were applied later to improve the results [9].

[10] and [11] generated paraphrases using English grammar rules. Both of them used the paraphrases to expand the training corpus in an SMT framework. While [11] focused on paraphrasing noun phrases, [10] proposed to generate paraphrases using English Resource Grammar (ERG). Although their methods could generate more grammatically correct paraphrases, only small gains were reported in terms of translation.

All the previous work listed above augmented the translation model by paraphrasing sentences, but are prone to errors incurred by their paraphrasing methods. However, one can also use paraphrases as the input source without changing the translation model. [12] generated a paraphrase lattice for each sentence by building a lattice based on the paraphrase model [8]. Translation is carried out by lattice decoding over the paraphrase lattice with additional features: paraphrase probability, language model score and paraphrase length. This method does not modify the translation model, but does provide flexibility in matching phrase pairs in the translation model.

Our method differs from previous work on three points: 1) we reorder the source input and translate each reordered phrase independently; 2) we assume dependency-wise permutations will not alter the original meaning in free word order languages such as Japanese; 3) we re-rank the translations from multiple reordered phrases using a discriminatively-learned model by incorporating the reordered phrase information. In this case, we can keep the translation table small without adding the phrases extracted from reordered bitext, but at the same time can improve the translation. Furthermore, we can provide more information in the re-ranking than the lattice decoding. In previous work, more resources, such as parallel corpora other than the one used for building translation model, were needed to build the paraphrase databases, but only a parser is needed for word reordering. Our approach also differs from [2]–[4], where the source is reordered so that it is in the same word order as the target, in that our reordered phrase word order preserves grammatical correctness in the source language. Our reordering generates multiple sources with different word orders and uses a re-ranking approach to select the translation that is better. The intuitiveness behind the two approaches is similar: a word order that is similar to the translation model can generate a better translation.

3. Translation by Reordered Phrases

The problems with free word order languages, such as Japanese, Korean and Finnish, is that the word order is too flexible. As a consequence, it is impossible to collect a bitext that can cover all possible grammatical permutations. Therefore, we propose using dependency-wise reordering of arguments to generate multiple permutations, and using a re-ranking approach to re-score the multiple translations. Our proposed method consists of three steps:

1. Generate multiple reordered phrases for each input sentence based on its dependency structure
2. Translate all reordered phrases independently using a standard phrase-based SMT system
3. Re-rank the reordered phrase translations using Ranking SVM

3.1 Problems with Phrase-Based SMT

A standard phrase-based SMT system was formalized by [13], [14], where alignments were first done using words, and then a phrase translation table was created on the basis of word alignments. During the translation process, the input is segmented into a number of phrases[†]. Then, each phrase is translated into the target phrase and the output phrases may be reordered to arrive at a grammatically correct sentence. The phrase translation model is based on the noisy channel model. The translation probability for translating a source sentence **f** into target language **e** is given as

$$\underset{\mathbf{e}}{\operatorname{argmax}}\, \mathbf{p}(\mathbf{e}|\mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}}\, \mathbf{p}(\mathbf{f}|\mathbf{e})\mathbf{p}(\mathbf{e}) \quad (1)$$

where $\mathbf{p}(\mathbf{e})$ is a language model and $\mathbf{p}(\mathbf{f}|\mathbf{e})$ is a translation model.

During the decoding process, the input sentence is segmented into a sequence of I phrases f_1^I and each f_i in f_1^I is translated into a target language phrase e_i. The target phrases may be reordered according to the language model. Usually, longer f_i phrases are preferred so that a fewer number of I phrases are segmented.

The nature of the phrase-based SMT model is that longer phrases are preferred over shorter ones during translation. By doing this, it is assumed that the generated target sentence will be more grammatically correct. For example, in Fig. 1, although both input sentences semantically are the same, the translations are different when different phrases are used. In this example, both input sentences are fortunately translated correctly.

However, in some cases, longer phrases do not definitely mean good for translation. In Fig. 2, when the longer

[†]Note that these phrases may not be linguistically correct.

Reference	there is an elevator on the right-hand side
Original	右手にエレベーターがあります
Translation	there is an elevator on your right
Phrases used	[there ||| あり ます] [is an elevator on your right ||| 右手 に エレベーター が]
Decoder's score	-1.10
Reordered	エレベーターが右手にあります
Translation	the elevator is on your right hand side
Phrases used	[the elevator is ||| エレベーター が] [on your right hand side ||| 右手 に あり ます]
Decoder's score	-1.00

Fig. 1 Phrase-based SMT output 1.

Reference	buses go directly to nara
Original	バスが直接奈良へ行きます
Translation	a bus i am going to nara
Phrases used	[a ||| が 直接] [bus ||| バス] [i am going to nara ||| 奈良 へ 行き ます]
Decoder's score	-1.83
Reordered	バスが奈良へ直接行きます
Translation	the bus will go directly to nara
Phrases used	[the bus ||| バス が] [will go ||| 行き ます] [directly ||| 直接] [to nara ||| 奈良 へ]
Decoder's score	-2.15

Fig. 2 Phrase-based SMT output 2.

phrase "奈良 へ 行き ます" is used for the process, the translation becomes worse. In this case, using shorter phrases could generate a better translation. The decoder's scores also could not give us a good idea on which input sentence should be used.

These errors are potentially caused by the free word order, where the coverage of all permutations is insufficient for the translation model since we cannot collect the bitext with all possible reorderings. Therefore, in the following section, we propose a method of reordering the words in the source sentence based on the dependency structure, and re-rank the reordered phrase translations using a discriminative approach by Ranking SVM, where a large number of features can be used.

3.2 Dependency Reordering

We employ a dependency parser to analyze the Japanese texts. A dependency structure for a sentence is usually segmented by arguments and predicates[†]. Each argument contains at least one content morpheme followed by a sequence of function morphemes. The order of the arguments is not determined by the predicates alone, which means that an argument can be freely placed in any location before its dependency head (a predicate) in the sentences[††]. Figure 3 shows a dependency structure. Below is the sentence in its common order in Japanese[†††].

- 私たちは 午後も このボートを 使うのですか。
  [we] [afternoon too] [this] [boat] *[going to use ?]
  are we going to use this boat in the afternoon too ?

[4] proposed rearranging the sequence of arguments so that it is similar to the target language. Besides a syntactic parser, they also employed a semantic role labeling system in order to find out the grammatical functions of the case markers. Some hand-crafted rules were used to reorder Japanese to match English word order. For instance: [nominative][predicate][accusative][locative]. In our example, the sentence will be reordered as:

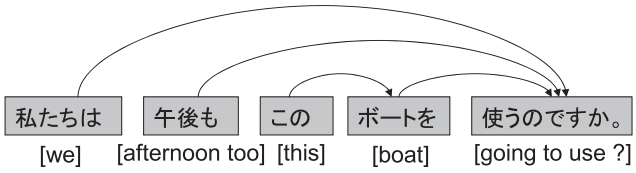

Fig. 3 A dependency structure.

- 私たちは 使うのですか。 このボートを 午後も
  [we] *[going to use ?] [this] [boat] [afternoon too]

Then, they used only these reordered source sentences to train a reordered translation model. However, the translations from the reordered input with the reordered translation model itself did not achieve better results than the original order. Therefore, these reordered sentence pairs were added to the original training set to re-train a larger translation model. This means that the reordering only helps to improve the translation model by reducing crossed word alignments during training. Even with training on both original and reordered sentence pairs, the translation is better when using the original order input sentences than the reordered input sentences. This shows that their reordering method does not seem to work well as input. Since their method provides only a fixed order, it does not give the flexibility to translate the same sentence in different word orders. We assume that a word order that could be found in the translation model built will have better translation. In other words, we want to do reordering so that the possibility of retrieving the proper phrase pairs is maximized using an existing translation model. Therefore, as opposed to a fixed order based on the grammatical functions of the case markers, we want to arrange the sentence in as many ways as possible for translation. Using the dependency structure shown in Fig. 3, we can reorder the arguments in any possible order. We only reorder the arguments, but not the predicates, because this pattern will not occur in our translation model.

Usually the sentence-final is the head of the sentence, which is the predicate, and the arguments before it are dependents of it. In this case, the former arguments can be reordered in such a way that the head always remains in the last position.

- 私たちは 午後も このボートを 使うのですか。
  [we] [afternoon too] [this] [boat] *[going to use ?]

[†]These arguments and predicates are referred to as *bunsetsu* in Japanese.

[††]However, in reality, there are some preferred orders.

[†††]The predicate is indicated by an asterisk throughout the paper.

- 私たちは このボートを 午後も 使うのですか。
  [we] [this] [boat] [afternoon too] *[going to use ?]
- 午後も 私たちは このボートを 使うのですか。
  [afternoon too] [we] [this] [boat] *[going to use ?]
- 午後も このボートを 私たちは 使うのですか。
  [afternoon too] [this] [boat] [we] *[going to use ?]
- このボートを 私たちは 午後も 使うのですか。
  [this] [boat] [we] [afternoon too] *[going to use ?]
- このボートを 午後も 私たちは 使うのですか。
  [this] [boat] [afternoon too] [we] *[going to use ?]

Even though the first pattern is more preferred as the subject comes before the other complements and adjuncts, and the object is just right before the predicate, all sentences are grammatically correct. Unlike the sentences generated by [4], most of the Japanese sentences generated with our approach will be grammatically correct after reordering. For n number of dependents, the possible number of reordering will be $n!$ ways. A language model can be used to show which pattern has a higher probability of being more commonly used, but it cannot show whether this pattern will generate a better translation than the others.

3.3 Ranking SVM

Our ranking algorithm is based on the ranking approach of [15] in which we seek the maximum scored output $\hat{\mathbf{e}}$ from a large n-best list

$$\hat{\mathbf{e}} = \underset{\mathbf{e} \in \mathrm{GEN}(\mathbf{f})}{\operatorname{argmax}} \mathbf{w}^\top \cdot \mathbf{h}(\mathbf{e}, \mathbf{f}) \tag{2}$$

where GEN$(\cdot)$ is normally an n-best list, a set of candidate translations, generated from the input sentence $\mathbf{f}$. However, in our approach, by reordering the input sentence $\mathbf{f}$, the n-best list is actually composed of the 1-best/n-best reordered phrase translations for each input sentence $\mathbf{f}$. $\mathbf{h}(\cdot)$ defines mapping from input/output sentence pair to feature functions, and $\mathbf{w}$ is a weight vector. In training the parameter vector $\mathbf{w}$, we employed an online large-margin learning for structured output classification [16]–[18] based on the margin infused relaxed algorithm (MIRA) [19]. First, we generate a large n-best list $\mathbf{E}$ for m input sentences $\mathbf{f}_{1\ldots m}$. For each iteration, we randomly choose an input sentence $\mathbf{f}_i$ and its corresponding n_i-best list $\mathbf{E}_i$. We seek a maximum scored hypothesized translation $\mathbf{E}_{ij}$ using the current weight $\mathbf{w}$

$$\mathbf{w}^\top \cdot \mathbf{h}(\mathbf{E}_{ij}) - \mathrm{b}(\mathbf{E}_{ij}) \tag{3}$$

where $\mathbf{h}(\mathbf{E}_{ij})$ and $\mathrm{b}(\mathbf{E}_{ij})$ are a feature vector representation and the BLEU score for $\mathbf{E}_{ij}$, respectively. Then, we update $\mathbf{w}$ by the value of $\mathbf{w}'$ which minimizes

$$\frac{\lambda}{2}\|\mathbf{w}' - \mathbf{w}\|^2 + \ell_{ij} - \mathbf{w}'^\top \cdot \Delta\mathbf{h}(\mathbf{E}_{ij}) \tag{4}$$

where ℓ_{ij} is the loss incurred by selecting the $\mathbf{E}_{ij}$ as the best translation calculated by the difference of BLEU from an oracle translation $\mathbf{E}_{i*}$

$$\ell_{ij} = \mathrm{b}(\mathbf{E}_{i*}) - \mathrm{b}(\mathbf{E}_{ij}) \tag{5}$$

and $\Delta\mathbf{h}(\mathbf{E}_{ij}) = \mathbf{h}(\mathbf{E}_{i*}) - \mathbf{h}(\mathbf{E}_{ij})$. $\lambda(> 0)$ is a constant to influence the fitness to the training data. Expression 4 first is solved by introducing its Lagrange dual

$$\frac{\lambda}{2}\|\mathbf{w}' - \mathbf{w}\|^2 + \alpha\left(\ell_{ij} - \mathbf{w}'^\top \cdot \Delta\mathbf{h}(\mathbf{E}_{i,j})\right) \tag{6}$$

where $\alpha \geq 0$ is a Lagrange multiplier. The problem in Expression 6 then is solved by taking its partial derivation with respect to $\mathbf{w}'$ to zero, leading to:

$$\mathbf{w}' = \mathbf{w} + \min\left(\frac{\ell_{ij} - \mathbf{w}^\top \cdot \Delta\mathbf{h}_{ij}}{\|\Delta\mathbf{h}_{ij}\|^2}, \frac{1}{\lambda}\right) \cdot \Delta\mathbf{h}_{ij} \tag{7}$$

Unlike the ranking SVM approach for training [20], our learning algorithm considers only a single pair of correct and incorrect translations in each iteration using the loss biased maximization in Expression 3 largely inspired by [18]. For the loss function ℓ_{ij} and the underlying BLEU score $\mathrm{b}(\cdot)$, we applied document-scaled BLEU, which computes BLEU by replacing one translation $\mathbf{E}_{i1}$ with another $\mathbf{E}_{ij}$ in a set of 1-best translations $\{\mathbf{E}_{i1}\}_{i=1\ldots m}$ [17]. Oracle translations are selected with respect to $\mathrm{b}(\cdot)$. When multiple oracle translations are found, we select the one which maximizes $\mathbf{w} \cdot \Delta\mathbf{h}(\mathbf{E}_{ij})$ [18].

3.4 Features

It is possible to use a large number of features in the Ranking SVM. These include real-valued features (the SMT decoder scores, reordered phrase scores, etc.) and some sparse binary features. The features are explained below in more detail.

3.4.1 Decoder Scores

The SMT decoder scores include unweighted scores for target language model, translation model, distortion model and word penalty. We use unweighted scores because the weights for each score will be re-assigned during the reranking learning.

3.4.2 Source and Target Words

The words in the source sentences and target hypotheses sentences are used. Word unigrams, bigrams and trigrams from the source words and target words, and all possible source-target word pairs are extracted.

3.4.3 Alignment

Source-target word pairs are extracted based on the word alignment between the source and the hypothesis by running Model 1 and HMM model in both directions. Then, the word alignment is extracted by combining the results from both directions using the grow-diag-final-and heuristic method. The word pairs are also extended to the previous word and the next word of the source side.

3.4.4 Reordered Phrase Scores

The unweighted score for the source language model of each input sentence, either the original sentence or the reordered phrase, is also used as a feature. The language model score can determine whether the input sentence is more preferable for the translation system. We also introduce an originality score as a feature. For the original sentence, the originality score is 1.0, and 0.5 for a reordered phrase.

3.4.5 Translation Phrases Used

There are two features in this group. The first feature is the number of phrase pairs taken to generate a translation. In general, the least number of phrases used the better, but this is not true in all cases. The second feature is the distance between the source phrase order and the used target phrase order. For example, if the source order is $s1\ s2\ s3$ and the translation has the order $t2\ t1\ t3$, then the score will be $|2-1| + |1-2| + |3-3| = 2$. This is to show the number of reorderings being done during translation.

4. Experiment Results

For the experiment, the Basic Travel Expression Corpus (BTEC) parallel corpus for Japanese-English is used [21]. There are approximately 160 K sentences in this corpus. The decoder used is an in-house phrase-based SMT model [13], Octavian. The parameter weights were tuned with MERT, using 508 sentences with 16 references. The language model was built using SRILM, 5-gram and interpolated with Kneser-Ney discounting. In this corpus, another set of 10K sentence test data with single references was prepared for testing. However, in this test data, only 2,469 sentences could have reordered phrases. The average length of the test set is about 10.5 words. We used 1,000 sentences for the SVM re-ranking training, 500 sentences for the SVM lambda λ value tuning, and 969 sentences for testing.

The Japanese text was segmented using a morphological analyzer, ChaSen[†] [22] and parsed by a dependency parser, CaboCha[††] [23]. If a sentence is composed of two or more sub-sentences (two or more predicates), the argument reordering can only be done within the sub-sentence. Besides, we do not reorder the arguments inside another argument for complex sentences. In other words, reordering is only done for the highest layer of arguments. The maximum number of arguments that can be reordered in our testing set is six ($6! = 720$ reordered phrases) and the minimum is two ($2! = 2$ reordered phrases).

Table 1 shows the experiment results for our proposed method. We evaluated the results using BLEU scores [24]. The baseline model used the original sentences as the source input. The table also shows the results of using the sentences (or reordered phrases) that have higher source language model probability (source lm only). This source language model is built from the same 160K training corpus. However, the results are worse than the baseline. We also selected the higher translation score for the target sentence (target tm only) but since the source is not the same, the comparison is not quite adequate. Finally, we combined both the source language model with the translation score (lm + tm) by summing up both scores and getting the higher one, but none of these naive approaches give better results. We also tested on a translation model built from a reordered training data (reordered trans-model). The dependency reordering method described in Sect. 3.2 was used to reorder the source sentences in the training data. This will cause multiple source sentences to align to a single target sentence. This approach is similar to [10] and [11] where the size of the training corpora is increased. We then used this reordered bitext to train an SMT system as usual. This approach shows some improvements on the test set, but the training data triples in size, and the number of phrase pairs generated is two times larger than the original translation model, slowing the translation process.

Finally, by using all the scores and other features in a re-ranking model, the selection is better and the translation results improved. We show the results of using only the basic features (basic: decoder scores, source and target words), plus word alignments (align: Sect. 3.4.3), the source reordered phrase scores (src: Sect. 3.4.4), and the statistics for the translation phrases used (phrase: Sect. 3.4.5). We chose the λ value that gives the highest BLEU score for the development set. The best λ value can be different for a dif-

Table 1 Japanese-English translation results using BLEU score.

	ja-en		
Model	λ dev	test	
baseline	39.47	40.18	+0.00
source lm only	39.16	39.70	-0.48
target tm only	39.11	40.63	+0.45
lm + tm	39.31	40.02	-0.16
reordered trans-model	39.23	41.10	+0.92
MERT with 1500	43.05	43.93	+3.75
reordered phrase 1-best re-ranking results			
basic	40.49	41.30	+1.12
basic+align	40.36	41.18	+1.00
basic+src	40.38	41.35	+1.17
basic+phrase	40.26	41.39	+1.21
basic+align+src+phrase	40.40	41.40	+1.22
n-best re-ranking results			
basic	41.81	43.59	+3.41
basic+align	**42.26**	43.36	+3.18
basic+src	41.99	43.64	+3.46
basic+phrase	42.02	43.75	+3.57
basic+align+src+phrase	42.07	43.66	+3.48
reordered phrase n-best re-ranking results			
basic	41.91	**44.30**	+4.12
basic+align	41.88	44.05	+3.87
basic+src	42.00	43.87	+3.69
basic+phrase	41.94	44.00	+3.82
basic+align+src+phrase	42.14	44.04	+3.96

[†]http://chasen-legacy.sourceforge.jp/

[††]http://chasen.org/~taku/software/cabocha/

ferent set of used features. We realized that adding features from the reordered phrase information and the translation phrases used could improve the selection results. There is a +1.22% BLEU score improvement for the test set over the baseline.

Using all the features described above for 1-best reordered phrase re-ranking generates the best translation output. Out of 969 test sentences, 251 (25.9%) sentences were translated using reordered phrases and 207 (82.5%) of the reordered phrases generated different translation output; only 44 (17.5%) reordered phrases generated the same translation output as the original source. This also means that most of the original sentences are still better than the reordered phrases.

We also used *n*-best output for re-ranking. In this experiment, 100-best translation output was generated for re-ranking. If we used only the original source with 100-best output for re-ranking, we obtained about +3.57% BLEU score improvement over the baseline. We also tuned the parameter weights with MERT using the training and λ devset for ranking SVM, which has 1,500 sentences with single reference (indicated by MERT with 1500 in Table 1). The result is almost comparable with the *n*-best re-ranking result, with +3.75% BLEU score improvement over the baseline. Since this data set is of a similar type as the test data, where more than one argument exists in the sentence for reordering, so it is not a surprise that the result is comparable to *n*-best re-ranking .

When we included the reordered phrases for 100-best re-ranking (total *n*-best = # of reordered phrases × 100-best), we obtained a +4.12% BLEU score improvement over the baseline, and 0.55% over the original source with 100-best re-ranking. By generating the *n*-best with the original source, some of the better translations could already be found in the *n*-best list, so the improvement for the reordered phrase *n*-best re-ranking is smaller than the reordered phrase 1-best re-ranking. However, when reordered phrase *n*-best re-ranking was used, extra features such as alignment, source reordered phrase score and the statistics for the translation phrases used are not useful for the re-ranking. Using only the basic features generated the best results.

[12] generated paraphrase lattice for decoding which could give better translation. Although we could also build a lattice based on the reordered phrases, no new information can be added besides the order of the phrases. Therefore, we do not think that lattice decoding is suitable for our approach. Since Octavian does not provide a lattice decoding function, we rebuilt a translation model using Moses [25], [26] with the same configuration. We did not incorporate any arc probability for the lattice besides the weight of the branches. The weight is divided equally between the branches. The baseline gives a BLEU percentage score of 44.13 and the lattice decoding is 42.59. Despite the lattice decoding, there is no improvement, and it is also slow compared with generating the *n*-best output. This is because the lattice can be huge when the number of arguments increases and the sentence is long.

Table 2 Japanese-Chinese and Japanese-Korean translation results using BLEU score.

	ja-zh		ja-ko	
Model	λ dev	test	λ dev	test
baseline	46.30	50.14	64.80	66.08
reordered trans-model	45.86	49.26	65.58	66.60
1-best reordered phrase re-ranking	46.80	50.57	64.66	65.68
n-best re-ranking	47.83	51.60	68.30	68.31
n-best reordered phrase re-ranking	47.76	51.48	67.63	67.32

4.1 Comparing with Other Target Languages

A similar experiment using the parallel Chinese and Korean portions of BTEC were also carried out. While Japanese (SOV) and Chinese (SVO) generally have different word order, the noun phrases are sometimes similar due to the Chinese characters used in both languages. Furthermore, there are quite a lot of expressions in Chinese that can be written as SOV order as well. On the contrary, Japanese and Korean (SOV) have the same word order and sometimes the translation reordering can be monotone. While Korean is almost the same as Japanese as a free word order language, Chinese is less free and English is the least free order language. In other words, we want to find out whether the argument reordering works for a language pair where the target also has a certain level of free word order or similarity to the source language.

Table 2 shows the experiment results. As we can see, using the reordered phrase re-ranking shows some slight improvements in BLEU for Japanese-Chinese translation but does not provide any help in Japanese-Korean translation. In fact, it deteriorates the translation results evaluated using BLEU as the word order of the translation is different from the source and also the reference. Comparing the translation results from Japanese-English, Japanese-Chinese and Japanese-Korean, we can conclude that using the argument reordering approach is not useful for translation pairs where the target language is also a free word order language and has a word order similar to the source language.

4.2 Human Evaluation

Besides the automatic evaluation, human assessment on the translation output was also carried out. We categorized the translations into five ranks as shown in Table 3. We compared only the translations from the baseline and the 1-best reordered phrase re-ranking with all features. Taking from the 1-best reordered phrase re-ranking could show more clearly the effectiveness of the reordering. Only translations that are different were selected for human evaluation. There are 207 sentences for Japanese-English, 136 sentences for Japanese-Chinese and only 36 sentences for Japanese-Korean.

Table 4 shows the human assessment results. We can see that the total number of good translations (SAB ranks) improved for the Japanese-English and Japanese-Chinese

language pairs, but worsened slightly for the Japanese-Korean language pair. This conforms to the automatic evaluations using the BLEU scores. We also had more perfect translation (S rank) using the reordered phrases. For Japanese-English, the improvement is from 16 to 31 sentences and for Chinese-English, the improvement is from 6 to 13 sentences. For all three language pairs, the reordered phrases generated more "better" (higher rank) translations than "worse" (lower rank) translations (right part of Table 4). More improvements can be seen in the Japanese-English translations. Although the automatic evaluation using BLEU score showed deterioration for Japanese-Korean translations, human evaluation shows that the difference is very small. This is because humans can accept a translation that has a different word order than the reference, but BLEU is strict about the ordering.

Figure 4 gives examples of where the reordered phrase translation is better than the original sentence. From the source language model probability, we know that a higher score does not equate to a better translation, and the same goes for the translation score. The number of phrase pairs used for translation also may not always be better with a smaller number. However, using all these features in a re-ranking model could select the better translation.

Table 3 Description of SABCD metric.

Rank	Description
S	Perfect, native-level translation
A	Grammatically correct, but slightly non-native translation
B	Grammatically incorrect but easily understandable translation
C	Grammatically incorrect and difficult to understand, may be missing a slight amount of information
D	Not understandable or missing important information

Table 4 Human assessment results.

Rank	S	+A	+B	+C	+D	better	equal	worse
Japanese-English								
Baseline	16	34	57	70	207	43	144	20
Proposed	31	46	65	83	207			
Japanese-Chinese								
Baseline	6	20	52	89	136	33	72	31
Proposed	13	26	53	88	136			
Japanese-Korean								
Baseline	15	23	25	28	36	10	18	8
Proposed	16	21	23	26	36			

5. Conclusions

For a free word order language, a sentence sometimes cannot be translated correctly when the word order is not found in the translation model. We proposed generating multiple reordered phrases by reordering the arguments without changing the content words, based on the dependency structure. These reordered phrases produced different translation outputs which can sometimes be better than the original ones by increasing the recall of matching phrase pairs in the translation model. We then used the re-ranking approach by applying Ranking SVM to re-score the translation output, providing a large number of sparse features. The experiment results showed that our approach could generate better translations with a BLEU score improvement of around 1.2% for 1-best re-ranking and 4.1% for n-best re-ranking. Currently, the source language model itself could not tell whether a reordered phrase will have better translation than the others. In the future, we would like to score the reordered phrases based on the translation model in order to find out whether it is a good reordered phrase for translation, so that we do not need to translate multiple reordered phrases and re-rank the

Reference	lm / tm	what track does it leave from ?
Original	-6.79	それは 何番線から 出ますか。
Translation	-4.03	what track ?
Phrases used		[what ||| それ は 何] [track ? ||| 番線 から 出 ます か 。]
Reordered	-4.62	何番線から それは 出ますか。
Translation	-12.99	what track does it leave ?
Phrases used		[what track does ||| 何 番線 から] [it ||| それ は] [leave ? ||| 出 ます か 。]
Reference		i 'd like to get my trousers pressed by ten tomorrow morning .
Original	-18.92	このズボンを 明朝十時までに プレスしてください。
Translation	-8.39	these pants pressed by ten o 'clock tomorrow .
Phrases used		[these pants ||| この ズボン] [pressed ||| プレス し て] [by ten o 'clock ||| 十 時 まで に] [tomorrow ||| を 明朝] [. ||| ください 。]
Reordered	-20.58	明朝十時までに このズボンを プレスしてください。
Translation	-6.88	please press these pants by ten o 'clock tomorrow morning .
Phrases used		[please press these pants ||| この ズボン を プレス し て ください] [by ten o 'clock ||| 十 時 まで に] [tomorrow morning ||| 明朝] [. ||| 。]
Reference		where and around what time will we be back ?
Original	-10.81	どこに 何時ごろ 戻ってきますか。
Translation	-5.18	what time will you come back ?
Phrases used		[what ||| どこ に] [time ||| 何時 ごろ] [will you come back ? ||| 戻っ て き ます か 。]
Reordered	-11.30	何時ごろ どこに 戻ってきますか。
Translation	-5.00	where and around what time will we be back ?
Phrases used		[where and around what time will we be back ? ||| 何時 ごろ どこ に 戻っ て き ます か 。]

Fig. 4 Better translations with reordered phrases. The table shows the source language model score (lm), the translation score (tm) and the phrase pairs used for translation.

output. As our method is quite general, it can be applied to other free word order languages such as Korean, Mongolian and Finnish, provided a dependency parser is available.

References

[1] A. Fujita and S. Sato, "Measuring the appropriateness of automatically generated phrasal paraphrases," J. Natural Language Processing, vol.17, no.1, pp.183–219, 2010.
[2] S. Nießen and H. Ney, "Morpho-syntactic analysis for reordering in statistical machine translation," Proc. MT Summit VIII, pp.247–252, 2001.
[3] M. Collins, P. Koehn, and I. Kucerova, "Clause restructuring for statistical machine translation," Proc. ACL, pp.531–540, 2005.
[4] M. Komachi, Y. Matsumoto, and M. Nagata, "Phrase reordering for statistical machine translation based on predicate-argument structure," Proc. IWSLT, pp.77–82, 2006.
[5] D. Genzel, "Automatically learning source-side reordering rules for large scale machine translation," Proc. COLING, pp.376–384, 2010.
[6] K. Visweswariah, R. Rajkumar, A. Gandhe, A. Ramanathan, and J. Navratil, "A word reordering model for improved machine translation," Proc. EMNLP, pp.486–496, 2011.
[7] C. Callison-Burch, P. Koehn, and M. Osborne, "Improved statistical machine translation using paraphrases," Proc. HLT-NAACL, pp.17–24, 2006.
[8] C. Bannard and C. Callison-Burch, "Paraphrasing with bilingual parallel corpora," Proc. ACL, pp.597–604, 2005.
[9] C. Callison-Burch, "Syntactic constraints on paraphrases extracted from parallel corpora," Proc. EMNLP, pp.196–205, 2008.
[10] F. Bond, E. Nicols, D.S. Appling, and M. Paul, "Improving statistical machine translation by paraphrasing the training data," Proc. IWSLT, pp.150–157, 2008.
[11] P. Nakov, "Improved statistical machine translation using monolingual paraphrases," Proc. ECAI, pp.338–342, 2008.
[12] T. Onishi, M. Utiyama, and E. Sumita, "Paraphrase lattice for statistical machine translation," Proc. ACL, pp.1–5, 2010.
[13] P. Koehn, F.J. Och, and D. Marcu, "Statistical phrase-based translation," Proc. HLT/NAACL, pp.81–88, 2003.
[14] F.J. Och and H. Ney, "The alignment template approach to statistical machine translation," Computational Linguistics, vol.30, no.4, pp.417–449, 2004.
[15] M. Collins and N. Duffy, "New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron," Proc. ACL, pp.263–270, 2002.
[16] R. McDonald, K. Crammer, and F. Pereira, "Online large-margin training of dependency parsers," Proc. ACL, pp.91–98, 2005.
[17] T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki, "Online large-margin training for statistical machine translation," Proc. EMNLP-CoNLL, pp.764–773, 2007.
[18] D. Chiang, Y. Marton, and P. Resnik, "Online large-margin training of syntactic and structural translation features," Proc. EMNLP, pp.224–233, 2008.
[19] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," J. Machine Learning Research, vol.7, pp.551–585, March 2006.
[20] T. Joachims, "Optimizing search engines using clickthrough data," KDD '02: Proc. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.133–142, 2002.
[21] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," IEEE Trans. Audio Speech Language, vol.14, no.5, pp.1674–1682, 2006.
[22] Y. Matsumoto, K. Takaoka, and M. Asahara, "ChaSen morphological analyzer version 2.4.0 User's manual," Nara Institute of Science and Technology, March 2007.
[23] T. Kudo and Y. Matsumoto, "Japanese dependency analyisis using cascaded chunking," Proc. CoNLL, pp.63–69, 2002.
[24] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, "BLEU: A method for automatic evaluation of machine translation," Proc. ACL, pp.311–318, 2002.
[25] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," Proc. ACL, pp.177–180, 2007.
[26] C. Dyer, S. Muresan, and P. Resnik, "Generalizing Word Lattice Translation," Proc. ACL, pp.1012–1020, 2008.

Chooi-Ling Goh received the M.E. and Ph.D. in Information Science from Nara Institute of Science and Technology, Japan, in 2003 and 2006, respectively. She is currently an expert researcher at National Institute of Information and Communications Technology, Japan. She is a member of ANLP. Her main research interests include machine translation and morphological analysis.

Taro Watanabe received the B.E. and M.E. degrees in information science from Kyoto Univ., Kyoto, Japan in 1994 and 1997, respectively, and obtained the Master of Science degree in language and information technologies from the School of Computer Science, Carnegie Mellon University in 2000. In 2004, he received the Ph.D. in informatics from Kyoto Univ., Kyoto, Japan. Dr. Watanabe is a researcher of National Institute of Information and Communications Technology. His research interests include natural language processing, machine learning and statistical 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 the group leader of NICT/MASTAR Project/Multilingual Translation Laboratory, and the visiting professor of Kobe University. His research interests include machine translation and e-Learning.