NUT-NTT Statistical Machine Translation System for IWSLT 2005

Kazuteru Ohashi and *Kazuhide Yamamoto*

Nagaoka University of Technology
1603-1, Kamitomioka, Nagaoka City
Niigata, 940-2188 Japan
{ohashi, ykaz}@nlp.nagaokaut.ac.jp

Kuniko Saito and *Masaaki Nagata*

NTT Cyber Space Laboratories
1-1 Hikarinooka, Yokoshuka-shi
Kanagawa, 239-0847 Japan
{saito.kuniko, masaaki.nagata}@labs.ntt.co.jp

Abstract

In this paper, we present a novel distortion model for phrase-based statistical machine translation. Unlike the previous phrase distortion models whose role is to simply penalize nonmonotonic alignments[1, 2], the new model assigns the probability of relative position between two source language phrases aligned to the two adjacent target language phrases. The phrase translation probabilities and phrase distortion probabilities are calculated from the N-best phrase alignment of the training bilingual sentences. To obtain N-best phrase alignment, we devised a novel phrase alignment algorithm based on word translation probabilities and N-best search. Experiments show that the phrase distortion model and phrase translation model improve the BLEU and NIST scores over the baseline method.

1. Introduction

In recent years, phrase-based translation models have become the mainstream of statistical machine translation, because they can represent context-based word selection and local word reordering better than word-based translation models. Previous phrased-based translation models[1, 2], however, are not effective for global phrase reordering, because their distortion model is too simplistic. As it was designed simply to penalize nonmonotonic phrase alignment, it is difficult to handle translations that require complex word reordering, such as between Japanese and English.

In this paper, we present a novel distortion model for phrase-based statistical machine translation. It models the probability of relative position between two source language phrases aligned to the two adjacent target language phrases. To obtain the distortion model, we first make a phrase alignment of each sentence pair in the training corpus. We then calculate the phrase distortion probability from the relative frequency of respective events in the phrase aligned training corpus. In order to cope with the sparse data problem, word reordering is classified into four states: monotone, monotone-gap, reverse, and reverse-gap. Phrases are also classified based on the part of speech of the first and last word.

We need phrase translation probabilities to get phrase alignment from the training corpus, but we need phrase alignment to get phrase translation probabilities. To solve this chicken and egg problem, we devised a novel phrase alignment algorithm using word translation probabilities and forward beam search. Phrase distortion probabilities mentioned above are calculated from the result of this phrase alignment.

The phrase alignment algorithm can easily be extended to obtain N-best phrase alignment using backward A* search, such as [3]. We found that phrase translation probabilities calculated from the result of this N-best phrase alignment improve the translation accuracy significantly.

In the following sections, we first explain our translation model including the phrase distortion model and phrase alignment algorithm. We then report the experiments' results and show the effectiveness of our phrase distortion model.

2. Baseline Translation Model

In the noisy channel approach to machine translation, we search for the target (English) sentence $\hat{e}$ that maximizes the probability of the target sentence e given the source (foreign) sentence f. By using Bayes rule, the posterior probability $p(e|f)$ can be decomposed into the product of target sentence probability $p(e)$ and source sentence probability given target sentence $p(f|e)$.

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

Here, the models for computing $p(e)$ and $p(f|e)$ are called the language model and translation model, respectively.

In phrase-based statistical machine translation, 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 [1] is the product of translation probability $\phi(\bar{f}_i|\bar{e}_i)$ and relative distortion probability $d(a_i - b_{i-1})$.

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

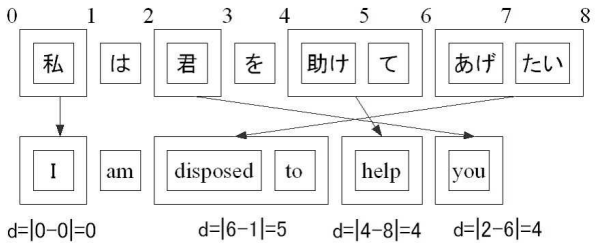

Figure 1: Example of relative distortion

where a_i denotes the start position of the source phrase that is 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.

Translation probability is calculated from the relative frequency of the respective source phrase given the target phrase.

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

where $count(\bar{f}, \bar{e})$ gives the frequency of the source phrase $\bar{f}$ aligned to the target phrase $\bar{e}$ in the parallel corpus. Note that, due to Bayes rule, the translation direction is inverted from a modeling standpoint.

The distortion model used in [1] is empirically defined as follows, with an appropriate value for parameter α.

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

Figure 1 illustrates the idea of relative distortion, using Japanese to English translation as an example. The target English sentence is generated from left to right by translating the source Japanese phrases in arbitrary order. Suppose we are generating target phrase "help" by translating the source phrase "助け て". The source phrase translated into the previous target phrase "disposed to" is "あげ たい". Since the start position of the source phrase for this target phrase a_i is 4, and the end position of the source phrase for previous target phrase b_{i-1} is 8, the relative distortion is $4 - 8 = -4$.

The purpose of the distortion model in Equation 3 is simply to penalize nonmonotonic phrase alignment. It cannot represent the general tendency of global phrase reordering, in terms of the distance and direction of the movement, as well as their dependency on phrase type. For example, for English to Japanese translation, the verb phrase generally moves toward the end of the sentence. In the next section, we present a novel phrase distortion model that considers these aspects.

3. Phrase Distortion Model

We define our phrase distortion model as the probability of relative distance between two source language phrases that are aligned to the two adjacent target language phrases,

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

where $\bar{e}_{i-1}$ and $\bar{e}_i$ are adjacent target phrases, $\bar{f}_{i-1}$ and $\bar{f}_i$ are source phrases aligned to $\bar{e}_{i-1}$ and $\bar{e}_i$, and d is the relative distance between $\bar{f}_{i-1}$ and $\bar{f}_i$.

Since the above distortion model involves too many parameters to estimate, we approximate it in several steps. First, we classify the relative distance d into four states:

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

We then classify each phrase by the part of speech of its head word. We define (arguably) the first word of each phrase as head word for English and Chinese, and the last word of each phrase as head word for Japanese.

Finally, we consider a series of distortion models that have increasingly complex dependencies.

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

where $class(\cdot)$ represents the classification of each phrase. When we classify each phrase by the part of speech of its head word, we identify the above five distortion models as type 1, 2s, 3s, 4s and 5s, respectively.

Figure 2 and Figure 3 show examples of phrase distortion models type 2s and type 3s, respectively, for Japanese to English translation. Here, monotone, monotone-gap, reverse, reverse-gap are represented by 1, 2, -1, -2, respectively. In Figure 3, the first three elements are d, $class(\bar{f}_i)$, $class(\bar{e}_{i-1})$, respectively. The fourth and fifth element are the distortion probability and frequency of this event in the training corpus.

Since we are not sure whether it is appropriate to define the head word of each phrase for each language a priori, we also tried "dual" distortion models, where $class(\cdot)$ of each phrase represented by both the first and the last word of each phrase. We call them type 2d, 3d, 4d, and 5d. An example of 3d is shown in Figure 4, where $class(\bar{f}_i)$ and $class(\bar{e}_{i-1})$ are represented by two POS tags.

```
...
-1 名詞-非自立 名詞-非自立 WRB WRB|0.34|17
-1 名詞-非自立 名詞-副詞可能 PRP PRP|0.75|3
-1 名詞-非自立 連体詞-連体詞 DT NNS|1|2
-1 名詞-副詞可能 記号-句点 NNP NNP|0.0526315789473684|1
-1 名詞-副詞可能 記号-句点 NNP TO|0.333333333333333|1
-1 名詞-副詞可能 記号-読点 . NN|1|1
...
```

Figure 4: Example of phrase distortion model type 3d

```
...
-1 名詞-非自立|0.456879958687386|9732
-1 名詞-副詞可能|0.380326288979142|5525
-1 連体詞-連体詞|0.0594823032223983|563
-2 フィラー-フィラー|0.578082191780822|422
-2 感動詞-感動詞|0.159919507575758|1351
-2 記号-句点|0.00304719568373694|1020
...
```

Figure 2: Example of phrase distortion model in type 2s

```
...
-1 名詞-非自立 WDT|0.676470588235294|69
-1 名詞-非自立 WP|0.360189573459716|152
-1 名詞-非自立 WRB|0.309219858156028|218
-1 名詞-副詞可能 ,|0.175824175824176|16
-1 名詞-副詞可能 .|0.216|27
-1 名詞-副詞可能 CC|0.130434782608696|3
...
```

Figure 3: Example of phrase distortion model type 3s

4. Phrase Alignment

The phrase distortion model in the previous section is computed from the Viterbi phrase alignment of the training corpus. In order to obtain this phrase alignment, we search for the segmentation of source and target sentences that maximizes the product of lexical translation probabilities $p(\bar{f}_i|\bar{e}_i)$,

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

Here, lexical translation probability [4] is an approximation of phrase translation probability based on the word translation probabilities estimated by using GIZA++[5],

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

where f_j and e_i are words in the phrases.

The phrase alignment is obtained by following these steps:

1. All pairs of one word from the source sentence and one word from the target sentence are considered as the phrase translation candidates.
2. If the lexical translation probability of a phrase translation candidate is less than the threshold, it is deleted.
3. Each phrase translation candidate is expanded toward its neighbors as described in [1].
4. If the lexical translation probability of the expanded phrase translation candidate is less than the threshold, it is deleted.
5. This expansion and deletion is repeated until no further expansion is possible.
6. Search for consistent phrase alignment among all combinations of the above phrase translation candidates.

We can obtain the Viterbi phrase alignment by using beam search from the beginning of the sentence to the end. We also can obtain the N-best phrase alignment by using A* search as described in [3].

Here, we must consider three parameters: phrase translation candidate threshold, beam width, and the number of N-best alignments. Preliminary tests have shown that the appropriate parameter is 1e-15 for phrase candidate threshold, 1000 for beam width, and 20 for the number of N-best. N-best phrase alignment is used for computing the phrase translation model, and Viterbi alignment is used for computing the phrase distortion model.

Figure 5 shows an example of the best 3 phrase alignments for a Japanese-English bilingual sentence. Each line represents a phrase translation candidate, where the first item is source phrase, second and third items are start and end positions of the phrase in the source sentence, fourth and fifth items are the parts of speech of the first and last words in the source phrase. After that, the same information for the target phrase is listed.

5. Corpus and Tools

We participated in Supplied Data + Tools Track in Japanese-English and Chinese-English translation because we need a part of speech tagger to obtain part of speech information

```
-
信号 は 赤 でし た|1|5|名詞-一般|助動詞-助動詞|the light was red|1|4|DT|JJ
。|6|6|記号-句点|記号-句点|.|5|5|.|.
2.71232e-06
-
信号 は|1|2|名詞-一般|助詞-係助詞|the light|1|2|DT|NN
赤 でし た|3|5|名詞-一般|助動詞-助動詞|was red|3|4|VBD|JJ
。|6|6|記号-句点|記号-句点|.|5|5|.|.
2.4524e-06
-
信号 は|1|2|名詞-一般|助詞-係助詞|the light|1|2|DT|NN
でし た|4|5|助動詞-助動詞|助動詞-助動詞|was|3|3|VBD|VBD
赤|3|3|名詞-一般|名詞-一般|red|4|4|JJ|JJ
。|6|6|記号-句点|記号-句点|.|5|5|.|.
2.38498e-06
```

Figure 5: Example of N-best phrase alignment for Japanese-English bilingual sentence

for our phrase distortion model. We did not use the word segmentation information of Japanese and Chinese provided in the supplied data because of the constraints of the POS tagger we used.

Word segmentation and POS tagging for Japanese was done by ChaSen[6]. As ChaSen's part of speech has a hierarchy, we used the first two layers. Word segmentation and POS tagging for Chinese was done by our own tool[7]. English is tokenized by a tool provided by LDC (tokenizer.sed)[8], and POS tagged by MXPOST[9]. Word translation probabilities are obtained by using GIZA++[5]. English are lowercased for training.

We used a back-off word trigram model as the language model. It is trained from the lowercased English side of the parallel training corpus using Palmkit[10].

For Japanese-English translation, we used a minimum error rate training tool provided by CMU[11]. The features used were the following:

- Phrase translation probability (both directions)[1]
- Lexical translation probability (both directions)[4]
- Word penalty[12]
- Phrase distortion probability

We didn't apply minimum error rate training to Chinese-English translation because we found no significant improvements for some reasons.

6. Experiments and Discussions

First, we compared our phrase extraction method with the conventional method described in [1]. Table 1 shows the NIST and BLEU scores for development set 2 in Japanese-English translation. We found that our phrase extraction method using N-best phrase alignment significantly improved the translation accuracy.

We then compared our phrase distortion model to the conventional distortion model[1]. Figure 6 shows the BLEU scores of the Japanese-English and Chinese-English translations created with various distortion models. Here, distortion model type 0 represents the conventional model [1]. Table 2 and Table 3 are NIST and BLEU scores for development set 2 of Japanese-English translation with various distortion models, before and after minimum error rate training. We found that, in general, distortion models type 2s and 3s yield a slight improvement in accuracy.

Table 1: Translation accuracy for development set 2 of Japanese-English with different phrase extraction methods

phrase extraction	NIST score	BLEU score
conventional	7.6162	0.3375
our method	8.8159	0.4471

Table 2: Translation accuracy for development set 2 of Japanese-English with different distortion models (before MER training)

distortion type	NIST score	BLEU score
0	8.7706	0.4050
1	8.9302	0.4219
2s	9.0435	0.4264
3s	8.9000	0.4179
4s	8.9419	0.4231
5s	8.8852	0.4168
2d	8.9904	0.4231
3d	8.9792	0.4214
4d	8.6711	0.3895
5d	8.7216	0.3959

In the experiments, the BLEU and NIST scores for distortion models 4d and 5d were generally very low. This is

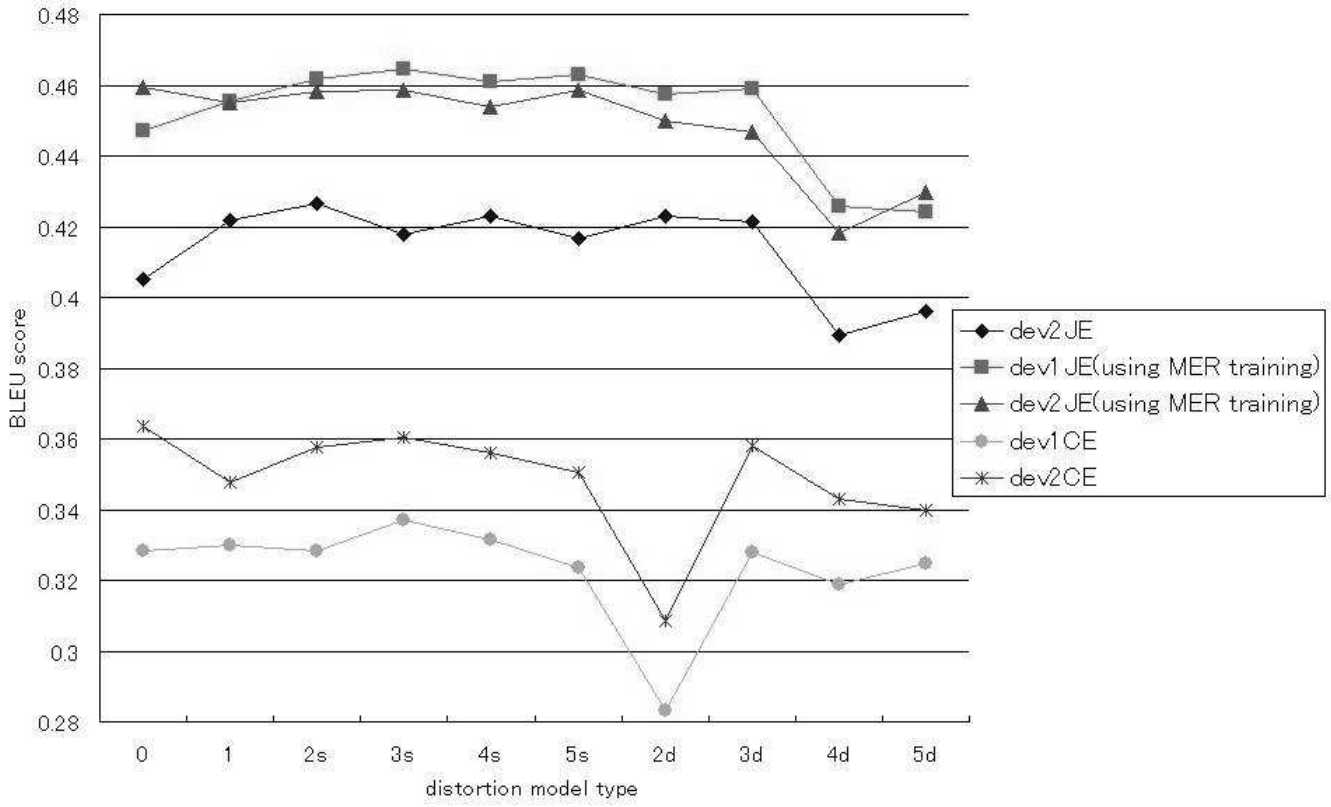

Figure 6: BLEU score of Japanese-English and Chinese-English translation with different distortion models

Table 3: Translation accuracy for development set 2 of Japanese-English with different distortion models (after MER training)

distortion type	NIST score	BLEU score
0	8.9551	0.4593
1	8.8916	0.4549
2s	8.9454	0.4581
3s	8.9846	0.4588
4s	8.9489	0.4539
5s	8.9995	0.4586
2d	8.8941	0.4500
3d	8.9219	0.4466
4d	8.8263	0.4181
5d	8.8829	0.4298

probably caused by data sparseness. The distortion model must consider 8 to 10 parts of speech using only the supplied data. The situation might be different if we had more training data.

We could not get phrase alignment for 1095 (5.5%) of the 20000 training sentences. In general, if the training parallel sentence is too long, we cannot get phrase alignment because of the large search space. As these sentences are not used for training at all, it probably hurt the performance significantly. Some countermeasure is needed, for example, limiting the search space for those long sentences by using the distortion model obtained from relatively short sentences.

In this experiment, the number of (N-best) phrase alignments for a sentence is fixed. This strategy is not the best because the number of plausible phrase alignments increases exponentially against sentence length. We must vary the number of alignments according to sentence length. It might be worth investigating other representation forms of phrase alignments, such as word graph.

7. Conclusion

In this paper, we present a novel phrase distortion model and a novel phrase alignment method for computing a more useful phrase distortion model. We show, by experiment, that the phrase distortion model described herein offers improved

translation accuracy over the baseline method.

8. References

[1] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *HLT-NAACL 2003: Main Proceedings*, M. Hearst and M. Ostendorf, Eds. Edmonton, Alberta, Canada: Association for Computational Linguistics, May 27 - June 1 2003, pp. 127–133.

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

[3] N. Ueffing, F. J. Och, and H. Ney, "Generation of word graphs in statistical machine translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Philadelphia: Association for Computational Linguistics, July 2002, pp. 156–163.

[4] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao, and A. Waibel, "The CMU statistical machine translation system," in *MT Summit IX*, New Orleans, USA, 23-27, September 2003.

[5] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," in *Computational Linguistics*, vol. 29, no. 1, 2003, pp. 19–51.

[6] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara, "Morphological analysis system chasen, ver.2.3.3," http://chasen.aist-nara.ac.jp/, 2003.

[7] K. Saito and M. Nagata, "Multi-language named entity recognition system based on hmm, acl2003, workshop on multilingual and mixed-language named entity recognition," 2003, pp. 41–48.

[8] R. MacIntyre, http://www.cis.upenn.edu/~treebank/tokenizer.sed, 1995.

[9] A. Ratnaparkhi, "Mxpost(maximum entropy pos tagger), ver.1.0," http://www.cis.upenn.edu/~adwait/statnlp.html, 1997.

[10] A. Ito, "Palmkit," http://palmkit.sourceforge.net/, 2002.

[11] A. Venugopal, http://www.cs.cmu.edu/~ashishv/mer.html, 2005.

[12] P. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine models, user manual and description for version 1.2," 2004.