# Phrase Alignment Based on Bilingual Parsing

**Akira Ushioda**

Software and Solution Laboratories
Fujitsu Laboratories Ltd.
4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki 211-8588
Japan
ushioda@jp.fujitsu.com

## Abstract

A novel approach is presented for extracting syntactically motivated phrase alignments. In this method we can incorporate conventional resources such as dictionaries and grammar rules into a statistical optimization framework for phrase alignment. The method extracts bilingual phrases by incrementally merging adjacent words or phrases on both source and target language side in accordance with a global statistical metric. The extracted phrases achieve a maximum F-measure of over 80 with respect to human judged phrase alignments. The extracted phrases used as training corpus for a phrase-based SMT shows better cross-domain portability over conventional SMT framework.

## 1 Introduction

In the phrase-based SMT framework (Marcu & Wong, 2002; Och & Ney, 2004; Chiang, 2005), extraction of phrase pairs is a key issue. Currently the standard method of extracting bilingual phrases is to use a heuristics such as diag-and (Koehn et. al., 2003). In this method starting with the intersection of word alignments of both translation directions additional alignment points are added according to a number of heuristics and all the phrase pairs which are consistent with the word alignments are collected.

Although this method is effective by itself it is very difficult to incorporate syntactic information in a straight manner because phrases extracted by this method have basically little syntactic significance. Especially if we intend to combine strength of conventional rule-based approach with that of SMT, it is essential that phrases, or translation units, carry syntactic significance such as being a constituent (Yamada & Knight, 2001).

Another drawback of the conventional method is that the phrase extraction process is deterministic and no quantitative evaluation is applied. Furthermore if the initial word alignments have errors, these errors propagate to the phrase alignment process. In doing so the burden of statistical optimization is imposed on the final decoding process. We propose in this paper a novel phrase alignment method in which we can incorporate conventional resources such as dictionaries and grammar rules into a statistical optimization framework for phrase alignment.

The outline of the proposed method, applied to Japanese-English bilingual corpus, is as follows.

1) The training bilingual corpus is first word-aligned by GIZA++ (Och & Ney, 2000).
2) A word translation model is learnt by relative frequency from the word-alignment and smoothed by a bilingual dictionary.
3) Chunking is performed on both sides.
4) The probability that an English word belongs to a Japanese chunk is evaluated from which an entropy score is computed.
5) The entropy score is used to guide the process of merging adjacent phrases of both languages.
6) The merging process terminates when the score takes a minimum value.

Although the above steps are purely guided by a statistical metric, some syntactic preferences or constraints can guide the search.

The objective of this work is to extract alignments of phrases which are linguistically motivated. However, there is no guarantee that even manually extracting, out of aligned sentences, bilingual phrases which correspond to each other in meaning results in a collection of pairs of source and target phrases which are both constituents. There might be cases in which a phrase in one language constitutes a constituent while the corresponding phrase in the other language does not. Therefore the basic strategy we adopt here is to try to extract bilingual phrases whose source language side at least constitutes a constituent. As for the target language side, a preference is given to constituent constructs.

## 2 Phrase Alignment Method

The phrase alignment method we propose here extracts bilingual phrases by incrementally merging adjacent words or phrases on both source and target languages in accordance with a global statistical metric along with syntactic constraints and preferences.

The merging process is guided by an entropy score which is calculated from the *alignment matrix*. Figure 1 shows an example of the alignment matrix for the following sentence pair:

(1a) 演算回路の記憶値の乗算と新しいデータの加算のループを繰り返すことにより，簡単な演算回路で現在のデータに重みを置いた平均値を算出可能とする。

(1b) *To calculate an average value weighed in the present data with a simple arithmetic circuit by repeating the loop of multiplication of the stored value in the arithmetic circuit and the addition of a new data.*

In the alignment matrix, English words are arranged in each row and Japanese chunks are arranged in each column. The value of the (i, j) element divided by the margin of the i-th row represents the probability that the translation of the i-th English word ($w_i$) appears in the j-th Japanese chunk ($j_j$). For example, the translation of $w_1$ (calculate) can be "**演算**", which appears in $j_0$ ("**演算**回路の記憶値の" ) and $j_8$ ("簡単な**演算**回路で"), or "**算出**", which appears in $j_{13}$ ("**算出**可能とする"), or "**算**" , which appears in $j_1$ and $j_3$ in addition to $j_0$, $j_8$ and $j_{13}$. Since "calculate" is more likely to be translated as "**算出**" than others, the (1, 13) element has larger value than other elements in the same row. Determiners, prepositions, conjunctions, and other function words are treated as stopwords and their elements are all assigned a value of zero. When there is more than one element with a positive value in the same row, these elements are shown in Figure 1 with a shaded square, and this means that the corresponding English word is ambiguous on the identity of the corresponding Japanese chunk. On the other hand, if there is only one element, say (p,q), with positive value in the same row, it is certain that the English word $w_p$ belongs to the Japanese chunk $j_q$. If there is one and only one nonzero element in each row and in each column, then we have a complete one-to-one matching between Japanese elements (phrases) and English elements (words or phrases). The intuition behind the proposed method is that by merging adjacent elements which constitute a phrase and tend to stay together in both languages, the alignment matrix approaches a one-to-one matching. Therefore if there is a global measure that shows how close the current alignment matrix is to a one-to-one matching, we can use it to guide the merging process. We use the entropy score which is described in the next section.

| [0] | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] | [9] | [10] | [11] | [12] | [13] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 To |
| 9 | 0.11 | 0 | 0.11 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 64 | 1 calculate |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 an |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 85 | 0 | 3 average |
| 96 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 96 | 0 | 4 value |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 weighed |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 in |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 the |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 34 | 0 | 0 | 0 | 0.04 | 8 present |
| 0 | 0 | 98 | 0 | 0 | 0 | 0 | 0 | 0 | 98 | 0 | 0 | 0 | 0 | 9 data |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 with |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 a |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 56 | 0 | 0 | 0 | 0 | 0 | 12 simple |
| 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 13 arithmetic |
| 99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 99 | 0 | 0 | 0 | 0 | 0 | 14 circuit |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 by |
| 0 | 0 | 0 | 0 | 0 | 53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 repeating |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 the |
| 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 loop |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 of |
| 0 | 72 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 multiplication |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21 of |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 the |
| 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 stored |
| 96 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 96 | 0 | 24 value |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 in |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 26 the |
| 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 27 arithmetic |
| 99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 99 | 0 | 0 | 0 | 0 | 0 | 28 circuit |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 29 and |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 30 the |
| 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 31 addition |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 32 of |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 a |
| 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 34 new |
| 0 | 0 | 98 | 0 | 0 | 0 | 0 | 0 | 0 | 98 | 0 | 0 | 0 | 0 | 35 data |

**[0]:演算回路の記憶値の**
**[1]:乗算と**
**[2]:新しいデータの**
**[3]:加算の**
**[4]:ループを**
**[5]:繰り返す**
**[6]:ことにより**
**[7]:,**
**[8]:簡単な演算回路で**
**[9]:現在のデータに**
**[10]:重みを**
**[11]:置いた**
**[12]:平均値を**
**[13]:算出可能とする**

Figure 1: An example of the alignment matrix

## 2.1 Without Syntactic Information

We begin by describing the proposed phrase alignment method in the case of incorporating no syntactic information. Figure 2 shows the framework of the phrase aligner. In the case of incorporating no syntactic information, *Syntactic Component* in the figure plays no role. We take here an example of translating from Japanese to English, but the framework presented here basically works for any language pair as long as a conventional rule-based approach is applicable.

As a preparation step, word alignments are obtained from a bilingual corpus by GIZA++ for both directions (source to target and target to source), and the intersection $A = A1 \cap A2$ of the two sets of alignments are taken. Then for each English word *e* and Japanese word *j*, the frequency *N(e)* of *e* in *A* and the co-occurrence frequency

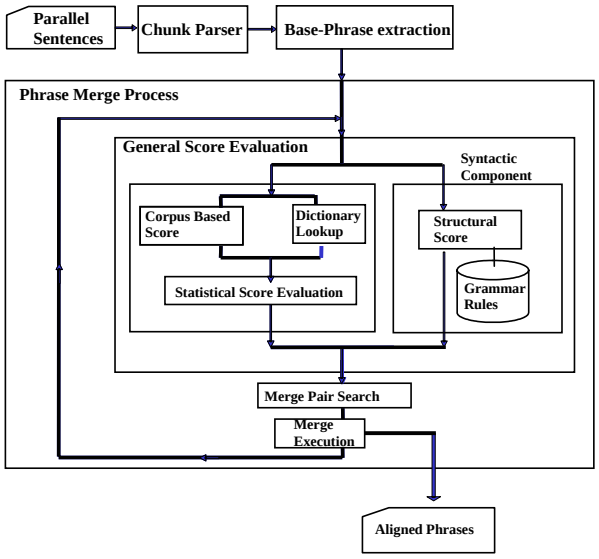



Figure 2: Framework of Phrase Aligner

*N(e , j)* of *e* and *j* in *A* are calculated. Furthermore, using a discrimination function $\delta(e, j)$ which determines whether *e* and *j* are a translation of each other with respect to a predefined bilingual dictionary, word based empirical translation probability is obtained as follows.

(2) $Pc(j|e) = (N(e,j) + \delta(e,j))/(N(e) + \Sigma_t \delta(e,t))$

$\delta(e, j)$ takes a value of 1 when *(e, j)* appears in the bilingual dictionary, and 0 otherwise.

An input to the phrase aligner is a pair **(J, E)** of Japanese and English sentences. The pair (**J**, **E**) is first chunk-parsed to extract base phrases, such as minimum noun phrases and phrasal verbs on both sides.

Let $\boldsymbol{J} = \boldsymbol{j}_1, \boldsymbol{j}_2, \ldots, \boldsymbol{j}_M$ be a series of Japanese chunks. These chunks are the minimum units for composing a final phrase alignment on Japanese side. Let $\mathbf{E} = w_1, w_2, \ldots, w_N$ be a series of English words. Then the probability that the translation of word $w_i$ appears in chunk $\boldsymbol{j}_j$ in the given sentence pair is given by (3)[1].

(3) $P(\boldsymbol{j}_j | w_i) = C_{ij} / \Sigma_j C_{ij}$

, where

(4) $C_{ij} = \Sigma_t Pc(t|w_i)\, P(t \text{ appears in } \boldsymbol{j}_j)$

is what we will call *an alignment matrix* which represents the relative likelihood that the translation of word $w_i$ appears in chunk $\boldsymbol{j}_j$ in comparison with other Japanese chunks, *t* is a translation candidate of $w_i$, and *P(t appears in* $\boldsymbol{j}_j$*)* is zero if $\boldsymbol{j}_j$ doesn't contain *t* as a substring and one if it does. Note that the values of $C_{ij}$ can be calculated form the parallel sentence pair and the empirical translation probability (2).

Similarly for Japanese phrases, we can calculate the probability $P(w_i | \boldsymbol{j}_j)$ that the translation of $\boldsymbol{j}_j$ is represented as $w_i$ as follows.

(5) $P(w_i \mid \boldsymbol{j}_j) = C_{ij} / \Sigma_i C_{ij}$

Given the translation probability (3), we can define the entropy $H(i)$ of the probability distribution $P(\cdot \mid w_i)$ as follows.

(6) $H(i) = -\Sigma_j P(\boldsymbol{j}_j \mid w_i) \log_2 P(\boldsymbol{j}_j \mid w_i)$

Since $\lim_{X \to 0} X \log_2 X = 0$, we define $H(i) = 0$ when $P(\boldsymbol{j}_j \mid w_i) = 0$ for all j.

In the proposed method, a statistical metric based on the entropy (6) is used for judging which adjacent phrases are to be merged. We calculate the change in the evaluation metric resulting from the merge just in the same way as we calculate the information gain (the reduction of entropy) of a decision tree when the dataset is divided according to some attribute, with the only difference that in a decision tree a dataset is incrementally *divided*, whereas in our method rows and columns are *merged*. We treat each row and each column of the alignment matrix as a dataset. The entire entropy, or uncertainty, of mapping English phrases to Japanese phrases is then given by:

(7) $H = \Sigma_i [\Sigma_j C_{ij}] H(i) / \Sigma_i \Sigma_j C_{ij}$

The entropy of mapping Japanese phrases to English phrases is obtained in the same way.

(8) $H_t = \Sigma_j [\Sigma_i C_{ij}] H(j) / \Sigma_i \Sigma_j C_{ij}$

Finally we define the total statistical metric, or an *evaluation score*, as the mean value of the two.

$H_{tot} = (H + H_t)/2$

**Phrase Extraction**

The merging process is terminated when the evaluation score $H_{tot}$ takes a minimum value. When the final value of the alignment matrix is obtained, then for each non-zero

[1] Interested readers are referred to (Ushioda, 2007) for more details of the derivation of equation (3).

element $C_{ij}$ the corresponding English phrase in the i-th row and the Japanese phrase in the j-th column are extracted and paired as an aligned phrase pair. Even after $H_{tot}$ reaches zero we can continue merging as long as $H_{tot}$ stays zero and a different set of phrase pairs can be extracted at each merging step while $H_{tot}$ stays zero. Whether rows are merged or columns are merged at each merging step is determined by the evaluation score. Since the merging process is easily trapped by the local minimum with a greedy search, a beam search is employed while keeping multiple candidates (instances of alignment matrices). The typical beam size employed is between 300 and 1000.

One of the advantages of the proposed method is that we can directly incorporate dictionary information into the scheme, which is quite effective for alleviating data sparseness problem especially in the case of small training corpus. Another distinctive feature of the method is that once word alignments are obtained and the empirical translation probability *Pc(j|e)* is calculated together with the dictionary information, the word alignments are discarded. This is how this method avoids deterministic phrase alignment, and keeps a possibility of recovering from word alignment errors.

**Multiple Correspondences**

As we saw in the example of Figure 1 there is very often more than one element with a positive value in the same row of the alignment matrix. Usually only one nonzero element is correct and others are erroneously assigned nonzero values due to an accidental string match between the Japanese chunks and the translation of the English word. However there is no simple way of preliminarily disambiguating the identity of the corresponding Japanese chunk.

To cope with this initial ambiguity, a separate initial alignment matrix is constructed for each combination of a nonzero element of a row so that each row has at most one nonzero element. If there are *n* words $w_1, w_2, \cdots w_n$ in the English sentence, and each word $w_i$ has $k_i$ possible corresponding Japanese chunks, then the number of combinations is $k_1 k_2 \cdots k_n$, which sometimes becomes huge. However, in the process of merging, most of the erroneous word alignments disappear in confrontation with correct word alignments. Figure 3 shows two examples of an initial alignment matrix candidate for the sentence pair (1) and phrase alignments obtained after the merging process. Since the evaluation score of (c) is zero, (a) is considered to be the correct initial alignment matrix. As a result, the initial ambiguity on the identity of the corresponding Japanese chunk for each English word is resolved.

In some cases, however, multiple correspondences between English words and Japanese chunks are intrinsic. Consider the following sentence pair.

(11a) 真空賦勢した管及び血液の取り出し中に添加剤を分配するための方法を提供する。

(11b) *To provide a tube energized in vacuum and establish a method for distributing additives during the process of taking out the blood.*

Figure 4 shows the phrase alignment result for this pair and Figure 5 shows the initial and final alignment matrices. As Figure 4 shows the Japanese verb "提供する" (f) is aligned with both "To provide" (t) and "and establish" (v). This is because in the clausal conjunction different verbs are used for different objects (a tube and a method) in English whereas the same verb (f) is used in Japanese. In those cases one-to-one correspondence can never be achieved through merging, but still the evaluation score is expected to lead the merging process to a correct alignment result.

## 2.2 With Syntactic Information

The proposed framework also has a capability of incorporating syntactic constraints and preferences in the process of merging. For example, suppose that there are two competing merging candidates; one is to merge (i-th row, i+1-th row) and the other is to merge (k-th column, k+1-th column), and that their evaluation scores are H1 and H2 respectively. Then if there are no syntactic constraints or preferences, the merging candidate which has the lower evaluation score is elected. But if there are syntactic constraints, the only merging candidate which satisfies the constraints is executed. When a syntactic preference is introduced, then the evaluation score is multiplied by some value which represents the degree of the strength of the preference. If we intend to extract only pairs of phrases which constitute a constituent, then we introduce a constraint which eliminates merging candidates that produce a phrase which crosses a constituent boundary. Although our goal is to fully integrate complete set of CFG rules into the merging scheme, we are still in the process of constructing the syntactic rules, and in the present work we employed only a small set of preferences and constraints. Table 1 illustrates some of the syntactic constraints and preferences employed in the present work.

Merging lines or columns in the alignment matrix can be viewed as a form of bottom-up parsing. When we trace the process of the merging, its history can be converted to

| [0] | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] | [9] | [10] | [11] | [12] | [13] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 To |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 64 | 1 calculate |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 an |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 85 | 0 | 3 average |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 96 | 0 | 4 value |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 weighed |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 in |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.4 | 0 | 0 | 0 | 0 | 7 the |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 34 | 0 | 0 | 0 | 0 | 8 present |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 98 | 0 | 0 | 0 | 0 | 9 data |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 with |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 a |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 56 | 0 | 0 | 0 | 0 | 0 | 12 simple |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 13 arithmetic |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 99 | 0 | 0 | 0 | 0 | 0 | 14 circuit |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 by |
| 0 | 0 | 0 | 0 | 0 | 53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 repeating |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 the |
| 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 loop |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 of |
| 0 | 72 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 multiplicatio |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21 of |
| 0.8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 the |
| 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 stored |
| 96 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 24 value |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 in |
| 0.8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 26 the |
| 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 27 arithmetic |
| 99 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 28 circuit |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 29 and |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 30 the |
| 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 31 addition |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 32 of |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 a |
| 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 34 new |
| 0 | 0 | 98 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 35 data |

(a)

| [0] | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] | [9] | [10] | [11] | [12] | [13] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 To |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 1 calculate |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 an |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 85 | 0 | 3 average |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 96 | 0 | 4 value |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 weighed |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 in |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.4 | 0 | 0 | 0 | 0 | 7 the |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 34 | 0 | 0 | 0 | 0 | 8 present |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 98 | 0 | 0 | 0 | 0 | 9 data |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 with |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 a |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 56 | 0 | 0 | 0 | 0 | 0 | 12 simple |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 13 arithmetic |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 99 | 0 | 0 | 0 | 0 | 0 | 14 circuit |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 by |
| 0 | 0 | 0 | 0 | 0 | 53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 repeating |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 the |
| 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 loop |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 of |
| 0 | 72 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 multiplication |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 21 of |
| 0.8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 the |
| 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 stored |
| 96 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 24 value |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 in |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 26 the |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | 27 arithmetic |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 99 | 0 | 0 | 0 | 0 | 0 | 28 circuit |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 29 and |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 30 the |
| 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 31 addition |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 32 of |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 a |
| 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 34 new |
| 0 | 0 | 98 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 35 data |

(b)

| [0] | [1] | [2] | [3] | [4] | [5] | [6] | [7] | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 64 | : | To calculate |
| 0 | 0 | 0 | 0 | 0 | 0 | 182 | 0 | : | an average value |
| 0 | 0 | 0 | 0 | 0 | 133 | 0 | 0 | : | weighed in the present data |
| 0 | 0 | 0 | 0 | 252 | 0 | 0 | 0 | : | with a simple arithmetic circuit |
| 0 | 0 | 0 | 151 | 0 | 0 | 0 | 0 | : | by repeating the loop |
| 0 | 72 | 0 | 0 | 0 | 0 | 0 | 0 | : | of multiplication |
| 321 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | : | of the stored value in the arithmetic circuit |
| 0 | 0 | 131 | 0 | 0 | 0 | 0 | 0 | : | and the addition of a new data |

**[0]: 演算回路の記憶値の**
**[1]: 乗算と**
**[2]: 新しいデータの加算の**
**[3]: ループを繰り返すことにより,**
**[4]: 簡単な演算回路で**
**[5]: 現在のデータに重みを置いた**
**[6]: 平均値を**
**[7]: 算出可能とする**

(c)

| [0] | [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | : | To calculate |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 182 | 0 | : | an average value |
| 0 | 0 | 0 | 0 | 0 | 0 | 133 | 0 | 0 | : | weighed in the present data |
| 0 | 0 | 0 | 0 | 0 | 252 | 0 | 0 | 0 | : | with a simple arithmetic circuit |
| 0 | 0 | 0 | 0 | 53 | 0 | 0 | 0 | 0 | : | by repeating |
| 0 | 0 | 0 | 97 | 0 | 0 | 0 | 0 | 0 | : | the loop |
| 0 | 72 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | : | of multiplication |
| 124 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | : | of the stored value |
| 0 | 0 | 0 | 0 | 0 | 196 | 0 | 0 | 0 | : | in the arithmetic circuit |
| 0 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 | : | and the addition |
| 0 | 0 | 119 | 0 | 0 | 0 | 0 | 0 | 0 | : | of a new data |

**[0]:演算回路の記憶値の**
**[1]:乗算と**
**[2]:新しいデータの**
**[3]:加算のループを**
**[4]:繰り返すことにより,**
**[5]:簡単な演算回路で**
**[6]:現在のデータに重みを置いた**
**[7]:平均値を**
**[8]:算出可能とする**

(d)

Figure 3: Two examples of an initial alignment matrix candidate for the sentence pair (1) and their merging results. (c) and (d) are the results of merging (a) and (b), respectively.

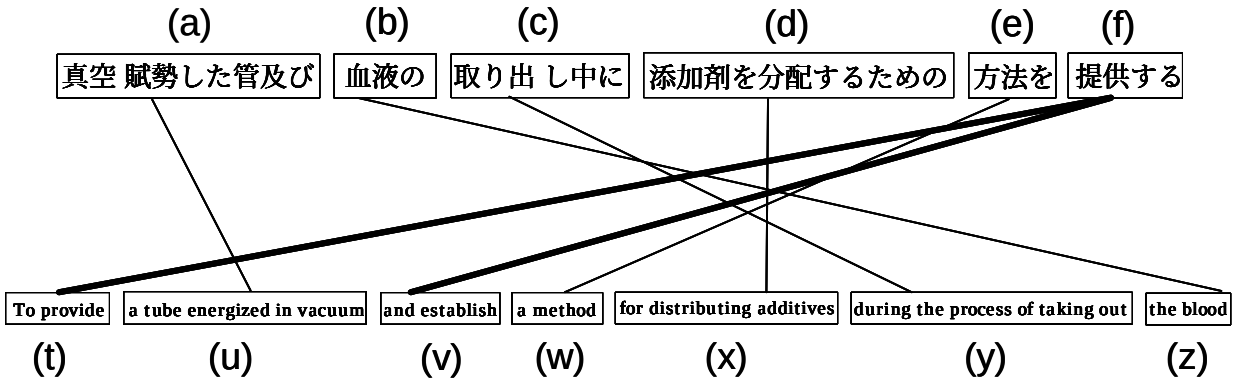



Figure 4: An example of intrinsic multiple correspondences.

| [0] | [1] | [2] | [3] | [4] | [5] | [6] | |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 To |
| 0 | 0 | 0 | 0 | 0 | 0 | 83 | 1 provide |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 a |
| 53 | 0 | 0 | 0 | 0 | 0 | 0 | 3 tube |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 energized |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 in |
| 91 | 0 | 0 | 0 | 0 | 0 | 0 | 6 vacuum |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 and |
| 0 | 0 | 0 | 0 | 0 | 0 | 23 | 8 establish |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 a |
| 0 | 0 | 0 | 0 | 0 | 95 | 0 | 10 method |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 for |
| 0 | 0 | 0 | 0 | 43 | 0 | 0 | 12 distributing |
| 0 | 0 | 0 | 70 | 0 | 0 | 0 | 13 additives |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 during |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 the |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 16 process |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 of |
| 0 | 0 | 68 | 0 | 0 | 0 | 0 | 18 taking |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 out |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 the |
| 0 | 94 | 0 | 0 | 0 | 0 | 0 | 21 blood |

**[0]:真空賦勢した管及び**
**[1]:血液の**
**[2]:取り出し中に**
**[3]:添加剤を**
**[4]:分配するための**
**[5]:方法を**
**[6]:提供する**

(a)

| [0] | [1] | [2] | [3] | [4] | [5] | | |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 83 | : | To provide |
| 144 | 0 | 0 | 0 | 0 | 0 | : | a tube energized in vacuum |
| 0 | 0 | 0 | 0 | 0 | 23 | : | and establish |
| 0 | 0 | 0 | 0 | 95 | 0 | : | a method |
| 0 | 0 | 0 | 113 | 0 | 0 | : | for distributing additives |
| 0 | 0 | 68 | 0 | 0 | 0 | : | during the process of taking out |
| 0 | 94 | 0 | 0 | 0 | 0 | : | the blood |

**[0]:真空賦勢した管及び**
**[1]:血液の**
**[2]:取り出し中に**
**[3]:添加剤を分配するための**
**[4]:方法を**
**[5]:提供する**

(b)

Figure 5: Initial (a) and final (b) alignment matrices for sentence (11)

a binary parse tree on both language sides. Since we are not yet incorporating grammar rules in our phrase alignment system, the merge history-induced inner-structures of the obtained bilingual phrases are not quite linguistically intuitive, although the obtained phrases themselves are intended to be linguistically motivated. However, even within the current setting, the obtained alignment matrix can be useful for guiding parsing process or correcting parse results via interplay between parsers of both sides through the alignment matrix. Figure 6 illustrates an example. If we suppose that the Japanese parse tree is more reliable than the English parse tree, then the alignment matrix can be used to convert Japanese tree structure into English one and to correct the PP-attachment error of the original English parse tree in which "*by forming*" is attached to "*to perform*" instead of the correct attachment site which is the conjunction of the preceding two clauses.

## 3 Experimets

This section describes experiments with the proposed phrase alignment method. For the evaluation of the obtained phrase alignments, two types of experiments are conducted. One is to evaluate the F-measure of the obtained phrase alignments with respect to a hand crafted golden standard. The second type is to measure the quality of phrase-based SMT which uses the obtained phrase pairs as a bilingual corpus. Each experiment is described in the following subsections. We used the test collection of a parallel patent corpus from the Patent Retrieval Task of the 3rd NTCIR Workshop (2002) for training word alignments. The corpus comprises of patent abstracts of Japan (1995-1999) and their English translation produced at Japan Patent Information Organization. We extracted 150 thousand sentence pairs from the PURPOSE part of the test collection of the year 1995. Each patent has its IPC category, from A through H. In-house English and Japanese parsers are used to chunk sentences and to make a constituent judgment. We also used in-house bilingual dictionary with 860 thousand word entries. For phrase alignment, we extracted 13,000 sentence pairs with English sentences of length smaller than 75 words, out of the sentence pairs in G-category (Physics) of the above word alignment training set. The sentence length is constrained to reduce the computational load. Table 2 summarizes the training corpora used. Out of 13,000 sentence pairs 208 thousand unique phrase pairs are extracted. More than one set of phrase alignments can often be extracted from one pair of aligned sentences when the evaluation score reaches zero.

Figure 7 shows examples of obtained phrase alignments. Japanese phrases acquired are mostly constituents, whereas many of English phrases are not, such as " by arranging", or "of infrared absorption ink". This is partly due to the fact that Japanese phrases are constructed out of base phrases, or chunks, whereas English phrases are constructed starting from individual words. Another reason is the fact that Japanese precedence rule takes precedence over English one as stated in Table 1.

### 3.1 Evaluation of Phrases with Human Judgment

Out of the 13,000 sentence pairs used for phrase alignments, 160 sentence pairs are randomly extracted for manual annotation. Although there have been a number of attempts to manually annotate word alignments, much less attempts have been made to construct a golden standard for phrase alignments. The major difficulty of aligning phrases is that there are many possible ways of aligning phrases, whereas word alignments have not much ambiguity.

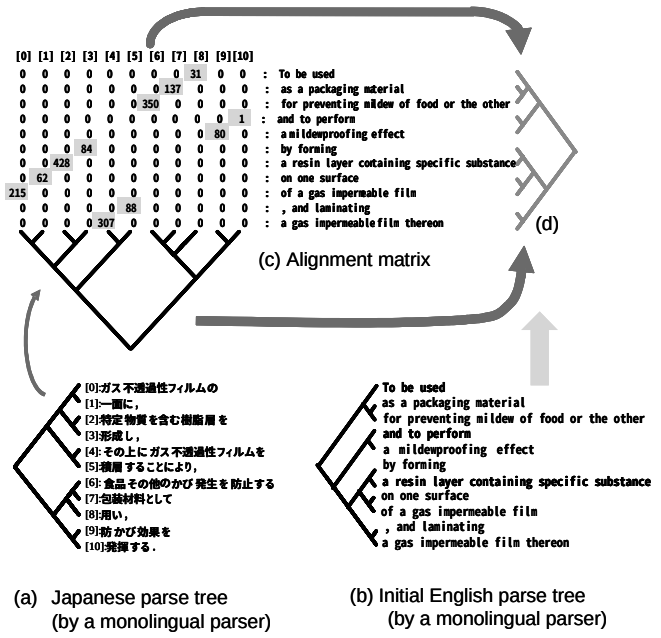



Figure 6: An example of correcting an English parse result by the combination of Japanese parse tree and the alignment matrix. In the initial English parse tree (b), the phrase "*by forming*" can be interpreted to be attached to "*and to perform*". Through the alignment matrix (c), the Japanese parse tree (a) can be automatically mapped to the English parse tree (d) which can for instance derive the correct interpretation of the attachment site of the phrase "*by forming*".

Since there is no obvious criterion to decide which phrase pairs are superior and which are not, we choose to extract all the possible ways of dividing a sentence pair into a set of bilingual phrases. Of course it is too much work for a human to exhaust all the possible combinations. However, there is a way of automatically generating all the possible phrase alignments from a result of manual work which is just repeating a simple task of dividing a phrase pair into pairs of sub-phrases. For example, consider a phrase pair in Figure 8. The phrase pair ("j1j2j3j4j5", "e1e2e3e4e5") is first divided into two phrase pairs, ("j1j2", "e4e5") and ("j3j4j5", "e1e2e3"). There are in total four possible division steps like this:

(12a) ("j1j2j3j4j5", "e1e2e3e4e5") ⇒ ("j1j2", "e4e5") , ("j3j4j5", "e1e2e3")
(12b) ("j1j2", "e4e5") ⇒("j1", "e5"), ("j2", "e4")
(12c) ("j3j4j5","e1e2e3") ⇒("j3","e3"), ("j4j5", "e1e2")
(12d) ("j4j5", "e1e2") ⇒ ("j4", "e2"), ("j5", "e1")

Given these four possible divisions, all the possible phrase alignments can be automatically calculated and the results are as follows.

("j1j2", "e4e5") ,("j3j4j5", "e1e2e3")
("j1j2", "e4e5"), ("j3","e3"), ("j4j5", "e1e2")
("j1j2","e4e5"),("j3","e3"),("j4","e2"), ("j5","e1")
("j1", "e5"), ("j2", "e4"), ("j3j4j5", "e1e2e3")
("j1","e5"),("j2","e4"),("j3","e3"),("j4j5", "e1e2")
("j1","e5"),("j2","e4"),("j3","e3"),("j4","e2"), ("j5","e1")

Therefore the task of human annotator is to keep dividing a phrase pair into pairs of sub-phrases. The procedure of the manual annotation is as follows.

1) Let the aligned sentence pair be a pair of aligned phrases.
2) Pick a pair of aligned phrases and try to divide it into two constituents so that each of the Japanese sub-phrases can be regarded as a translation of either of the English sub-phrases. An Example is given in Figure 9(a) and 9(b).
3) If 2) succeeds, repeat steps 2) through 4). If 2) fails, then try to divide the picked aligned pair of phrases into three, four, or more constituents in turn so that each of Japanese sub-phrases can be regarded as a translation of either of the English sub-phrases.
4) If 3) succeeds, repeat steps 2) through 4). Otherwise stop dividing the current pair of phrases and go through steps 2) through 4) with the next pair of phrases. If no more pair of phrases is available for dividing, terminate and output the set of division steps.

Figure 9 shows an example of dividing a pair of sentences into aligned phrases. The set {(a), (b)} constitutes one division step like (12a), as is also the case with sets {(c), (d)} and {(e), (f)}. From manually created division steps for the 160 sentence pairs, all the possible phrase alignments are generated and stored as a set of golden standard. Outputs of phrase aligner for these 160 sentences are then compared with the golden standard. For each phrase alignment in the golden standard, F-measure is calculated with the system output, and the maximum value among all the phrase alignments of the golden standard is recorded as the F-measure of the system output. The mean value of the F-measures of all the 160 sentences was 80.4. The average number of phrases in a sentence for the golden standard phrase alignments which give the maximum F-measure was 6.0. Therefore it is not the case that the most simple phrase alignment, which is a partition of a sentence into two parts, is earning high F-measures. In order to examine the contribution of simple phrase alignments, F-measures are calculated by gradually eliminating

| | Constraint | Preference |
|---|---|---|
| Japanese | ･conjunctions and punctuations are merged with the preceding entities | ･when the score ties, a merge which creates a constituent takes precedence |
| English | ･conjunctions, prepositions and punctuations are merged with the following entities<br>･merging across base-phrase boundary is pro-hibited | ･when the score ties, a merge which creates a constituent takes precedence. If the English pref-erence conflicts with the Japanese prece-dence, the latter takes precedence. |

Table 1: Syntactic constraints and preferences

| Training | year | size(sent) | IPC CAT |
|---|---|---|---|
| Word Alignment | 1995 | 150,000 | A-H |
| Phrase Alignment | 1995 | 13,000 | G |

Table 2: Training set description

from golden standard phrase alignments with small number of phrases. Table 3 shows the result. There are no big drops until MinNum = 4 , and after that F-measure declines rather rapidly. This also suggests that golden standard phrase alignments with 2 or three phrases are not playing a major role in the evaluation of the system outputs.

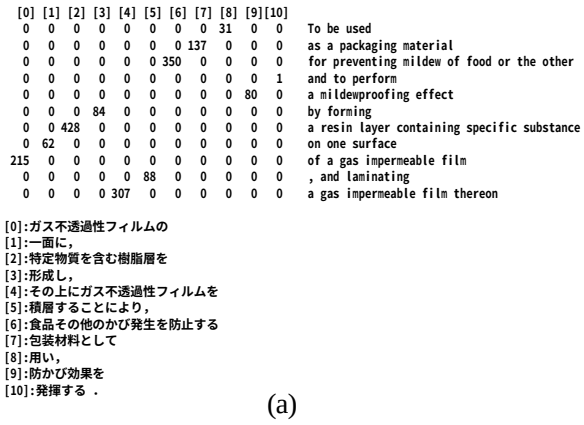
### 3.2 Evaluation of Phrases with SMT

The extracted phrase alignments were also evaluated with an SMT engine. We used Pharaoh (Koehn, 2004) as the baseline. Although our goal is to use obtained phrase alignments as translation units of Rule-based/SMT hybrid systems, we haven't yet processed large amount of parallel corpora, and the decoding scheme which takes advantage of the constituent oriented phrase alignments is still under development. Therefore, instead of testing the phrase alignments as translation units, we tested the cross-domain portability of the obtained phrase alignments. One of the major merits of a syntactic constituent is its generalization capability. N-gram statistics extracted from a large collection of data in a specific domain is a powerful resource within the same domain, but quite often fails to adopt to quite different domains. Constituents, or grammatical categories, on the other hand, cannot easily be tuned to a specific domain, but possess a generalization capability. In this experiment we trained Pharaoh using parallel sentences in one domain, namely IPC-G category (Physics), and tested the decoder in different domains. The training corpus we used for a baseline setting is the 13,000 sentence pairs in IPC-G category listed in Table 2 . We then used a set of aligned phrases extracted from the 13,000 sentence pairs for training Pharaoh (PhrAlign). The phrases are used alone and not mixed with the original parallel sentences. For testing, a set of 500 sentence pairs are randomly extracted from each IPC category. For development, another set of 500 sentence pairs are extracted from IPC-G category. Table 4 shows the result. PhrAlign outperforms Baseline in all the categoris. Especially in category E, PhrAlign scores 1.49 points higher than Baseline, which is relative percentage of 16% increase from Baseline. Since the training corpus is fairly small it is possible that the difference of the two cases decreases as the training data is increased, but this result suggests a generalizing capability of the syntactically oriented phrase alignments.

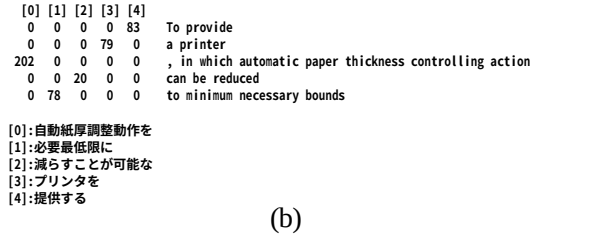
## 4 Related work

The inversion transduction grammar formalism (Wu, 1997) is one of the pioneering approaches for stochastically extracting bilingual phrases with constituent structure. A concept of bilingual parsing, where the input is a sentence pair rather than a sentence, is introduced in this framework. By allowing the inverse order of the right-hand-side of productions, the expressiveness of the grammar is shown to be considerably enhanced. In order to control the computational complexity, however, several severe constraints are applied, which makes it difficult to apply ITG to free-word-order languages like Japanese. This formalism is also not intended to be robust against the translation lexicon inadequacies: sentences containing more than one word absent form the translation lexicon are rejected in the reported experiment. The proposed method, on the other hand, is quite robust to a sparse alignment matrix because of the utilization of statistical word-alignment and the robustness of the chunkers.

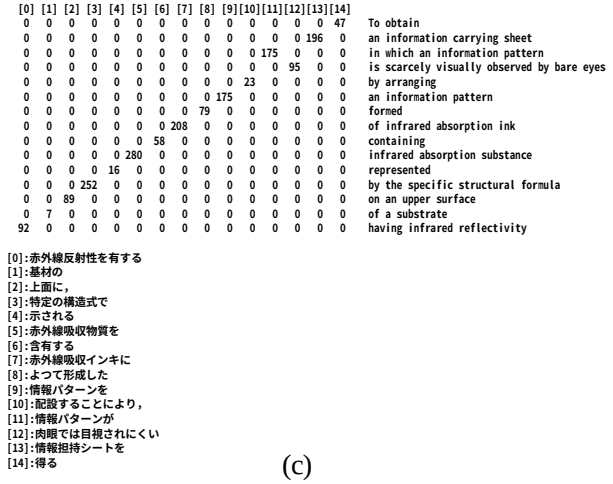
Integrated Segmentation and Alignment (Zhang and Vogel, 2005), or ISA, is probably most similar in concept to the proposed approach. ISA employs a greedy algorithm, called CGA, to extract phrase pairs out of a bilingual corpus. CGA extends the competitive linking algorithm (Melamed, 1997), a greedy word alignment algorithm with one word-to-one word assumption, to allow for combining



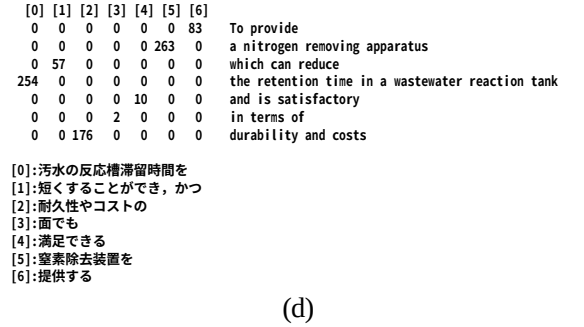

(a)




(b)




(c)




(d)

Figure 7: Examples of obtained phrase alignments

| Min num of phrases | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| F-measure | 80.4 | 78.4 | 78.4 | 72.6 | 69.6 | 64.6 |

Table 3: F-measure with minimum number of phrases in the golden standard varied

the detected "sure" word pair (a seed) with its neighbors to form a group. ISA uses $\chi^2$ statistics to measure the mutual translation likelihood between words, and the word pair with the highest $\chi^2$ value is selected as a seed. Neighboring words to be joined with the seed are also greedily searched on the basis of $\chi^2$ values. Although both approaches use a statistical measure for the decision of agglomeration, CGS uses a word-to-word association for the judgment of local grouping, whereas the proposed approach uses a sentence level, or global, association metric for the judgment of merging, which makes the merging judgment justifiable not only for the merged phrase pairs, but also for the other words and phrases in the sentence pair. The n-best search in the proposed method also avoids the greediness of the merging process. Another difference is that in order to make the computation tractable, ISA employs a "locality assumption" which requires that a source phrase of adjacent words only be aligned to a target phrase composed of adjacent words. This assumption is again not suitable for language pairs of a quite different word order like the pair of Japanese and English.

## 5 Conclusion

A novel approach is presented for extracting syntactically motivated phrase alignments. In this method we

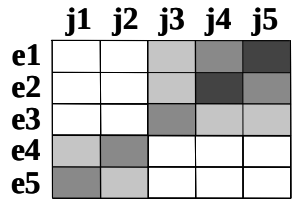



Figure 8: Multilayered Phrase Correspondences

| IPC CAT | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Baseline | 7.94 | 11.43 | 10.24 | 7.42 | 9.29 | 11.38 | 14.66 | 12.03 |
| PhrAlign | 8.91 | 11.78 | 10.85 | 8.37 | 10.78 | 12.48 | 15.70 | 13.08 |

Table 4: Bleu score of the baseline and the proposed method.

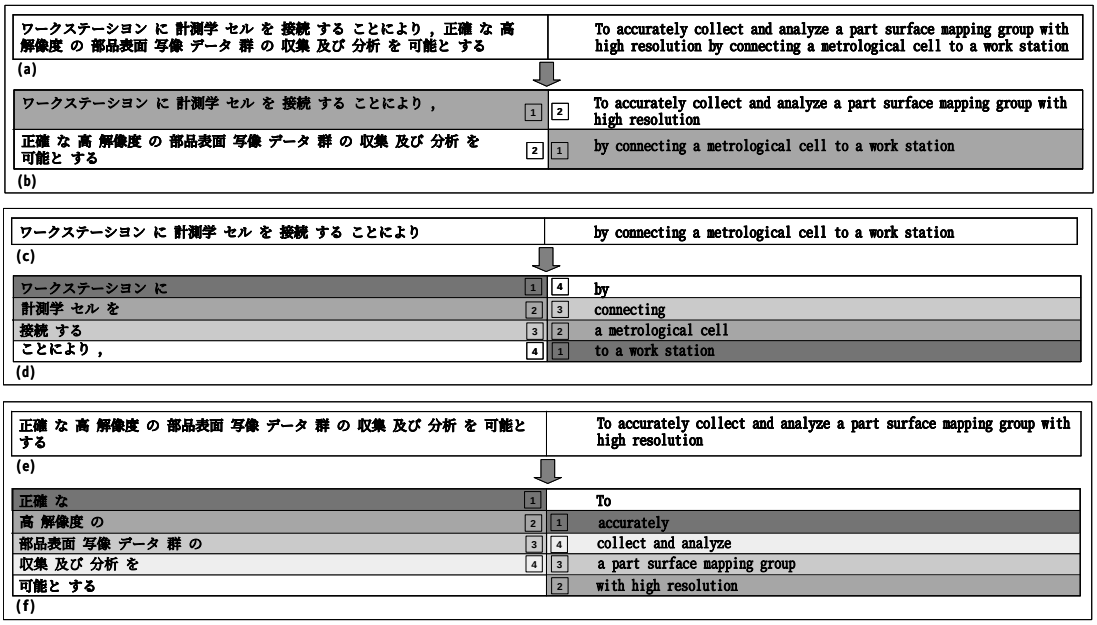



Figure 9: Manual Annotation of Phrase Alignments

can incorporate conventional resources such as dictionaries and grammar rules into a statistical optimization framework for phrase alignment. The method extracts bilingual phrases by incrementally merging adjacent words or phrases on both source and target language sides in accordance with a global statistical metric along with constraints and preferences composed by combining statistical information, dictionary information, and also grammatical rules. The extracted phrases achieved a maximum F-measure of over 80 with respect to human judged phrase alignments. The extracted phrases used as a training corpus for a phrase-based SMT showed better cross-domain portability over conventional SMT framework.

## References


Chiang, David (2005). A hierarchical phrase-based model for statistical machine translation. In Proceedings of the 43rd Annual Meeting of the ACL ( pp263-270).

Koehn, Philipp, Franz Josef Och,, and Daniel Marcu (2003). Statistical Phrase-Based Translation. In Proceedings of HLT-NAACL.

Koehn, Philipp (2004). Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In 6th Conference of the Association for Machine Translation in the Americas, AMTA.

Melamed, I. Dan (1997). A Word-to-Word Model of Translational Equivalence. In Proceedings of the Eighth Conference of the European Chapter of the Association for Computational Linguistics (pp.490-497).

Marcu, Daniel and William Wong (2002). A Phrase-based Joint Probability Model for Statistical Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp.133-139).

NTCIR Workshop (2002). http://research.nii.ac.jp/ntcir/ntcir-ws3/work-en.html.

Och , Franz Josef and Hermann Ney (2000). Improved statistical alignment models. In Proceedings of the 38th Annual Meeting of the ACL (pp.440-447).

Och, Franz-Josef and Hermann Ney (2004). The alignment template approach to statistical machine translation. Computational Linguistics, 30(4), 417--450.

Ushioda, Akira (2007). Phrase Alignment for Integration of SMT and RBMT Resources. In Proceedings of MT Summit XI Workshop on Patent Translation (to appear), Copenhagen, Denmark.

Wu, Dekai (1997). Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. Computational Linguistics, 23(3), 377-403.

Yamada , Kenji and Kevin Knight (2001). A syntax-based statistical translation model. In Proceedings of the 39th Annual Meeting of the ACL (pp.523-530).

Zhang, Ying and Stephan Vogel (2005). Competitive Grouping in Integrated Phrase Segmentation and Alignment Model. In Proceedings of the ACL Workshop on Building and Using Parallel Texts (pp 159—162).