Phrase Alignment for Integration of SMT and RBMT Resources

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 sides in accordance with a global statistical metric. Phrase alignments are extracted from parallel patent documents using this method. The extracted phrases used as training corpus for a phrase-based SMT showed 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 called 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. For a statistical optimization to be in effect, it is preferable that the initial word alignments are numerical, not zero/one. Let's take a simplified example to obtain an intuition behind the proposed method. Consider the following Japanese-English parallel sentences.

(1a) ジョンは　　　　白球を　　　　　　投げた
(John- Nominative)　(white ball-Accusative)　(threw)

(1b) John threw white balls

Figure 1 shows the degrees of correspondence (scores) between each Japanese word/phrase and English word/phrase. The score in each cell is just an illustrative figure. In Figure 1(a) each Japanese word in (1a) is arranged in a row and each English word in (1b) is arranged in a column. The dominant cells are shadowed and they are considered to show a clear correspondence. For a pair of languages with similar word order, the corresponding cells tend to align diagonally, but for languages like Japanese and English which have quite different word order, the corresponding cells are scattered. Nonetheless, when we look at local correspondences like words within a phrase, the corresponding cells come to next to each other. In this representation, when we obtain

1	2	3	
98	1	1	1 John
0	2	98	2 threw
0	98	2	3 white
0	97	3	4 balls
ジョンは (John-Nom)	白球を(white ball-Acc)	投げた (threw)	

(a)

1	2	3	
98	1	1	1 John
0	2	98	2 threw
0	195	5	3 white balls
ジョンは	白球を	投げた	

(b)

1	2	3	
98	1	1	1 John
0	100	100	2 threw white
0	97	3	3 balls
ジョンは	白球を	投げた	

(c)

Figure 1: Phrase alignment example

a one-to-one correspondence in which one and only one dominant cell appears in each row and column, we can judge that we obtained a phrase alignment. It is rarely the case that we obtain a one-to-one correspondence at the initial stage (1-a). However when we repeat merging a pair of adjacent words (or phrases) on Japanese side and English side, and adding the score of merged rows or columns, then we will eventually arrive at a one-to-one correspondence, in a worst case leaving only one row and one column. In the example of Figure 1, when we merge two adjacent English words “white” and “balls”, we reach a one-to-one correspondence (Figure 1-b). This is because “白球を” and “white balls” constitute a pair of phrases with no excess or deficiency on either side. On the other hand, when we merge “threw” and “white”, the matrix goes away from the one-to-one correspondence. We present in the next section a formal framework of the proposed method.

2. Phrase Alignment Method

Although our objective in this work is to extract alignments of phrases which are linguistically motivated, there might be cases in which a phrase in one language in a pair 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.

The phrase alignment method we propose here 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 optionally grammatical rules.

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 conventional rule-based approach is applicable.

As a preparation step, word alignments are obtained from a bilingual corpus by GIZA++ (Och & Ney, 2000) 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 $\boldsymbol{e}$ and Japanese word $\boldsymbol{j}$, the frequency $N(e)$ of e in A and the co-occurrence frequency $N(e\ ,\ j)$ 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.

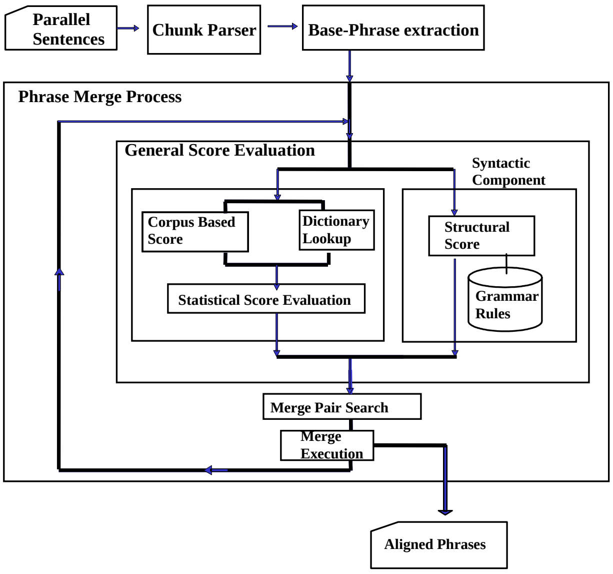


Figure 2: Framework of Phrase Aligner

An input to the phrase aligner is a pair **(J, E)** of Japanese and English sentence. The pair (**J**, **E**) is first chunk-parsed to extract base phrases, such as minimum noun phrases and phrasal verbs on both sides. Let $\boldsymbol{J = j_1, j_2, \ldots, j_M}$ 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. We now consider the probability that the translation of word w_i appears in chunk $\boldsymbol{j}_{\mathrm{j}}$ in the given sentence pair using the empirical translation probabilities $Pc(j|e)$. From the assumption that the translation of word w_i always appears somewhere in the Japanese sentence,

$$(3)\quad \Sigma_j \Sigma_t P(t|w_i)\ P(t\ \ \text{appears in}\ \mathbf{j}_j) = 1$$

, where t is the translation candidate of w_i, $P(t|\ w_i\ )$ is the probability that w_i is translated to t in the given sentence pair, and $P(t\ \ appears\ in\ \boldsymbol{j}_j\ )$ is the probability that t appears in $\boldsymbol{j}_j$. Since the sentence pair is given and fixed here, $P(t\ \ appears\ in\ \boldsymbol{j}_j)$ is zero if $\boldsymbol{j}_j$ doesn’t contain t as a substring and one if it does. Precisely speaking, there is a possibility that t appears not as a translation of w_i even if $\boldsymbol{j}_j$ contains t as a substring, but we define $P(t\ \ appears\ in\ \boldsymbol{j}_j)$ as stated above. We also make an assumption that the translation probability $P(t|w_i)$ in the given sentence pair is proportional to the empirical translation probability defined in (2). That is,

$$(4)\qquad P(t|\ w_i)=\lambda Pc(t|\ w_i)$$

for some constant λ. From (3) and (4), the probability that the translation of word w_i appears in chunk $\boldsymbol{j}_j$ is given as follows.

$$\begin{aligned}(5)\ P(\mathbf{j}_j|w_i) &= \Sigma_t P(t|\ w_i)P(t\ \text{appears in}\ \mathbf{j}_j) \\ &= \Sigma_t\ P(t\ \text{appears in}\ \mathbf{j}_j)*Pc(t|\ w_i)/\Sigma_j C_{ij} \\ &= C_{ij}\ /\ \Sigma_j C_{ij}\end{aligned}$$

, where

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

is called a *bilingual phrase matrix* which represents the relative likelihood that the translation of word w_i appears in chunk $\mathbf{j}_j$ in contrast to other Japanese chunks. Note that the values of C_{ij} can be calculated given the parallel sentence pair and the empirical translation probability.

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

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

Next we consider the degree of uncertainty as to in which Japanese chunk the translation of w_i appears. For example, if $P(\mathbf{j}_j \mid w_i) = 1$ then it is certain that the translation of w_i appears in $\mathbf{j}_j$, that is, the entropy of the probability distribution $P(\cdot \mid w_i)$ is zero. The entropy $H(i)$ of the probability distribution $P(\cdot \mid w_i)$ in general is given as follows.

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

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

In the proposed method, a statistical metric based on the entropy (8) 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 bilingual phrase matrix as a dataset. The entire entropy, or uncertainty, of mapping English phrases to Japanese phrases is then given by:

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

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

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

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

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

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

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). Then if there are no syntactic constraints or preferences, the merging candidate which has 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.

	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 ・merging across base-phrase boundary is prohibited	・when the score ties, a merge which creates a constituent takes precedence. If the English preference conflicts with the Japanese precedence, the latter takes precedence.

Table 1: Syntactic constraints and preferences

3. Experiments on Parallel Patent Corpus

This section describes experiments with the proposed phrase alignment method on a parallel patent corpus.

We used the test collection of 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. The description of the IPC categories is given in Table 2. 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 of the above word alignment training set. The sentence length is constrained to reduce the computation load. Table 3 summarizes the training corpora used.

Category	A	B	C
Description	HUMAN NECESSITIES	PERFORMING OPERATIONS; TRANSPORTING	CHEMISTRY; METALLURGY
Category	D	E	F
Description	TEXTILES; PAPER	FIXED CONSTRUC-TIONS	MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING
Category	G	H	
Description	PHYSICS	ELECTRICITY	

Table 2: IPC Categories

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

Table 3: Training set description

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 3 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.

The extracted phrase alignments were 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 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 be tuned easily 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, and tested the decoder in different domains. The training corpus we used is the 13,000 sentence pairs in IPC-G category listed in Table 3 for as a baseline setting.

We also used a set of aligned phrases extracted form 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 of the year 1996. For development another set of 500 sentence pairs are extracted from the IPC-G category of the year 1996. Table 4 shows the result. PhrAlign outperforms Baseline in all the categories. 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.

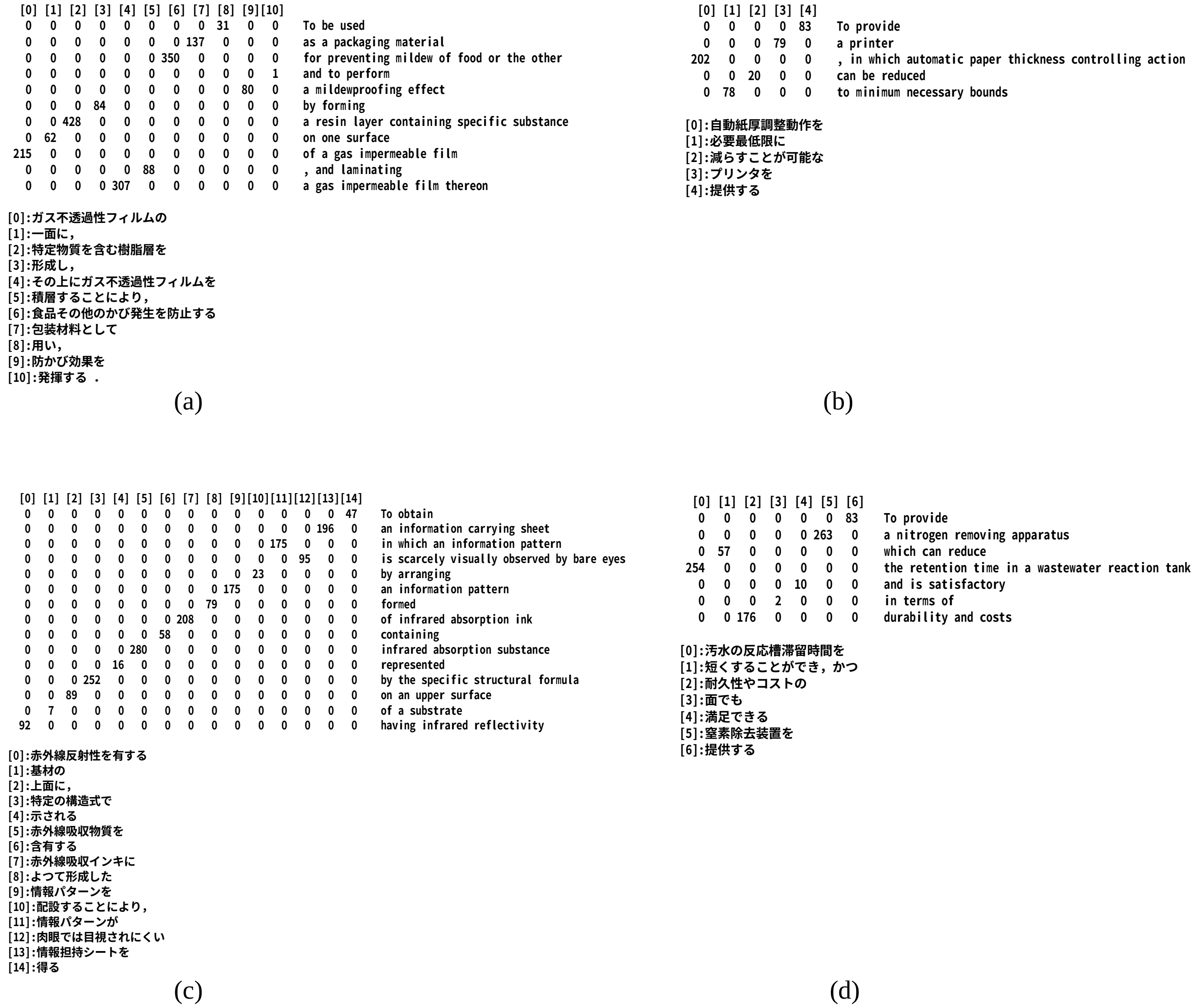
4. Conclusion

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 sides in accordance with a global statistical metric along with constraints and preferences composed by combining statistical information, dictionary information, and also grammatical rules. Phrase alignments are extracted from parallel patent corpus using the method. The extracted phrases used as training corpus for a phrase-based SMT shows 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.



```
  [0] [1] [2] [3] [4] [5] [6] [7] [8] [9][10]
   0   0   0   0   0   0   0   0  31   0   0    To be used
   0   0   0   0   0   0   0 137   0   0   0    as a packaging material
   0   0   0   0   0   0 350   0   0   0   0    for preventing mildew of food or the other
   0   0   0   0   0   0   0   0   0   0   1    and to perform
   0   0   0   0   0   0   0   0   0  80   0    a mildewproofing effect
   0   0   0  84   0   0   0   0   0   0   0    by forming
   0   0 428   0   0   0   0   0   0   0   0    a resin layer containing specific substance
   0  62   0   0   0   0   0   0   0   0   0    on one surface
 215   0   0   0   0   0   0   0   0   0   0    of a gas impermeable film
   0   0   0   0   0  88   0   0   0   0   0    , and laminating
   0   0   0   0 307   0   0   0   0   0   0    a gas impermeable film thereon

[0]:ガス不透過性フィルムの
[1]:一面に，
[2]:特定物質を含む樹脂層を
[3]:形成し，
[4]:その上にガス不透過性フィルムを
[5]:積層することにより，
[6]:食品その他のかび発生を防止する
[7]:包装材料として
[8]:用い，
[9]:防かび効果を
[10]:発揮する .
```

(a)

```
  [0] [1] [2] [3] [4]
   0   0   0   0  83    To provide
   0   0   0  79   0    a printer
 202   0   0   0   0    , in which automatic paper thickness controlling action
   0   0  20   0   0    can be reduced
   0  78   0   0   0    to minimum necessary bounds

[0]:自動紙厚調整動作を
[1]:必要最低限に
[2]:減らすことが可能な
[3]:プリンタを
[4]:提供する
```

(b)

```
  [0] [1] [2] [3] [4] [5] [6] [7] [8] [9][10][11][12][13][14]
   0   0   0   0   0   0   0   0   0   0   0   0   0   0  47    To obtain
   0   0   0   0   0   0   0   0   0   0   0   0   0 196   0    an information carrying sheet
   0   0   0   0   0   0   0   0   0   0   0 175   0   0   0    in which an information pattern
   0   0   0   0   0   0   0   0   0   0   0   0  95   0   0    is scarcely visually observed by bare eyes
   0   0   0   0   0   0   0   0   0   0  23   0   0   0   0    by arranging
   0   0   0   0   0   0   0   0   0 175   0   0   0   0   0    an information pattern
   0   0   0   0   0   0   0   0  79   0   0   0   0   0   0    formed
   0   0   0   0   0   0   0 208   0   0   0   0   0   0   0    of infrared absorption ink
   0   0   0   0   0   0  58   0   0   0   0   0   0   0   0    containing
   0   0   0   0   0 280   0   0   0   0   0   0   0   0   0    infrared absorption substance
   0   0   0   0  16   0   0   0   0   0   0   0   0   0   0    represented
   0   0   0 252   0   0   0   0   0   0   0   0   0   0   0    by the specific structural formula
   0   0  89   0   0   0   0   0   0   0   0   0   0   0   0    on an upper surface
   0   7   0   0   0   0   0   0   0   0   0   0   0   0   0    of a substrate
  92   0   0   0   0   0   0   0   0   0   0   0   0   0   0    having infrared reflectivity

[0]:赤外線反射性を有する
[1]:基材の
[2]:上面に，
[3]:特定の構造式で
[4]:示される
[5]:赤外線吸収物質を
[6]:含有する
[7]:赤外線吸収インキに
[8]:よつて形成した
[9]:情報パターンを
[10]:配設することにより，
[11]:情報パターンが
[12]:肉眼では目視されにくい
[13]:情報担持シートを
[14]:得る
```

(c)

```
  [0] [1] [2] [3] [4] [5] [6]
   0   0   0   0   0   0  83    To provide
   0   0   0   0   0 263   0    a nitrogen removing apparatus
   0  57   0   0   0   0   0    which can reduce
 254   0   0   0   0   0   0    the retention time in a wastewater reaction tank
   0   0   0   0  10   0   0    and is satisfactory
   0   0   0   2   0   0   0    in terms of
   0   0 176   0   0   0   0    durability and costs

[0]:汚水の反応槽滞留時間を
[1]:短くすることができ，かつ
[2]:耐久性やコストの
[3]:面でも
[4]:満足できる
[5]:窒素除去装置を
[6]:提供する
```

(d)

Figure 3: Examples of obtained phrase alignments

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 baseline and the proposed method

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.

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.

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).