Extracting Phrasal Alignments from Comparable Corpora by Using Joint Probability SMT Model

Tadashi Kumano[†,‡] **Hideki Tanaka**[†]
[†]NHK Science and Technical Research Laboratories
Tokyo, JAPAN 157-8510
{kumano.t-eq,tanaka-h.ja}@nhk.or.jp

Takenobu Tokunaga[‡]
[‡]Department of Computer Science
Tokyo Institute of Technology
Tokyo, JAPAN 152-8552
take@cl.cs.titech.ac.jp

Abstract

We propose a method of extracting phrasal alignments from comparable corpora by using an extended phrase-based joint probability model for statistical machine translation (SMT). Our method does not require preexisting dictionaries or splitting documents into sentences in advance. By checking each alignment for its reliability by using log-likelihood ratio statistics while searching for optimal alignments, our method aims to produce phrasal alignments for only parallel parts of the comparable corpora. Experimental result shows that our method achieves about 0.8 in precision of phrasal alignment extraction when using 2,000 Japanese-English document pairs as training data.

1 Introduction

Comparable corpora as a source of translation knowledge have attracted the attention of many researchers. Comparable corpora are composed of document pairs describing the same topic in different languages. They are not *parallel* (mostly word-to-word translated) corpora composed of good bilingual sentence pairs, but still contain various levels of parallelism, such as words, phrases, clauses, sentences, and discourses, depending on the corpora characteristics. Compared with parallel corpora, comparable corpora are much easier to build from commonly available documents, such as news article pairs describing the same event in different languages.

Recently, many studies on automatic acquisition of parallel parts from noisy non-parallel corpora have been conducted to acquire larger training corpora for statistical machine translation (SMT). One of the recent studies tried to find parallel sentences (Zhao and Vogel, 2002; Munteanu and Marcu, 2002; Fung and Cheung, 2004), and another tried to extract sub-sentential parallel fragments (Munteanu and Marcu, 2006). To detect the parallel parts, most of these studies required good statistical bilingual dictionaries, which are extracted from parallel corpora. Here we face "the chicken or the egg" problem. Previous studies use preexisting parallel corpora as bootstraps to prepare dictionaries, but it would be better to obtain lexical translation knowledge and extract parallel parts (eliminate unrelated parts) from comparable corpora simultaneously without parallel corpora.

In this paper, we propose an extension of the phrase-based joint probability model for SMT proposed by Marcu and Wong (2002). Our method can extract phrase alignments directly from comparable document pairs, without preexisting dictionaries or preprocessing of training data such as splitting it into sentences or extracting parallel parts. To prevent from producing alignments between unrelated phrases while searching for optimal alignments, we check each alignment as to

Original Japanese script:	Script translated into English:
1: 地震が続いている伊豆諸島できょう午前六時四十二分頃強い地震があり式根島で震度五弱を観測しました。 (There was a strong earthquake in the Izu Islands at 6:42 this morning, and the quake was measured the intensity of five-minus on the Japanese scale of seven at Shikine Island. A series of earthquakes have recently occurred around Izu Islands.)	1: A strong earthquake jolted Shikine Island, one of the Izu islands south of Tokyo, early on Thursday morning.
2: このほか震度四が新島、神津島、震度三が利島、三宅島、また関東各地や静岡県の一部で震度二や一の揺れを観測しました。 (The measurements of the quake at other places are as follows: intensities of four at Niijima and Kozu Islands, three at Toshima and Miyake Islands, and two or one at several places in the Kanto area and a part of Shizuoka Prefecture.)	2: The Meteorological Agency says the quake measured five-minus on the Japanese scale of seven.
3: この地震による津波の心配はありません。 (Official says there will be no fear of tsunamis caused by this earthquake.)	3: The quake affected other islands nearby.
4: 気象庁の観測によりますと震源地は新島・神津島の近海で震源の深さは十キロ、地震の規模を示すマグニチュードは五点一と推定されています。 (According to the observation of the Meteorological Agency, the center of the earthquake was 10 kilometers under the the sea bottom near Niijima and Kozu Islands, and the magnitude was 5.1.)	4: Seismic activity began in the area in late July, and 17 quakes of similar or stronger intensity have occurred.
5: 六月末から地震活動が始まった伊豆諸島では活動が活発な状態とやや落ち着いた状態を繰り返していて、先月三十日も三宅島で震度六弱の強い地震を一回観測した他震度五強の地震が二回起きました。 (Intermittent seismic activity began in the Izu Islands in late July, and the recent quakes were observed on the 30th of last month, once with an intensity of six-minus at Miyake Island and twice with an intensity of five-minus nearby.)	5: Officials are warning of more similar or stronger earthquakes around Niijima and Kozu Islands.
6: これらの地震を含めて一連の地震活動では神津島や新島、三宅島で震度六弱の強い揺れを四回観測したのを含めてこれまでに震度五弱以上の地震が十七回起きています。 (17 quakes with intensities of five-minus or higher including the recent ones have occurred during the activity, including four strong quakes with intensities of six-minus observed at Kozu, Niijima and Miyake Islands.)	6: Tokyo police say there have been no reports of damage from the latest quake.

Figure 1: Example article pair from the NHK Japanese-English news corpus

whether it is a statistically reliable translation by using log-likelihood ratio (LLR) statistics. The experimental results on our extension of Marcu-Wong's Model 1 shows that it is effective for extracting phrase alignments from comparable corpora. Those phrasal alignments are useful in applications other than machine translation. For example, we are developing a comparable translation retrieval system for supporting professional translators. The system will be more effective if it is able to show how a part in a source document is translated in a counterpart in response to the user's requests.

Section 2 introduces the Japanese-English broadcast news corpus, which is the target of our proposing method, and explains our tasks. Section 3 explains our improvements to the phrase-based joint probability model of Marcu and Wong in order to apply it to comparable corpora. After that, we show the results of our preliminary alignment experiment and discuss the effectiveness of our method in Section 4. Section 5 refers to related works and Section 6 concludes our paper.

2 Alignment Task for NHK Japanese-English News Corpus

We have been studying possible alignment methods for our comparable corpus, the NHK

Japanese-English news corpus, which is composed of pairs of Japanese news scripts and their manual translations into English broadcasted by NHK (Japan Broadcasting Corporation)[1]. The articles in Japanese and English in our corpus respectively have about 5 and 8 sentences on average.

An example article pair is shown in Figure 1 (The Japanese article is provided with a literal English translation for convenience). This example shows that the article pair shares the same topic, but each article describes the topic in a different style. Some articles have partially different content from their counterparts. Therefore, few parallel sentence pairs can be found in this corpus. At the level of words or shorter collocations, many useful translations can be found. However, words or phrases in a sentence are often translated into different sentences in the counterpart language. Thus, if you estimate word or phrase alignments from this type of comparable corpora, you have to search the whole document of the counterpart language.

3 Extension of Phrase-Based Joint Probability Model

Marcu and Wong (2002) proposed a joint probability model. It models how source and target sentences are simultaneously generated by *concepts*. Many of the phrase-based SMT models require word-level alignments for extracting phrases from combinations of the alignments. On the other hand, their training method can learn word and phrase alignments at the same time for searching for optimal alignments among possible partial word sequences in sentence pairs. There was a report that the joint probability model achieved better performance on SMT, especially for small-sized training data (Birch et al., 2006).

The formulation of Marcu-Wong model can be simply extended to non-parallel corpora by adding a means of handling monolingual phrases appearing independently of any counterpart. The search for optimal phrase alignments in their training method can be straightforwardly viewed as finding the parallel parts in a comparable document pairs. Therefore, we choose to employ their joint probability model for comparable corpora.

The main difficulty of the extension is the arbitrariness of deciding how many portions in each of the document pairs should be considered as unrelated to the counterpart document. We try to resolve the difficulty with the help of the log-likelihood ratio statistics to distinguish reliably correlated translations from unrelated parts.

3.1 Model Formulation

The original joint probability model assumes that every part of the sentences on the source and target sides is composed of phrases generated from *concepts*. We extended the model so that comparable document pairs have not only parallel phrases that share *concepts* but also non-parallel phrases that are independent of the counterpart document.

We consider a concept so that they can generate a monolingual phrase only on either side of a document pair. Under this definition, we can use the following formula, which is the same as the Marcu-Wong method, to express the probability of generating a document pair $(\boldsymbol{e}, \boldsymbol{f})$ which may have non-parallel phrases:

$$p(\boldsymbol{e}, \boldsymbol{f}) = \sum_{C \in \{\mathcal{C} | L(\boldsymbol{e}, \boldsymbol{f}, C)\}} \prod_{c_i \in C} t(\vec{e}_i, \vec{f}_i), \quad (1)$$

where

$\vec{e}$, $\vec{f}$: source and target phrases which are empty ( ) or consist of sequences of one or more words,

c_i: a concept to generate a pair of source and target phrases $(\vec{e}, \vec{f})$ only one side of which can be . Each concept produces a unique pair of phrases (or a monolingual phrase), so we indicate a concept as a pair of phrases like $(\vec{e}, \vec{f})$.

In this model, a document pair can be linearized with various degrees of parallelness from completely independent (when every c_i is monolingual) to completely parallel (when every c_i is bilingual).

[1] http://www.nhk.or.jp/english/

3.2 Training Procedure

Our training procedure consists of the following steps similar to those of the Marcu-Wong method:

1. Initialize distributions.
2. For each document pair, produce an initial alignment by linking phrases so as to create bilingual or monolingual concepts that have high t for all words in the document pairs. Then hillclimb towards the Viterbi alignment by breaking and merging concepts, swapping words between concepts, and moving words across concepts, so as to maximize the product of t.
3. Update distributions with the results of hillclimbing in step 2.
4. Iterate step 2.–3. several times.

We use a suffix array data structure for counting phrase occurrences (Callison-Burch et al., 2005), so we don't need to select only the limited number of high-frequency n-grams as phrase candidates.

In the following sections we give a detailed explanation of our extensions to the steps of the Marcu-Wong method.

3.2.1 Initializing Distributions

***t*-distribution** We define a phrase as a continuous sequence of zero or more words which does not extend more than one sentence. Under this definition, a document consisting of w words and s non-empty sentences can be partitioned into i non-empty phrases in $\binom{w-s}{i-s}$ ways, because the document has $w \quad s$ partitionable word boundaries and $i \quad s$ times of partitioning makes s pieces into i fragments[2]. Given that any phrases in $\boldsymbol{e}$ consisting of w_e words and s_e non-empty sentences can be mapped to any phrase in $\boldsymbol{f}$ consisting of w_f words and s_f non-empty sentences, there are $A(w_e, s_e, w_f, s_f)$ ways of alignments that can be built between $(\boldsymbol{e}, \boldsymbol{f})$:

$$A(w_e, s_e, w_f, s_f) = \sum_{k=0}^{\min(w_e, w_f)} k! \sum_{i=\max(k, s_e)}^{w_e} \sum_{j=\max(k, s_f)}^{w_f} \binom{w_e \quad s_e}{i \quad s_e} \binom{i}{k} \binom{w_f \quad s_f}{j \quad s_f} \binom{j}{k}. \quad (2)$$

In this formula, k denotes the number of bilingual concepts that $(\boldsymbol{e}, \boldsymbol{f})$ shares, and i and j denote the number of phrases which $\boldsymbol{e}$ and $\boldsymbol{f}$ are partitioned into, which follows that $\boldsymbol{e}$ and $\boldsymbol{f}$ have $i \quad k$ and $j \quad k$ phrases generated from monolingual concepts, respectively.

When the EM training starts without any information, all of the $A(w_e, s_e, w_f, s_f)$ alignments that can be built between the document pair $(\boldsymbol{e}, \boldsymbol{f})$ can be assumed to occur with the same probability. Under these conditions, the probability that a bilingual concept $(\vec{e}, \vec{f})$ occurs to generate non-empty phrases $\vec{e}$ and $\vec{f}$ consisting of l_e and l_f words in the document pair $(\boldsymbol{e}, \boldsymbol{f})$ is

$$\frac{A(w_e \quad l_e, s_e + \delta_e, w_f \quad l_f, s_f + \delta_f)}{A(w_e, s_e, w_f, s_f)}. \quad (3)$$

If $\vec{e}$ is placed in the middle of a sentence so that its removal separates the sentence into two non-empty parts, then $\delta_e = 1$; if $\vec{e}$ shares a single end with a sentence so that its removal from the sentence leaves a single non-empty sequence, then $\delta_e = 0$; and if $\vec{e}$ covers the whole of a sentence, then $\delta_e = \quad 1$ (δ_f likewise).

Similarly, the probability that a monolingual concept $(\vec{e}, \quad)$ occurs to generate a non-empty phrase $\vec{e}$ consisting of l_e words in the document pair $(\boldsymbol{e}, \boldsymbol{f})$ is:

$$\frac{A(w_e \quad l_e, s_e + \delta_e, w_f, s_f)}{A(w_e, s_e, w_f, s_f)} \quad (4)$$

(and likewise for concept $(\quad, \vec{f})$).

We can consider the probabilities (3) and (4) for each concept as the expected counts for which the concept contributes to the generation of the document pairs. We collect these counts for each document pair in a corpus,

[2] Although it is not theoretically essential to do so, we strictly enumerate the ways of partitioning, unlike in the Marcu-Wong method which approximates them by using the Stirling number.

and then obtain an initial joint distribution t by normalizing the counts to obtain probabilities. The use of a suffix array data structure for counting phrases enables us to calculate each t probability on the fly while EM training without a prepared table. The only thing we have to calculate beforehand is the total counts as a normalization factor.

***o*-distribution** In addition to the t-distribution, we need a distribution of phrase cooccurrence counts o, for checking the correlation between the bilingual phrase pairs described in the next section.

We consider a pair of bilingual phrases $\vec{e}$ and $\vec{f}$ in a document pair $(\boldsymbol{e}, \boldsymbol{f})$ to be cooccurring phrases if they are potentially generable by a bilingual concept; i.e. the pair is generated by a bilingual concept, or each of the pair is separately generated by a monolingual concept. In addition, we assume that only smaller number of cooccurrences between a and b are observed when $\vec{e}$ (we call each of them $\vec{e_1}, \ldots, \vec{e_a}$) in $\boldsymbol{e}$ appears a times and $\vec{f}$ (we call each of them $\vec{f_1}, \ldots, \vec{f_b}$) in $\boldsymbol{f}$ appears b times. There are $(a + \sum_{n=1}^{b-1} \sum_{c=1}^{a-n} c)$ ways of alignments between $(\boldsymbol{e}, \boldsymbol{f})$ where the same number of $\vec{e}$ and $\vec{f}$ are generated from monolingual concepts in each side of the document pair (assuming $a > b$), so the cooccurrence counts for a pair $(\vec{e}, \vec{f})$ cooccurring in $(\boldsymbol{e}, \boldsymbol{f})$ can be calculated as follows:

$$\left(1 + \frac{a + \sum_{n=1}^{b-1} \sum_{c=1}^{a-n} c}{ab}\right) \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{A(w_e \quad l_e, s_e + \delta_{e_i}, w_f \quad l_f, s_f + \delta_{f_j})}{A(w_e, s_e, w_f, s_f)}. \tag{5}$$

We collect the counts of each document pair in a corpus to obtain the initial cooccurrence distribution o. As in the calculation of the t-distribution, we only need to prepare the total counts before EM training.

3.2.2 Producing Alignments with Log-Likelihood Ratio (LLR) Checking

To produce the alignments in step 2, we statistically check the bilingual concepts by using log-likelihood ratio (LLR) statistics (Dunning, 1993) so as to produce only concepts of reliably correlated phrase pairs (Moore, 2004; Munteanu and Marcu, 2006). Note that monolingual concepts are all available without checking. The checking procedure for a concept $(\vec{e}, \vec{f})$ is as follows:

1. Prepare the o of the following pairs: $o(\vec{e}, \vec{f})$, $o(\vec{e}, \neg\vec{f})$ (total counts for $\vec{e}$ and any phrases except $\vec{f}$), $o(\neg\vec{e}, \vec{f})$ and $o(\neg\vec{e}, \neg\vec{f})$. Then calculate the $LLR(\vec{e}, \vec{f})$ by using the formula in Figure 2.

2. If the $LLR(\vec{e}, \vec{f})$ exceeds the threshold, the occurrences of $\vec{e}$ and $\vec{f}$ are considered to be reliably correlated. The correlation can be classified as positive if both $ad \quad bc > 0$ in the matrix in Figure 2 and $t(\vec{e}, \vec{f}) > t(\vec{e}, \quad) \cdot t(\quad, \vec{f})$, negative if $ad \quad bc < 0$, and else unreliably correlated.

3. If the LLR value is smaller than the threshold, we cannot make a reliable decision as to whether the occurrences of $\vec{e}$ and $\vec{f}$ are correlated or not.

We produce bilingual concepts only from phrase pairs that are considered to have positive correlation.

$$LLR(\vec{e}, \vec{f}) = 2 \log \frac{B(a|a+b, \frac{a}{a+b}) B(c|c+d, \frac{c}{c+d})}{B(a|a+b, \frac{a+c}{a+b+c+d}) B(c|c+d, \frac{a+c}{a+b+c+d})}$$

$$B(k|n,p) = \binom{n}{k} p^k (1 \quad p)^{n-k}$$

	$\vec{e}$	$\neg\vec{e}$
$\vec{f}$	a	b
$\neg\vec{f}$	c	d

cooccurrence count matrix

Figure 2: Log-Likelihood Ratio Statistics (Dunning, 1993)

3.2.3 Updating Distributions

We update the t- and o-distributions in the same way as the Marcu-Wong method; we calculate the probabilities for each alignment generated during the hillclimbing process over all document pairs in a corpus, and then collect counts over all concepts and coocurrences

in these alignments. The detailed procedure differs from the original as follows because of LLR checking.

***t*-distribution** In the generated alignments, unreliably correlated bilingual concepts are never found because they are suppressed producing by LLR checking. Word sequences that can be generated by such unreliably correlated bilingual concepts are mostly composed of monolingual concepts. Therefore we use the following procedure for updating the t-distribution:

1. For each document pair, collect counts for each concept for all alignments.

2. Distribute the counts for every monolingual concept in the result of step 1 to the every monolingual and unreliably correlated bilingual concepts in proportion to the current t-distribution to obtain smoothed counts for a document pair.

3. Collect these smoothed counts for all document pairs in a corpus.

4. Obtain the updated t-distribution for the next iteration by normalizing the counts.

In our implementation of the suffix array data structure, the difference from the initial distribution is stored in the table for each document pair. Every count for positive and negative correlated bilingual concepts is stored in the table since they cannot be directly calculated from the initial distribution. On the otherhand, the counts for the rest can be obtained by multiplying their initial counts by a factor for each document pair, which is also held in the table.

***o*-distribution** From the definition of phrase cooccurrences described in Section 3.2.1, we approximate the updated cooccurrence counts of $(\vec{e}, \vec{f})$ in $(\boldsymbol{e}, \boldsymbol{f})$ by the following equation ($a, b, \vec{e_i}, \vec{f_j}$ are the same as in Section 3.2.1):

$$\sum_{i=1}^{a}\sum_{j=1}^{b} t(\vec{e_i}, \vec{f_j}|(\boldsymbol{e}, \boldsymbol{f})) + \frac{a + \sum_{n=1}^{b-1}\sum_{c=1}^{a-n} c}{ab} \sum_{i=1}^{a}\sum_{j=1}^{b} t(\vec{e_i}, \ |(\boldsymbol{e}, \boldsymbol{f}))\, t(\ , \vec{f_j}|(\boldsymbol{e}, \boldsymbol{f}))\,. \quad (6)$$

We can easily calculate these conditional probabilities from the difference table for t-distribution if the table also hold the total alignment probability of the document pairs.

4 Experiments

We conducted a series of preliminary experiments using our model to align phrases from the NHK Japanese-English broadcast news corpus, which is composed of document pairs of Japanese news scripts and their manual translation into English. The Japanese documents in the corpus were segmented into morpheme tokens with part-of-speech tags by Chasen[3], the morphological analyzer for Japanese. Each experiment was given different conditions as to the size of corpora, LLR thresholds, and the times of iterations as in Table 1. Note that the smaller corpus is the subset of the larger one.

One human evaluator evaluated the quality of the phrase alignments by marking all alignments from the 10 randomly selected article pairs in each of the above experiments. He marked according to three grades:

correct(○): the extracted phrase pair is parallel without no extra or absent words,

partly correct(△): the extracted phrase pair has extra or absent word(s) but almost all content words are parallel,

incorrect( ): otherwise.

Table 2 shows the number of alignments for each grade, the average number of words in the aligned phrases, and coverage (how many words of each document were covered by the aligned phrases).

Table 3 shows some phrase alignments that have higher LLR scores in the article pair

[3] `http://chasen.naist.jp/hiki/ChaSen/`

No.	Corpus Size # of document pairs (# of tokens / types)	LLR Threshold[4]	Iteration Times
1	1,000 (J: 287,597 / 10,855) (E: 161,976 / 10,521)	3.841 (95%)	1
2			3
3			5
4		2.706 (80%)	
5		0.4549 (50%)	
6	2,000 (J: 578,374 / 18,182) (E: 312,353 / 17,905)	3.841 (95%)	3

Table 1: Experimental Conditions

Condition No.		1	2	3	4	5	6
Evaluation	○	32/7	65/19	102/32	188/59	164/44	173/46
(# of alignments	△	8/4	28/15	35/21	61/46	66/42	53/33
(tokens/types))		42/22	33/19	26/20	216/166	357/258	38/25
rate of ○ or △ (token/type)		.488/.371	.738/.642	.840/.726	.535/.389	.392/.250	.856/.760
Phrase Length	J	1.02	1.09	1.21	1.29	1.23	1.33
(# of words)	E	1.01	1.10	1.19	1.18	1.10	1.24
Coverage	J	.029	.049	.071	.210	.254	.122
(rate in words)	E	.051	.088	.124	.341	.403	.211

Table 2: Results of evaluation

shown in Figure 1 from the experiment for the condition 6.

Comparing the evaluations of the experimental conditions 3 to 5, it is apparent that LLR checking seems to be useful for selecting parallel segments from comparable corpora.

Comparing the conditions 1 to 3, we see that the iteration improves the quality of alignments, but is not very effective for finding new longer alignments as expected. This may be because our method of updating distributions is inappropriate.

Comparing the conditions 3 and 6, we see that a larger corpus size made coverage better and phrase lengths longer but did not change the precision by much. This means that LLR checking guarantees the correctness of phrasal alignments according to the LLR thresholds.

5 Related Work

The studies on acquiring translation knowledge from non-parallel corpora started with extracting lexical translations (e.g. (Fung and Yee, 1998; Rapp, 1999)). To find translations, they generally exploit the tendency that equivalent words have similar contextual words in corpora of different languages. These methods are powerful in terms of their applicability even to unrelated bilingual corpora, but they provide very poor coverage.

Extracting parallel segments of longer than lexical level from non-parallel corpora have been studied afterward. As for the challenges to exploit comparable corpora, there have been some efforts on extracting parallel sentences (Zhao and Vogel, 2002; Munteanu and Marcu, 2002). Both studied used a statistical bilingual dictionary obtained from a parallel corpus as bootstraps to extract more parallel

[4]The asymptotic distribution of LLR statistics will follows $^2(1)$, so if the LLR score of a phrase pair exceeds a threshold whose $^2(1)$ probability is p, the phrase pair is considered to be correlated with an approximate probability of p.

Japanese	English	Log Prob.	LLR	Judge
地震	quake	14.2	12.8	○
地震	earthquakes	15.3	10.1	○
気象庁	The Meteorological Agency	15.9	8.11	○
以上/の	more	15.1	7.89	○
地震	jolted	14.6	7.83	
強い	strong	14.9	4.17	○
を/観測/し (*observe(d)*)	eaethquake	16.8	4.11	

Table 3: Example of phrase alignments extracted in the experiment No.6

sentences and bilingual lexicons from comparable corpora. Fung and Cheung (2004) used a multi-level bootstrapping to improve alignments at the levels of document, sentence, and word pairs and thereby avoid the use of preexisting knowledge sources such as dictionaries.

These methods of parallel sentence extraction have a limitation in that few sentence pairs can be extracted from corpora that are far from parallel. Munteanu and Marcu (2002) proposed a method of extracting sub-sentential parallel fragments from comparable corpora. It first selects sentence pairs which are likely to share some parallel fragments from a bilingual dictionary of broad coverage, then detects parallel fragments within each of the sentence pairs by another precise bilingual dictionary.

These studies aim to *mine* corpora for clean parallel parts in order to acquire further knowledge for proposes such as SMT. On the other hand, our approach directly acquires phrase alignments from comparable document pairs. We obtain lexical translation knowledge and extract parallel parts from comparable corpora simultaneously.

6 Conclusion

We described a method of extracting phrasal alignments from comparable corpora by using an extended phrase-based joint probability model for statistical machine translation. Our method can extract phrasal alignments directly from comparable document pairs composed of about 5–8 sentences without preexisting resources or splitting them into sentences. The experiments showed that our method achieves about 0.8 in precision of phrasal alignment extraction when using 2,000 document pairs of Japanese-English news articles as training data, thanks to its use of the alignment checking process using log-likelihood ratio statistics.

The experiments indicated plenty of room for our method to be improved, e.g.:

> As mentioned before, our method of updating distributions is far from theoretically well-grounded, which may affect performance.
>
> Computation cost is high, especially for the hillclimbing search. We need to make practical improvements to the process (e.g. (Birch et al., 2006)). Calculating distributions on the fly also costs very much, which spoil the merit of the suffix array data structure in part.
>
> Our method cannot recognize discontinuous segments as phrases. It is common that a continuous phrase in English does not have a Japanese counterpart of discontinuous segments because of the difference in language structure. We would like to improve the model so that it can handle discontinuous phrasal segments.
>
> Our method highly depends on the size of each document in a training corpus. Because we find statistical prominence in the cooccurrences distribution to find reliable phrase correspondences, expan-

sion of each cooccurrence window will decrease the performance of our method. We need to test our method for longer documents.

We would like to make a much finer evaluation by manually constructing an evaluation set in the near future. The proposed model highly depends The proposed model is an enhancement of Marcu-Wong's Model 1 and it does not contain a constraint on word or phrase order. We would like to enhance our method by taking order into consideration, and apply it to statistical machine translation.

Acknowledgements

We would like to express our deep gratitude to Mr. Tomoki Kozuru from Kanji Information System Co.,Ltd., who worked together with us in implementing the experimental system.

References

Alexandra Birch, Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Constraining the phrase-based, joint probability statistical translation model. In *Proceedings on the HLT-NAACL 2006 Workshop on Statistical Machine Translation*, pages 154–157.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 255–262.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Pascale Fung and Percy Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 1051–1057.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pages 414–420.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 133–139.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 333–340.

Dragos Stefan Munteanu and Daniel Marcu. 2002. Processing comparable corpora with bilingual suffix trees. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 289–295.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 81–88.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '98)*, pages 519–526.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the IEEE International Conference on Data Mining (ICDM 2002)*, pages 745–748.