A Clustered Global Phrase Reordering Model for Statistical Machine Translation

Masaaki Nagata
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Souraku-gun
Kyoto, 619-0237 Japan
nagata.masaaki@labs.ntt.co.jp

Kuniko Saito
NTT Cyber Space Laboratories
1-1 Hikarinooka, Yokoshuka-shi
Kanagawa, 239-0847 Japan
saito.kuniko@labs.ntt.co.jp

Kazuhide Yamamoto, Kazuteru Ohashi*
Nagaoka University of Technology
1603-1, Kamitomioka, Nagaoka City
Niigata, 940-2188 Japan
ykaz@nlp.nagaokaut.ac.jp, ohashi@nlp.nagaokaut.ac.jp

Abstract

In this paper, we present a novel global reordering model that can be incorporated into standard phrase-based statistical machine translation. Unlike previous local reordering models that emphasize the reordering of adjacent phrase pairs (Tillmann and Zhang, 2005), our model explicitly models the reordering of long distances by directly estimating the parameters from the phrase alignments of bilingual training sentences. In principle, the global phrase reordering model is conditioned on the source and target phrases that are currently being translated, and the previously translated source and target phrases. To cope with sparseness, we use N-best phrase alignments and bilingual phrase clustering, and investigate a variety of combinations of conditioning factors. Through experiments, we show, that the global reordering model significantly improves the translation accuracy of a standard Japanese-English translation task.

1 Introduction

Global reordering is essential to the translation of languages with different word orders. Ideally, a model should allow the reordering of any distance, because if we are to translate from Japanese to English, the verb in the Japanese sentence must be moved from the end of the sentence to the beginning just after the subject in the English sentence.

Standard phrase-based translation systems use a word distance-based reordering model in which non-monotonic phrase alignment is penalized based on the word distance between successively translated source phrases without considering the orientation of the phrase alignment or the identities of the source and target phrases (Koehn et al., 2003; Och and Ney, 2004). (Tillmann and Zhang, 2005) introduced the notion of a *block* (a pair of source and target phrases that are translations of each other), and proposed the block orientation bigram in which the local reordering of adjacent blocks are expressed as a three-valued orientation, namely Right (monotone), Left (swapped), or Neutral. A block with neutral orientation is supposed to be less strongly linked to its predecessor block: thus in their model, the global reordering is not explicitly modeled.

In this paper, we present a global reordering model that explicitly models long distance reordering[1]. It predicts four type of reordering patterns, namely MA (monotone adjacent), MG (monotone gap), RA (reverse adjacent), and RG (reverse gap). There are based on the identities of the source and target phrases currently being translated, and the previously translated source and target phrases. The parameters of the reordering model are estimated from the phrase alignments of training bilingual sentences. To cope with sparseness, we use N-best phrase alignments and bilingual phrase clustering.

In the following sections, we first describe the global phrase reordering model and its param-

*Graduated in March 2006

[1]It might be misleading to call our reordering model "global" since it is at most considers two phrases. A truly global reordering model would take the entire sentence structure into account.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 713–720,
Sydney, July 2006.

eter estimation method including N-best phrase alignments and bilingual phrase clustering. Next, through an experiment, we show that the global phrase reordering model significantly improves the translation accuracy of the IWSLT-2005 Japanese-English translation task (Eck and Hori, 2005).

2 Baseline Translation Model

In statistical machine translation, the translation of a source (foreign) sentence f is formulated as the search for a target (English) sentence $\hat{e}$ that maximizes the conditional probability $p(e|f)$, which can be rewritten using the Bayes rule as,

$$\hat{e} = \arg\max_e p(e|f) = \arg\max_e p(f|e)p(e)$$

where $p(f|e)$ is a translation model and $p(e)$ is a target language model.

In phrase-based statistical machine translation, the source sentence f is segmented into a sequence of I phrases $\bar{f}_1^I$, and each source phrase $\bar{f}_i$ is translated into a target phrase $\bar{e}_i$. Target phrases may be reordered. The translation model used in (Koehn et al., 2003) is the product of translation probability $\phi(\bar{f}_i|\bar{e}_i)$ and distortion probability $d(a_i - b_{i-1})$,

$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1}) \quad (1)$$

where a_i denotes the start position of the source phrase translated into the i-th target phrase, and b_{i-1} denotes the end position of the source phrase translated into the $(i-1)$-th target phrase.

The translation probability is calculated from the relative frequency as,

$$\phi(\bar{f}|\bar{e}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{f}} count(\bar{f}, \bar{e})} \quad (2)$$

where $count(\bar{f}, \bar{e})$ is the frequency of alignments between the source phrase $\bar{f}$ and the target phrase $\bar{e}$.

(Koehn et al., 2003) used the following distortion model, which simply penalizes non-monotonic phrase alignments based on the word distance of successively translated source phrases with an appropriate value for the parameter α,

$$d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|} \quad (3)$$

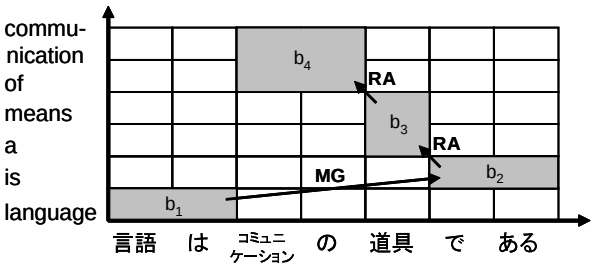

Figure 1: Phrase alignment and reordering

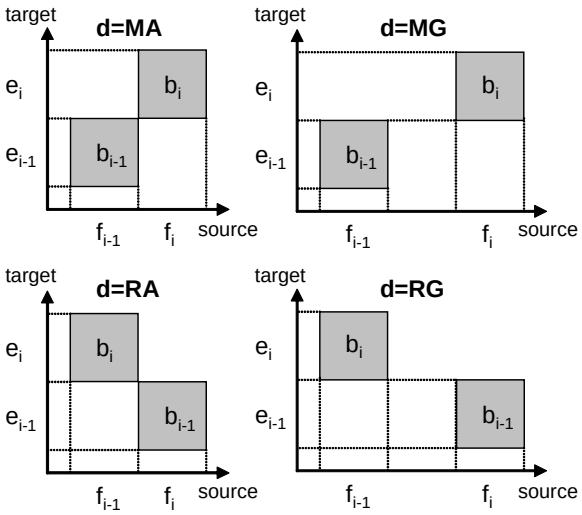

Figure 2: Four types of reordering patterns

3 The Global Phrase Reordering Model

Figure 1 shows an example of Japanese-English phrase alignment that consists of four phrase pairs. Note that the Japanese verb phrase "で_ある" at the the end of the sentence is aligned to the English verb "is" at the beginning of the sentence just after the subject. Such reordering is typical in Japanese-English translations.

Motivated by the three-valued orientation for local reordering in (Tillmann and Zhang, 2005), we define the following four types of reordering patterns, as shown in Figure 2,

- monotone adjacent (MA): The two source phrases are adjacent, and are in the same order as the two target phrases.
- monotone gap (MG): The two source phrases are not adjacent, but are in the same order as the two target phrases.
- reverse adjacent (RA): The two source phrases are adjacent, but are in the reverse order of the two target phrases.

	J-to-E	C-to-E
Monotone Adjacent	0.441	0.828
Monotone Gap	0.281	0.106
Reverse Adjacent	0.206	0.033
Reverse Gap	0.072	0.033

Table 1: Percentage of reordering patterns

- reverse gap (RG): The two source phrases are not adjacent, and are in the reverse order as the two target phrases.

For the global reordering model, we only consider the cases in which the two target phrases are adjacent because, in decoding, the target sentence is generated from left to right and phrase by phrase. If we are to generate the i-th target phrase $\bar{e}_i$ from the source phrase $\bar{f}_i$, we call $\bar{e}_i$ and $\bar{f}_i$ the current block b_i, and $\bar{e}_{i-1}$ and $\bar{f}_{i-1}$ the previous block b_{i-1}.

Table 1 shows the percentage of each reordering pattern that appeared in the N-best phrase alignments of the training bilingual sentences for the IWSLT 2005 Japanese-English and Chinese-English translation tasks (Eck and Hori, 2005). Since non-local reorderings such as monotone gap and reverse gap are more frequent in Japanese to English translations, they are worth modeling explicitly in this reordering model.

Since the probability of reordering pattern d (intended to stand for 'distortion') is conditioned on the current and previous blocks, the global phrase reordering model is formalized as follows:

$$p(d|\bar{e}_{i-1}, \bar{e}_i, \bar{f}_{i-1}, \bar{f}_i) \tag{4}$$

We can replace the conventional word distance-based distortion probability $d(a_i - b_{i-1})$ in Equation (1) with the global phrase reordering model in Equation (4) with minimal modification of the underlying phrase-based decoding algorithm.

4 Parameter Estimation Method

In principle, the parameters of the global phrase reordering model in Equation (4) can be estimated from the relative frequencies of respective events in the Viterbi phrase alignment of the training bilingual sentences. This straightforward estimation method, however, often suffers from sparse data problem. To cope with this sparseness, we used N-best phrase alignment and bilingual phrase clustering. We also investigated various approximations of Equation (4) by reducing the conditional factors.

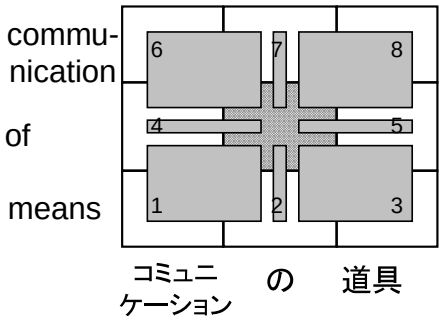

Figure 3: Expansion of a phrase pair

4.1 N-best Phrase Alignment

In order to obtain the Viterbi phrase alignment of a bilingual sentence pair, we search for the phrase segmentation and phrase alignment that maximizes the product of the phrase translation probabilities $p(\bar{f}_i|\bar{e}_i)$,

$$(\hat{\bar{f}}_1^I, \hat{\bar{e}}_1^I) = \arg\max_{\bar{f}_1^I, \bar{e}_1^I} \prod_{i=1}^{I} p(\bar{f}_i|\bar{e}_i) \tag{5}$$

Phrase translation probabilities are approximated using word translation probabilities $p(f_j|e_i)$ and $p(e_i|f_j)$ as follows,

$$p(\bar{f}|\bar{e}) = \prod_j \sum_i p(f_j|e_i) + p(e_i|f_j) \tag{6}$$

where f_j and e_i are words in the target and source phrases.

The phrase alignment based on Equation (5) can be thought of as an extension of word alignment based on the IBM Model 1 to phrase alignment. Note that bilingual phrase segmentation (phrase extraction) is also done using the same criteria. The approximation in Equation (6) is motivated by (Vogel et al., 2003). Here, we added the second term $p(e_i|f_j)$ to cope with the asymmetry between $p(f_j|e_i)$ and $p(e_i|f_j)$. The word translation probabilities are estimated using the GIZA++ (Och and Ney, 2003).

The above search is implemented in the following way:

1. All source word and target word pairs are considered to be initial phrase pairs.

2. If the phrase translation probability of the phrase pair is less than the threshold, it is deleted.

3. Each phrase pair is expanded toward the eight neighboring directions as shown in Figure 3.

4. If the phrase translation probability of the expanded phrase pair is less than the threshold, it is deleted.

5. The process of expansion and deletion is repeated until no further expansion is possible.

6. The consistent N-best phrase alignment are searched from all combinations of the above phrase pairs.

The search for consistent Viterbi phrase alignments can be implemented as a phrase-based decoder using a beam search whose outputs are constrained only to the target sentence. The consistent N-best phrase alignment can be obtained by using A* search as described in (Ueffing et al., 2002). We did not use any reordering constraints, such as IBM constraint and ITG constraint in the search for the N-best phrase alignment (Zens et al., 2004).

The thresholds used in the search are the following: the minimum phrase translation probability is 0.0001. The maximum number of translation candidates for each phrase is 20. The beam width is 1e-10, the stack size (for each target candidate word length) is 1000. We found that, compared with the decoding of sentence translation, we have to search significantly larger space for the N-best phrase alignment.

Figure 3 shows an example of phrase pair expansion toward eight neighbors. If the current phrase pair is (の, of), the expanded phrase pairs are (コミュニケーション_の, means_of), (の, means_of), (の_道具, means_of), (コミュニケーション_の, of), (の_道具, of), (コミュニケーション_の, of_communication), (の, of_communication), and (の_道具, of_communication).

Figure 4 shows an example of the best three phrase alignments for a Japanese-English bilingual sentence. For the estimation of the global phrase reordering model, preliminary tests have shown that the appropriate N-best number is 20. In counting the events for the relative frequency estimation, we treat all N-best phrase alignments equally.

For comparison, we also implemented a different N-best phrase alignment method, where

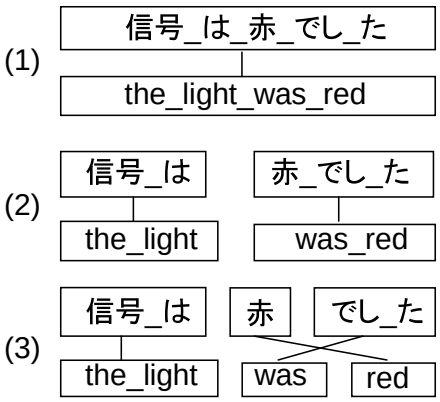

Figure 4: N-best phrase alignments

phrase pairs are extracted using the standard phrase extraction method described in (Koehn et al., 2003). We call this conventional phrase extraction method "grow-diag-final", and the proposed phrase extraction method "ppicker" (this is intended to stand for phrase picker).

4.2 Bilingual Phrase Clustering

The second approach to cope with the sparseness in Equation (4) is to group the phrases into equivalence classes. We used a bilingual word clustering tool, mkcls (Och et al., 1999) for this purpose. It forms partitions of the vocabulary of the two languages to maximize the joint probability of the training bilingual corpus.

In order to perform bilingual phrase clustering, all words in a phrase are concatenated by an underscore '_' to form a pseudo word. We then use the modified bilingual sentences as the input to mkcls. We treat all N-best phrase alignments equally. Thus, the phrase alignments in Figure 4 are converted to the following three bilingual sentence pairs.

```
信号_は_赤_でし_た
the_light_was_red
信号_は　赤_でし_た
the_light  was_red
信号_は　赤　でし_た
the_light  was  red
```

Preliminary tests have shown that the appropriate number of classes for the estimation of the global phrase reordering model is 20.

As a comparison, we also tried two phrase classification methods based on the part of speech of the head word (Ohashi et al., 2005). We defined (arguably) the first word of each English phrase and the last word of each Japanese phrase as the

shorthand	reordering model
baseline	$\alpha^{\lvert a_i - b_{i-1} - 1 \rvert}$
ϕ	$p(d)$
e[0]	$p(d\lvert\bar{e}_i)$
f[0]	$p(d\lvert\bar{f}_i)$
e[0]f[0]	$p(d\lvert\bar{e}_i, \bar{f}_i)$
e[-1]f[0]	$p(d\lvert\bar{e}_{i-1}, \bar{f}_i)$
e[0]f[-1,0]	$p(d\lvert\bar{e}_i, \bar{f}_{i-1}, \bar{f}_i)$
e[-1]f[-1,0]	$p(d\lvert\bar{e}_{i-1}, \bar{f}_{i-1}, \bar{f}_i)$
e[-1,0]f[0]	$p(d\lvert\bar{e}_{i-1}, \bar{e}_i, \bar{f}_i)$
e[-1,0]f[-1,0]	$p(d\lvert\bar{e}_{i-1}, \bar{e}_i, \bar{f}_{i-1}, \bar{f}_i)$

Table 2: All reordering models tried in the experiments

head word. We then used the part of speech of the head word as the phrase class. We call this method "1pos". Since we are not sure whether it is appropriate to introduce asymmetry in head word selection, we also tried a "2pos" method, where the parts of speech of both the first and the last words are used for phrase classification.

4.3 Conditioning Factor of Reordering

The third approach to cope with sparseness in Equation (4) is to approximate the equation by reducing the conditioning factors.

Other than the baseline word distance-based reordering model and the Equation (4) itself, we tried eight different approximations of Equation (4) as shown in Table 2, where, the symbol in the left column is the shorthand for the reordering model in the right column.

The approximations are designed based on two intuitions. The current block ($\bar{e}_i$ and $\bar{f}_i$) would probably be more important than the previous block ($\bar{e}_{i-1}$ and $\bar{f}_{i-1}$). The previous target phrase ($\bar{e}_{i-1}$) might be more important than the current target phrase ($\bar{e}_i$) because the distortion model of IBM 4 is conditioned on $\bar{e}_{i-1}$, $\bar{f}_{i-1}$ and $\bar{f}_i$. The appropriate form of the global phrase reordering model is decided through experimentation.

5 Experiments

5.1 Corpus and Tools

We used the IWSLT-2005 Japanese-English translation task (Eck and Hori, 2005) for evaluating the proposed global phrase reordering model. We report results using the well-known automatic evaluation metrics Bleu (Papineni et al., 2002).

IWSLT (International Workshop on Spoken Language Translation) 2005 is an evaluation campaign for spoken language translation.Its task domain encompasses basic travel conversations. 20,000 bilingual sentences are provided for training. Table 3 shows the number of words and the size of vocabulary of the training data. The average sentence length of Japanese is 9.9 words, while that of English is 9.2 words.

	Sentences	Words	Vocabulary
Japanese	20,000	198,453	9,277
English	20,000	183,452	6,956

Table 3: IWSLT 2005 Japanese-English training data

Two development sets, each containing 500 source sentences, are also provided and each development sentence comes with 16 reference translations. We used the second development set (devset2) for the experiments described in this paper. This 20,000 sentence corpus allows for fast experimentation and enables us to study different aspects of the proposed global phrase reordering model.

Japanese word segmentation was done using ChaSen[2] and English tokenization was done using a tool provided by LDC[3]. For the phrase classification based on the parts of speech of the head word, we used the first two layers of the Chasen's part of speech tag for Japanese. For English part of speech tagging, we used MXPOST[4].

Word translation probabilities are obtained by using GIZA++ (Och and Ney, 2003). For training, all English words are made in lower case. We used a back-off word trigram model as the language model. It is trained from the lowercased English side of the training corpus using a statistical language modeling toolkit, Palmkit [5].

We implemented our own decoder based on the algorithm described in (Ueffing et al., 2002). For decoding, we used phrase translation probability, lexical translation probability, word penalty, and distortion (phrase reordering) probability. Minimum error rate training was not used for weight optimization.

The thresholds used in the decoding are the following: the minimum phrase translation probability is 0.01. The maximum number of translation

[2]http://chasen.aist-nara.ac.jp/
[3]http://www.cis.upenn.edu/~treebank/tokenizer.sed
[4]http://www.cis.upenn.edu/~adwait/statnlp.html
[5]http://palmkit.sourceforge.net/

	ppicker		grow-diag-fi nal	
	class	lex	class	lex
baseline	0.400	0.400	0.343	0.343
ϕ	0.407	0.407	0.350	0.350
f[0]	0.417	0.410	0.362	0.356
e[0]	0.422	0.416	0.356	0.360
e[0]f[0]	0.422	0.404	0.355	0.353
e[0]f[-1,0]	0.407	0.381	0.346	0.327
e[-1,0]f[0]	0.410	0.392	0.348	0.341
e[-1,0]f[-1,0]	0.394	0.387	0.339	0.340

Table 4: BLEU score of reordering models with different phrase extraction methods

candidates for each phrase is 10. The beam width is 1e-5, the stack size (for each target candidate word length) is 100.

5.2 Clustered and Lexicalized Model

Figure 5 shows the BLEU score of clustered and lexical reordering model with different conditioning factors. Here, "class" shows the accuracy when the identity of each phrase is represented by its class, which is obtained by the bilingual phrase clustering, while "lex" shows the accuracy when the identity of each phrases is represented by its lexical form.

The clustered reordering model "class" is generally better than the lexicalized reordering model "lex". The accuracy of "lex" drops rapidly as the number of conditioning factors increases. The reordering models using the part of speech of the head word for phrase classification such as "1pos" and "2pos" are somewhere in between.

The best score is achieved by the clustered model when the phrase reordering pattern is conditioned on either the current target phrase $\bar{e}_i$ or the current block, namely phrase pair $\bar{e}_i$ and $\bar{f}_i$. They are significantly better than the baseline of the word distance-based reordering model.

5.3 Interaction between Phrase Extraction and Phrase Alignment

Table 4 shows the BLEU score of reordering models with different phrase extraction methods. Here, "ppicker" shows the accuracy when phrases are extracted by using the N-best phrase alignment method described in Section 4.1, while "grow-diag-final" shows the accuracy when phrases are extracted using the standard phrase extraction algorithm described in (Koehn et al., 2003).

It is obvious that, for building the global phrase reordering model, our phrase extraction method is significantly better than the conventional phrase extraction method. We assume this is because the proposed N-best phrase alignment method optimizes the combination of phrase extraction (segmentation) and phrase alignment in a sentence.

5.4 Global and Local Reordering Model

In order to show the advantages of explicitly modeling global phrase reordering, we implemented a different reordering model where the reordering pattern is classified into three values: monotone adjacent, reverse adjacent and neutral. By collapsing monotone gap and reverse gap into neutral, it can be thought of as a local reordering model similar to the block orientation bigram (Tillmann and Zhang, 2005).

Figure 6 shows the BLEU score of the local and global reordering models. Here, "class3" and "lex3"represent the three-valued local reordering model, while "class4" and "lex4"represent the four-valued global reordering model. "Class" and "lex" represent clustered and lexical models, respectively. We used "grow-diag-final" for phrase extraction in this experiment.

It is obvious that the four-valued global reordering model consistently outperformed the three-valued local reordering model under various conditioning factors.

6 Discussion

As shown in Figure 5, the reordering model of Equation (4) (indicated as e[-1,0]f[-1,0] in shorthand) suffers from a sparse data problem even if phrase clustering is used. The empirically justifiable global reordering model seems to be the following, conditioned on the classes of source and target phrases:

$$p(d|class(\bar{e}_i), class(\bar{f}_i)) \quad (7)$$

which is similar to the block orientation bigram (Tillmann and Zhang, 2005). We should note, however, that the block orientation bigram is a joint probability model for the sequence of blocks (source and target phrases) as well as their orientations (reordering pattern) whose purpose is very different from our global phrase reordering model. The advantage of the reordering model is that it can better model global phrase reordering using a four-valued reordering pattern, and it can be easily

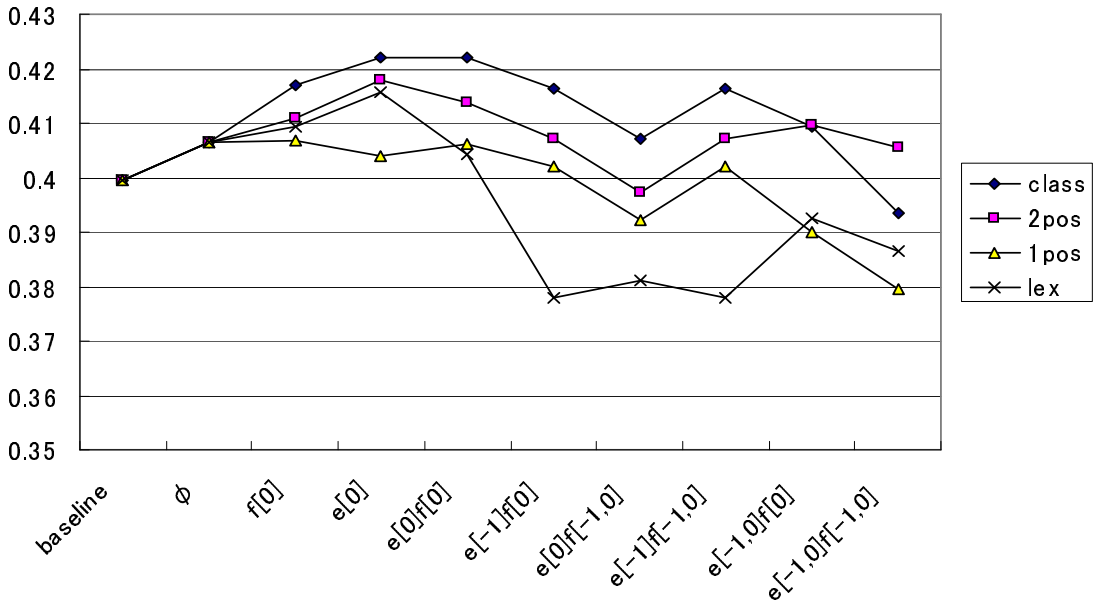

Figure 5: BLEU score for the clustered and lexical reordering model with different conditioning factors

incorporated into a standard phrase-based translation decoder.

The problem of the global phrase reordering model is the cost of parameter estimation. In particular, the N-best phrase alignment described in Section 4.1 is computationally expensive. We must devise a more efficient phrase alignment algorithm that can globally optimize both phrase segmentation (phrase extraction) and phrase alignment.

7 Conclusion

In this paper, we presented a novel global phrase reordering model, that is estimated from the N-best phrase alignment of training bilingual sentences. Through experiments, we were able to show that our reordering model offers improved translation accuracy over the baseline method.

References

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT 2005)*, pages 11–32.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL-03)*, pages 127–133.

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

Franz Josef Och and Herman Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Franz Josef Och, Christoph Tillman, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/WVLC-99)*, pages 20–28.

Kazuteru Ohashi, Kazuhide Yamamoto, Kuniko Saito, and Masaaki Nagata. 2005. NUT-NTT statistical machine translation system for IWSLT 2005. In *Proceedings of International Workshop on Spoken Language Translation*, pages 128–133.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Lnguistics (ACL-02)*, pages 311–318.

Christoph Tillmann and Tong Zhang. 2005. A localized prediction model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 557–564.

Nicola Ueffi ng, Franz Josef Och, and Hermann Ney. 2002. Generation of word graphs in statistical machine translation. In *Proceedings of the Conference*

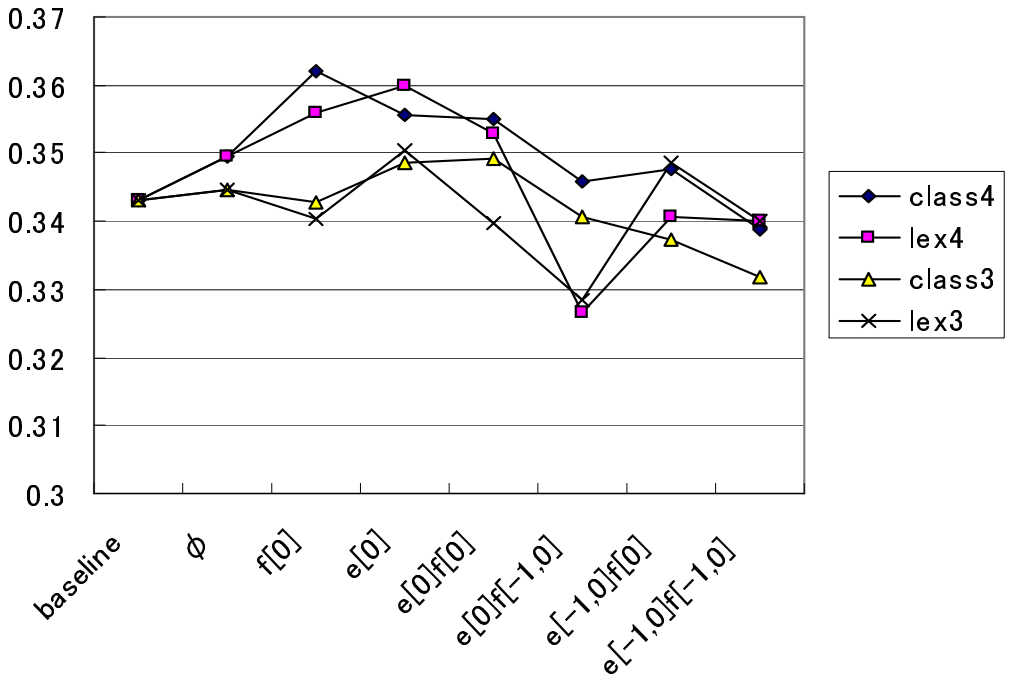

Figure 6: BLEU score of local and global reordering model

on Empirical Methods in Natural Language Processing (EMNLP-02), pages 156–163.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, and Alex Waibel. 2003. The CMU statistical machine translation system. In *Proceedings of MT Summit IX*.

Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of 20th International Conference on Computational Linguistics (COLING-04)*, pages 205–211.