Statistical Machine Translation Based on Hierarchical Phrase Alignment

Taro Watanabe, Kenji Imamura, Eiichiro Sumita
ATR Spoken Languge Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 JAPAN
{taro.watanabe, kenji.imamura, eiichiro.sumita}@atr.co.jp

Abstract

This paper describes statistical machine translation improved by applying hierarchical phrase alignment. The hierarchical phrase alignment is a method to align bilingual sentences phrase-by-phrase employing the partial parse results. Based on the hierarchical phrase alignment, a translation model is trained on a chunked corpus by converting hierarchically aligned phrases into a sequence of chunks. The second method transforms the bilingual correspondence of the phrase alignments into that of translation model. Both of our approaches yield better quality of the translaiton model.

1 Introduction

A statistical machine translation (SMT), first introduced by Brown et al. (1993), represents a translation process as a noisy channel model that consists of a source-channel model, a translation model and a prior, language model of target language texts. This transformed the problem of machine translation into a maximum posteriori solution to the source-channel paradigm.

The translation model is based on word-for-word translation and limited to allow only one channel source word to be aligned from a channel target word. Although phrasal correspondence is implicitly implemented into some translation models by means of distortion, careful parameter training is required. In addition, the training procedure relies on the EM algorithm, which can converge to an optimal solution but does not assured the global maximum parameter assignment. Furthermore, the translation models are represented by the numbers of parameters, so that easily suffered from the overfitting problem. In order to overcome these problems, simpler models, such as word-for-word translation models (Brown et al. 1993) or HMM models (Och & Ney 2000), have been introduced to determine the initial parameters and to bootstrap the training.

This paper describes two methods to overcome the above problems by using hierarchical phrase alignment (Imamura 2001). Hierarchical phrase alignment (HPA) is a method to align bilingual texts phrase-by-phrase from partial parse results. One method converts the hierarchically aligned phrasal texts into a pair of sequences of chunks of words, treating the word-for-word translation model as a chunk-for-chunk translation model. The second method computes the parameters for the translation model from the computed phrase alignments and uses the parameters as a starting point for training iterations.

The experimental results on Japanese-to-English translation indicated that the model trained from the parameters derived from the hierarchical phrase alignment could further improve the quality of translation from 61.3% to 70.0% in subjective evaluation. This results suggested

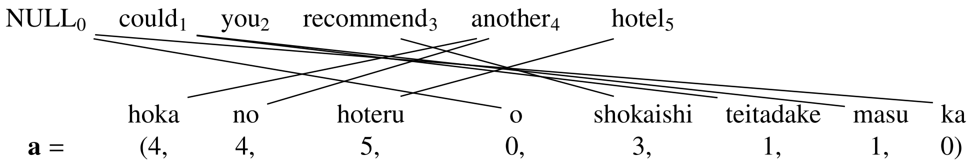
NULL₀ could₁ you₂ recommend₃ another₄ hotel₅
hoka no hoteru o shokaishi teitadake masu ka
a = (4, 4, 5, 0, 3, 1, 1, 0)

Figure 1: Example of alignment

that the hierarchical phrase alignment could boost the training parameters better than other stochastic based models.

The next section briefly describes statistical machine translation, mainly concentrating on, so called IBM 4. Then, after the explanation of hierarchical phrase alignment, the detailed procedure of applying the phrase aligned text to statistical machine translation is presented. Section 5 gives experimental results on Japanese-to-English translation followed by a discussion on the proposed methods.

2 Statistical Machine Translation

Statistical machine translation regards machine translation as a process of translating a source language text ($\mathbf{f}$) into a target language text ($\mathbf{e}$) with the following formula:

$$\mathbf{e} = \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f})$$

The Bayes Rule is applied to the above to derive:

$$\mathbf{e} = \arg\max_{\mathbf{e}} P(\mathbf{f}|\mathbf{e})P(\mathbf{e})$$

The translation process is treated as a noisy channel model, like those used in speech recognition in which there exists $\mathbf{e}$ transcribed as $\mathbf{f}$, and a translation is to infer the best $\mathbf{e}$ from $\mathbf{f}$ in terms of $P(\mathbf{f}|\mathbf{e})P(\mathbf{e})$. The former term, $P(\mathbf{f}|\mathbf{e})$, is a translation model representing some correspondence between bilingual text. The latter, $P(\mathbf{e})$, is the language model denoting the likelihood (or plausibility) of the channel source text. In addition, a word correspondence model, called alignment $\mathbf{a}$, is introduced to the translation model to represent a positional correspondence of the target and source words:

$$\mathbf{e} = \arg\max_{\mathbf{e}} \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e})P(\mathbf{e})$$

An example of an alignment is shown in Figure 1, where the English sentence "could you recommend another hotel" is mapped onto the Japanese "hoka no hoteru o shokaishi teitadake masu ka", and both "hoka" and "no" are aligned to "another", etc. The NULL symbol at index 0 is also a lexical entry in which no morpheme is aligned from the channel target morpheme, such as "masu" and "ka" in this Japanese example.

2.1 IBM Model 4

Many models have been suggested to denote the $P(\mathbf{f}, \mathbf{a}|\mathbf{e})$, including the so-called IBM Models 1 to 5 (Brown et al. 1993) and HMM model (Och & Ney 2000). The IBM Model 4 main forcus in this paper, is composed of the following models (refer to Figure 2):

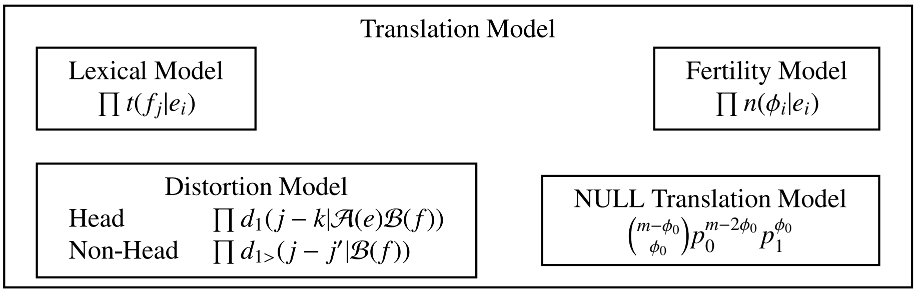
Translation Model
Lexical Model
$\prod t(f_j|e_i)$
Fertility Model
$\prod n(\phi_i|e_i)$
Distortion Model
Head $\prod d_1(j-k|\mathcal{A}(e)\mathcal{B}(f))$
Non-Head $\prod d_{1>}(j-j'|\mathcal{B}(f))$
NULL Translation Model
$\binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0}$

Figure 2: Translation Model (IBM Model 4)

- Lexical Model — $t(f|e)$: Word-for-word translation model, representing the probability of a source word f being translated into a target word e.
- Fertility Model — $n(\phi|e)$: Representing the probability of a source word e generating ϕ words.
- Distortion Model — d : The probability of distortion. In Model 4, the model is decomposed into two sets of parameters:
  - $d_1(j-k|\mathcal{A}(e),\mathcal{B}(f))$: Distortion probability for head words. The head word is the first of the target words generated from a source word by the fertility model . The head word position j is determined by the word classes of the previous source word, $\mathcal{A}(e)$, and target word, $\mathcal{B}(f)$, relative to the center of the previous source word, k.
  - $d_{>1}(j-j'|\mathcal{B}(f))$: Distortion probability for non-head words. The position of a non-head word j is determined by the word class and relative to the previous target word generated from the same source word (j).
- NULL Translation Model — p_1 : A fixed probability of inserting a NULL word after determining each target word f ($p_0 = 1 - p_1$).

For details, refer to Brown et al. (1993).

2.2 Problems in Statistical Machine Translation

In statistical machine translation, there exists three key problems as described below (Ney 2001):

Modeling Problem As this model suggests, a target word can be aligned to only a single source word. This restriction prohibits, for instance in Figure 1, “teitadake” from being mapped to both “could” and “you”, but allows only “could” to be mapped, and the other remaining source word, “you”, is treated as a zero fertility word. Och et al. (1999) introduced the concept of a translation template that could capture the phrase level correspondence, though the model relied on the HMM based translation model and could not be directly applied to fertility models such as the IBM Model 4.

Training Problem Training for the various parameters, t, n, p_1, d_1, $d_{>1}$ relies on the EM algorithm, which optimizes the log-likelihood of the model over a given bilingual corpus. The EM algorithm can find an optimal solution, although it cannot assure finding the globally best one. As the number of parameters is larger than those of speech it will become easily stuck at the local optimum solution.

To overcome this problem, simpler models, such as the IBM Model 1 or 2, were introduced to provide word-for-word translation models with uniform alignment probability (Brown et al. 1993). Those models first compute only the lexical model, or rough alignment probability. Och & Ney (2000) presented a HMM Model in which alignment was dependent on the previous word's alignment. The simpler models, like those briefly described above, can be used as intermediate models to train and improve the IBM Model 4 or 5 by using a boosting strategy. However, even computationally cheaper models rely on the EM-algorithm, so it is still not assured that they can discover the optimal model.

Search Problem This problem is not handled here, although the search problem is a critical issue for the success of statistical machine translation. The decoder, or the search system, should induce the source string from a sequence of target words by utilizing clues from a large numbers of parameters. Basically, if the vocabulary size is 10,000 and the output sentence length is 10, then 10000^{10} possible candidates must be enumerated. In addition, since the source sentence length is unknown to the decoder, the search system should also infer the total length of output at the same time. For details of the search problem, refer to Germann et al. (2001); Knight (1999); Och et al. (2001).

3 Hierarchical Phrase Alignment

Hierarchical phrase alignment, proposed by Imamura (2001), computes the correspondence of sub-trees between source language and target language parse trees based on partial parse results. A phrase alignment is defined as an equivalent sequence of words between bilingual sentences, and it may be a sequence of words representing noun phrases and/or verb phrases etc. For instance, the sentence pairs,

E: I have just arrived in Kyoto .
J: kyoto ni tsui ta bakari desu .

consists of three phrase alignments:

in Kyoto	—	kyoto ni
arrived in Kyoto	—	kyoto ni tsui
have just arrived in Kyoto	—	kyoto ni tsui ta bakari desu

The phrase alignments are first computed by tagging and parsing the bilingual sentences. After the parse, word level alignment, or word-linkage, is computed and the partial parses (non-terminals) are pruned out if there exists no word-linkage. The remaining partial parses are compared by using the similarity of the syntactic categories. If multiple candidates of a sentence or auxiliary verb phrases are acquired, the partial parse with maximum coverage is selected. For other syntactic categories, those of minimal coverage are chosen for phrase alignment. Figure 3 illustrates an example of phrase alignment for the above example, in which phrase alignments **NP(1)**, **VMP(2)**, **VP(3)**, **VP(4)**, **AUXVP(5)**, and **S(6)** were computed with the word-linkage "arrived"—"tsui" and "Kyoto"—"kyoto".

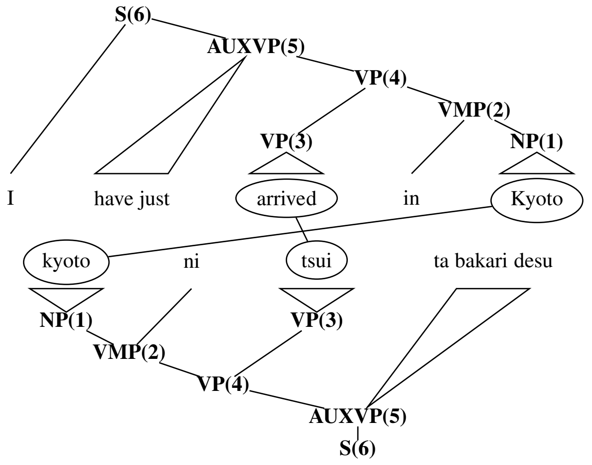

Figure 3: Example of phrase alignment

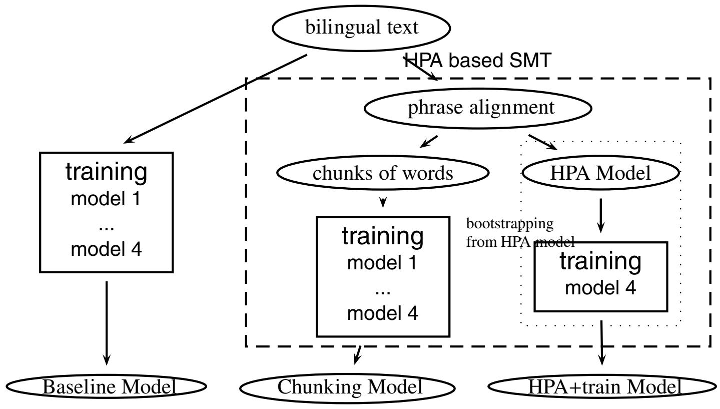

Figure 4: Statistical machine translation based on hierarchical phrase alignment

4 Applying Phrase Alignment to Statistical Machine Translation

Figure 4 illustrates the procedures for transforming the correspondence in hierarchical phrase alignment into that of statistical machine translation.

One is to convert the hierarchical phrase alignments into chunks of words and use the sequence of chunks as bilingual texts for training (chunking model). This enlarges the size of vocabulary, although the multiple target word associations are implicitly implemented into chunks. This method is expected to resolve the first problem of statistical machine translation, that is the model problem where the model restricts only one source word aligned from a target word. In addition, improved quality of the various parameters is expected. For instance, the *t* parameters, the lexical model, will be improved because of the strong correspondence of phrase alignments and chunks. The distortion model, *d* parameters, will also be improved as the length of the bilingual sentence is shortened.

The second method is to compute the translation model parameters from the phrase alignments (HPA model) and use the model parameters in a bootstrapping strategy for training

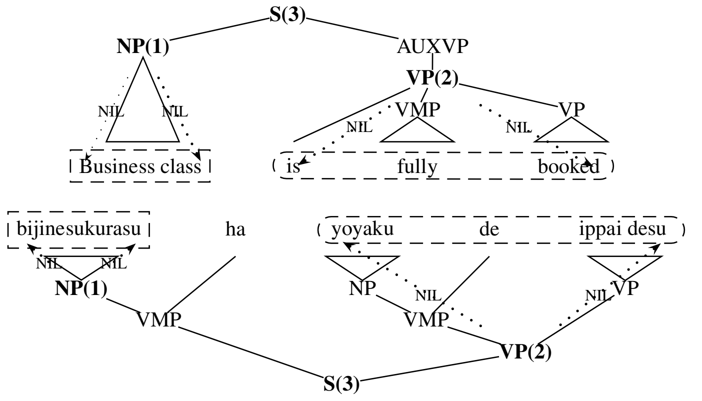

Figure 5: Example of converting phrase alignment into chunks

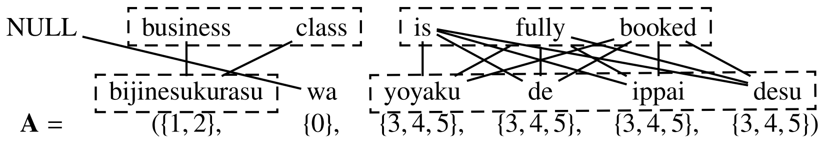

Figure 6: Example of restricted alignment

(HPA+train model). From the phrase alignment, since the sequence of words in a chunk is already aligned, the subset of alignments can be obtained, and the computed alignments can be applied to derive the parameters for the translation model.

4.1 Phrase Alignment as Chunk of Words

The hierarchical structures were converted to a sequence of chunks by inserting NIL symbols before and after the phrase alignments. After the insertion, the chunks were obtained by separating the sequence of words by NILs. For example, in Figure 5, NIL symbols are inserted before and after the phrase alignments **NP(1)**, **VP(2)**, and **S(3)**. Then, chunks of "business class" and "is fully booked" are derived for English, and "bijinesukurasu" and "yoyaku de ippai desu" are extracted for Japanese.

The chunking model can be computed by first transforming the The sequence of words in a corpora into the corresponding sequence of chunks. the derived texts are consumed as the inputs for training/decoding by treating each chunk as a word.

4.2 Phrase Alignment as Translation Model Parameters

Given a sequence of chunks of words derived by the above procedure, restricted alignments are first computed. The restricted alignments are subsets of all possible alignments given a bilingual sentence but are limited by the boundary hypothesized by the aligned chunks of words. For instance, Figure 6 illustrates an example of restricted alignments where the target word "bijinesukurasu" is aligned to only "business" or "class", and so on.

In computing the HPA model, t, n, d_1, $d_{>1}$ and p_0 parameters can be derived by simply enumerating all the restricted alignments and accumulating counts that correspond to one iteration

Table 1: Corpus

	English	Japanese
number of sentences	145,432	
number of words	835,048	896,302
vocabulary size	13,162	20,348
average sentence length	5.74	6.16
trigram perplexity	36.03	32.93

perplexity: evaluated by trigram language model using Clarkson & Rosenfeld (1997).

Table 2: Aligned chunks

	English	Japanese
number of chunks	7,604	6,750
vocabulary size (of chunks)	2,166	1,624
average number of chunks per sentence	0.759	0.673
average number of words per chunk	2.21	2.52

of training. For example, the t parameters can be computed by first accumulating the counts:

$$tc(f|e; \mathbf{f}, \mathbf{e}, \mathbf{A}) = \sum_{\mathbf{a} \in \mathbf{A}} P(\mathbf{a}|\mathbf{f}, \mathbf{e}) \sum_{i,j} \delta(f, f_j)\delta(e, e_{a_j})$$

Then, t parameters are computed as:

$$t(f|e) \leftarrow \sum_{s \in train} tc(f|e; \mathbf{f}_s, \mathbf{e}_s, \mathbf{A}_s)$$

where $P(\mathbf{a}|\mathbf{f}, \mathbf{e})$ is considered uniform distribution.

Given the HPA model as initial parameters, the HPA+train model can be trained further with the EM-algorithm as described in Brown et al. (1993).

5 Experiments

5.1 Corpus

The corpus for this experiment consists of 145,432 bilingual sentences of English and Japanese extracted from a large-scale travel conversation corpus. The statistics of the corpus are shown in Table 1. The corpus was split into three parts: a training set of 125,537 sentence pairs, a validation set of 9,872, and a test set of 10,023.

From the training set, 355,293 hierarchical phrase alignments were extracted, and 177,242 chunks for English and 159,884 chunks for Japanese were created. Among the chunks, those with frequency higher than 10 were selected. The final statistics on chunks are shown in Table 2.

5.2 Models

In this experiments, four models were created for Japanese-to-English translation: One was the baseline model, trained from the IBM Models 1 to 4, including the HMM model as suggested by Och & Ney (2000) [1]. The next model, the chunking model, was derived from chunks of

[1] the training program is based on GIZA++ (Och 2001)

Table 3: Experimental results of translation

Model	WER	PER	SE			
			A	B	C	D
baseline	0.702	0.592	12.7%	33.3%	14.7%	38.7%
chunking	0.640	0.531	21.3%	28.0%	16.7%	34.0%
HPA	0.645	0.581	17.3%	32.0%	15.3%	35.3%
HPA+train	0.710	0.593	16.0%	32.0%	22.0%	30.0%

WER: word error rate
PER: position independent word error rate
SE: subjective evaluation (A: perfect, B: fair, C: acceptable, D: nonsense)

baseline: baseline model
chunking: chunking model
HPA: IBM Model 4 with parameters derived from HPA
HPA+train: IBM Model 4 boosted from HPA model

words by transforming the hierarchical phrase alignments into chunks. The sequence of chunks are treated as a sequence of words, and trained as the baseline model. The third model, the HPA model, was also computed from the hierarchical phrase alignments, but by converting them into translation model's parameters. The HPA model was also evaluated to see if the derived model were trained properly on the training corpus. The last model, the HPA+train model, was obtained by further training the HPA model only for IBM Model 4 iterations. The baseline, chunking and HPA+train models were all trained on the training set, and their training loop was terminated when the perplexity for the validation set indicated the lowest scores (cross validation).

5.3 Results

The four models were tested on 150 bilingual sentences, 50 pairs each for Japanese input sentence lengths of 6, 8 and 10. For this experiment, a decoder was developed based on a stack decoding algorithm presented in Berger et al. (1996) with the very weak admissible heuristic function described in Och et al. (2001). All of the sentences were evaluated by word error rate (WER) and subjective evaluation (SE), with criteria ranging from A (best) to D (worst)[2] (Sumita et al. 1999). The evaluation measure, word error rate (WER), cannot take into account the fact that a channel source string can be re-ordered differently; hence we also introduced the position-independent word error rate (PER), in which the positional disfluencies were not considered (Och et al. 2001).

Table 3 summarizes the results of WER, PER and SE. The results with varying input sentence lengths are presented in Table 4, with WER, PER and SE with ranks from A to C. Some samples of translation results are illustrated in Figure 7 with the subjective evaluation rank for the Baseline Model and the HPA+train model.

6 Discussion

From Table 3, the chunking model greatly reduced both of WER and PER, and improved the translation quality based on the subjective evaluation. The chunking seems to successfully

[2] the meanings of the symbol are follows: A — perfect: no problem in either information or grammar; B — fair: easy to understand but some important information is missing or it is grammatically flawed; C — acceptable: broken but understandable with effort; D — nonsense: important information has been translated incorrectly.

Table 4: Experimental results of translation by sentence length

Model	WER			PER			SE(A+B+C)		
length	6	8	10	6	8	10	6	8	10
baseline	0.666	0.675	0.766	0.568	0.607	0.600	66.0%	64.0%	52.0%
chunking	0.545	0.570	0.806	0.484	0.489	0.620	78.0%	72.0%	48.0%
HPA	0.595	0.657	0.684	0.553	0.607	0.584	72.0%	66.0%	56.0%
HPA+train	0.643	0.726	0.762	0.558	0.625	0.597	78.0%	72.0%	60.0%

SE(A+B+C): subjective evaluation with the rank of A, B or C

input:	steki no yaki guai wa dou sa re masu ka		
output:	baseline:	(D)	can you steak
	HPA+train:	(A)	how do you like your steak
input:	gorufujo no yoyaku deki masu ka		
output:	baseline:	(C)	can i make a reservation
	HPA+train:	(A)	could you make a reservation for the golf course
input:	shikago kara shiatoru made donokurai jikan ga kakari masu ka		
output:	baseline:	(A)	how long does it take to seattle from chicago
	HPA+train:	(B)	do you how long will it take to seattle from chicago
input:	seki no kakuho dewa kuregruemo saiko no tokoro o onegaishimasu (please be sure to secure the best available seats for us)		
output:	baseline:	(B)	i would like a seat in a great place please
	HPA+train:	(D)	my best regards to your seat find a place please
input:	shoshinsha na no dakedo sankashi temo ii desu ka (i am a beginner may i join)		
output:	baseline:	(D)	do you have may but take beginner
	HPA+train:	(D)	it is but i am a beginner

Figure 7: Samples of translation

model the multiple source words aligned to one target word. However, this model was only good for shorter sentences, and the quality degraded for longer input sentences (refer to Table 4). This is partly due to the data sparseness problem for the vocabulary size was increased by the chunked words and also increases is the number of possible translations. Accordingly, it became more difficult for the decoder to discover the best translation.

As with the chunking model, the HPA model reduced WER and PER and improved the subjective evaluation rating. This indicates that the parameters derived from the hierarchical phrase alignment could correctly assign parameters better than those derived from the baseline model.

The last model, the HPA+train model, could not decrease the values for WER and PER, but still greatly increased the score of the subjective evaluation. This showed that the HPA+train model, trained further from the HPA model could achieved better parameter assignment. In comparing this model to the chunking model, the results of the subjective evaluation were very close, though the HPA+train model is better for longer input length.

The results above demonstrate that the translation model parameters derived from the hierarchical phrase alignment were better than those acquired only by EM-training on a given corpus.

One of the advantages of using hierarchical phrase alignment is that it is neutral to language pairs: They can share the same parsing system with a simple algorithm for aligning texts. The next advantage is the robustness to the input, since the phrase alignments are extracted from partial parse results. These advantages were not available in alignment methods of Yamamoto & Matsumoto (2000); Kaji et al. (1992). In addition, hierarchical phrase alignment is different from other chunking methods bacause it can preserve the correspondence in bilingual texts. Although the proposed method here did not use the higher level structures in the hierarchically aligned phrases, it will be challenging to incorporate those alignments restricted by non-terminal correspondences.

The quality of translation is expected to be improved by including the restricted alignments into training steps. This idea is based on pegging (Brown et al. 1993) or from the work of Och & Ney (2000), in which a subset of all the alignments were used for training based on the viterbi alignment, the best scored alignment, and neighboring alignments. Instead of limiting the alignments to those from trained parameters, the alignments obtained from hierarchical phrase alignments will be able to guide the training to a better model.

It will also be interesting to test a combination of the chunking model and the restricted alignments. From the results, the chunking could help improve quality by translating language into A-ranked sentences, while the HPA+train model could serve to improve overall quality by suppressing the translation into D-ranked sentences.

Acknowledgement

We wish to thank our colleagues for useful discussions, especially for Dr. Yamamoto and Mr. Shirai of ATR Spoken Language Translation Laboratories. We are also grateful to Prof. Okuno and Prof. Kawahara of Kyoto University for their advice in pursuitting this research.

References

Berger, A., P. Brown, S. Pietra, V. Pietra, J. Gillett, A. Kehler & R. Mercer: 1996, ‘Language translation apparatus and method of using context-based translation models’, Tech. rep., United States Patent, Patent Number 5510981.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra & Robert L. Mercer: 1993, ‘The mathematics of statistical machine translation: Parameter estimation’, *Computational Linguistics*, **19**(2): 263–311.

Clarkson, Philip & Ronald Rosenfeld: 1997, ‘Statistical language modeling using the CMU-cambridge toolkit’, in *Proc. Eurospeech ’97*, Rhodes, Greece, pp. 2707–2710.

Germann, Ulrich, Michael Jahr, Kevin Knight, Daniel Marcu & Kenji Yamada: 2001, ‘Fast decoding and optimal decoding for machine translation’, in *Proceedings of ACL-01*, Toulouse, France.

Imamura, Kenji: 2001, ‘Hierarchical phrase alignment harmonized with parsing’, in *Proc. of NLPRS 2001, Tokyo*.

Kaji, H., Y. Kida & Y. Morimoto: 1992, ‘Learning translation templates from bilingual texts’, in *Proceeding of COLING 92*, pp. 672–678.

Knight, Kevin: 1999, ‘Decoding complexity in word-replacement translation models’, *Computational Linguistics*, **25**(4): 607–615.

Ney, Hermann: 2001, ‘Stochastic modeling: From pattern classification to language translation’, in *Proceedings of the ACL-2001 Workshop on Data-Driven Machine Translation*, Toulouse, France, pp. 33–37.

Och, F. J. & H. Ney: 2000, ‘Improved statistical alignment models’, in *ACL 2000*, Hongkong, China, pp. 440–447.

Och, Franz Josef: 2001, ‘Giza++: Training of statistical translation models’, http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html.

Och, Franz Josef, Christoph Tillmann & Hermann Ney: 1999, ‘Improved alignment models for statistical machine translation’, in *Proceedings of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora*, University of Maryland, College Park, MD.

Och, Franz Josef, Nicola Ueffing & Hermann Ney: 2001, ‘An efficient a* search algorithm for statistical machine translation’, in *Proceedings of the ACL-2001 Workshop on Data-Driven Machine Translation*, Toulouse, France, pp. 55–62.

Sumita, Eiichiro, Setsuo Yamada, Kazuhide Yamamoto, Michael Paul, Hideki Kashioka, Kai Ishikawa & Satoshi Shirai: 1999, ‘Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach’, in *Machine Translation Summit VII*, pp. 229–235.

Yamamoto, Kaoru & Yuji Matsumoto: 2000, ‘Acquisition of phrase-level bilingual correspondence using dependency structure’, in *Proceedings of COLING-2000*, pp. 933–939.