Statistical Phrase Alignment Model
Using Dependency Relation Probability

Toshiaki Nakazawa   **Sadao Kurohashi**

Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
nakazawa@nlp.kuee.kyoto-u.ac.jp   kuro@i.kyoto-u.ac.jp

Abstract

When aligning very different language pairs, the most important needs are the use of structural information and the capability of generating one-to-many or many-to-many correspondences. In this paper, we propose a novel phrase alignment method which models word or phrase dependency relations in dependency tree structures of source and target languages. The dependency relation model is a kind of tree-based reordering model, and can handle non-local reorderings which sequential word-based models often cannot handle properly. The model is also capable of estimating phrase correspondences automatically without any heuristic rules. Experimental results of alignment show that our model could achieve F-measure 1.7 points higher than the conventional word alignment model with symmetrization algorithms.

1 Introduction

We consider that there are two important needs in aligning parallel sentences written in very different languages such as Japanese and English. One is to adopt structural or dependency analysis into the alignment process to overcome the difference in word order. The other is that the method needs to have the capability of generating phrase correspondences, that is, one-to-many or many-to-many word correspondences. Most existing alignment methods simply consider a sentence as a sequence of words (Brown et al., 1993), and generate phrase correspondences using heuristic rules (Koehn et al., 2003). Some studies incorporate structural information into the alignment process *after* this simple word align-

ment (Quirk et al., 2005; Cowan et al., 2006). However, this is not sufficient because the basic word alignment itself is not good.

On the other hand, a few models have been proposed which use structural information from the beginning of the alignment process. Watanabe et al. (2000) and Menezes and Richardson (2001) proposed a structural alignment methods. These methods use heuristic rules when resolving correspondence ambiguities. Yamada and Knight (2001) and Gildea (2003) proposed a tree-based probabilistic alignment methods. These methods reorder, insert or delete sub-trees on one side to reproduce the other side, but the constraints of using syntactic information is often too rigid. Yamada and Knight flattened the trees by collapsing nodes. Gildea cloned sub-trees to deal with the problem. Cherry and Lin (2003) proposed a model which uses a source side dependency tree structure and constructs a discriminative model. However, there is the defect that its alignment unit is a word, so it can only find one-to-one alignments. Nakazawa and Kurohashi (2008) also proposed a model focusing on the dependency relations. Their model has the constraint that content words can only correspond to content words on the other side, and the same applies for function words. This sometimes leads to an incorrect alignment. We have removed this constraint to make more flexible alignments possible. Moreover, in their model, some function words are brought together, and thus they cannot handle the situation where each function word corresponds to a different part. The smallest unit of our model is a single word, which should solve this problem.

In this paper, we propose a novel phrase alignment method which models word or phrase dependency relations in dependency tree structures of source and target languages. For a pair of correspondences which has a parent-child relation on one side, the dependency relation on the other side is defined as the relation between the two correspondences. It is a kind of tree-based reordering model, and can capture non-local reorderings which sequential word-based models often cannot handle properly. The model is also capable of estimating phrase correspondences automatically without heuristic rules. The model is trained in two steps: Step 1 estimates word translation probabilities, and Step 2 estimates phrase translation probabilities and dependency relation probabilities. Both Step 1 and Step 2 are performed iteratively by the EM algorithm. During the Step 2 iterations, word correspondences are grown into phrase correspondences.

2   Proposed Model

We suppose that Japanese is the source language and English is the target language in the description of our model. Note that the model is not specialized for this language pair, and it can be applied to any language pair.

Because our model uses dependency tree structures, both source and target sentences are parsed beforehand. Japanese sentences are converted into dependency structures using the morphological analyzer JUMAN (Kurohashi et al., 1994), and the dependency analyzer KNP (Kawahara and Kurohashi, 2006). MSTparser (McDonald et al., 2005) is used to convert English sentences. Figure 1 shows an example of dependency structures. The root of a tree is placed at the extreme left and words are placed from top to bottom.

2.1   Overview

This section outlines our proposed model in comparison to the IBM models, which are the conventional statistical alignment models.

In the IBM models (Brown et al., 1993), the best alignment $\hat{\mathbf{a}}$ between a given source sentence $\mathbf{f}$ and its target sentence $\mathbf{e}$ is acquired by the following equation:

$$\begin{aligned} \hat{\mathbf{a}} &= \underset{\mathbf{a}}{\operatorname{argmax}}\, p(\mathbf{f}, \mathbf{a}|\mathbf{e}) \\ &= \underset{\mathbf{a}}{\operatorname{argmax}}\, p(\mathbf{f}|\mathbf{e}, \mathbf{a}) \cdot p(\mathbf{a}|\mathbf{e}) \end{aligned} \quad (1)$$
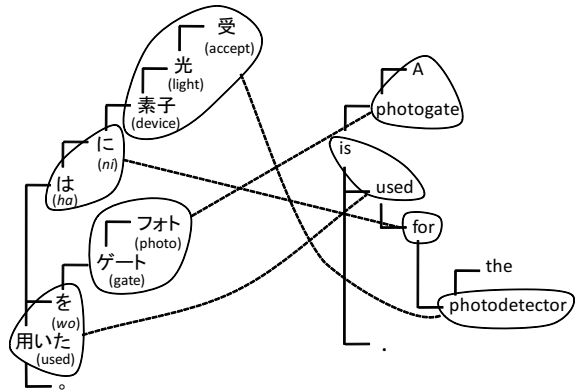

Figure 1: An example of a dependency tree and its alignment.

where $p(\mathbf{f}|\mathbf{e}, \mathbf{a})$ is called *lexicon probability* and $p(\mathbf{a}|\mathbf{e})$ is called *alignment probability*.

Suppose $\mathbf{f}$ consists of n words $f_1, f_2, ..., f_n$, and $\mathbf{e}$ consists of m words $e_1, e_2, ..., e_m$ and a NULL word (e_0). The alignment mapping $\mathbf{a}$ consists of associations $j \rightarrow i = a_j$ from source position j to target position $i = a_j$. The two probabilities above are broken down as:

$$p(\mathbf{f}|\mathbf{e}, \mathbf{a}) = \prod_{j=1}^{J} p(f_j|e_{a_j}) \quad (2)$$

$$p(\mathbf{a}|\mathbf{e}) = \prod_{i=1}^{I} p(\Delta j|e_i) \quad (3)$$

where Δj is a relative position of words in the source side which corresponds to e_i. Equation 2 is the product of the word translation probabilities, and Equation 3 is the product of relative position probabilities.

In the proposed model, we refine the IBM models in three ways. First, as for Equation 2, we consider phrases instead of words. Second, as for Equation 3, we consider dependencies of words instead of their positions in a sentence.

Finally, the proposed model can find the best alignment $\hat{\mathbf{a}}$ by not using $\mathbf{f}$-to-$\mathbf{e}$ alone, but simultaneously with $\mathbf{e}$-to-$\mathbf{f}$. That is, Equation 1 is modified as follows:

$$\begin{aligned} \hat{\mathbf{a}} &= \underset{\mathbf{a}}{\operatorname{argmax}} \quad p(\mathbf{f}|\mathbf{e}, \mathbf{a}) \cdot p(\mathbf{a}|\mathbf{e}) \cdot \\ & \qquad\qquad p(\mathbf{e}|\mathbf{f}, \mathbf{a}) \cdot p(\mathbf{a}|\mathbf{f}) \end{aligned} \quad (4)$$

Since our model regards a phrase as a basic unit, the above formula is calculated in a straightforward way. In contrast, the IBM models can consider a many-to-one alignment by combining one-to-one

alignments, but they cannot consider a one-to-many or many-to-many alignment.

The models are estimated by EM-like algorithm which is very similar to (Liang et al., 2006). The important difference is that we are using tree structures.

We maximize the data likelihood:

$$\max_{\theta_{ef},\theta_{fe}} \sum_{\mathbf{f},\mathbf{e}} (\log p_{ef}(\mathbf{f},\mathbf{e};\theta_{ef}) + \log p_{fe}(\mathbf{f},\mathbf{e};\theta_{fe}))$$

(5)

In the E-step, we compute the posterior distribution of the alignments with the current parameter θ:

$$q(\mathbf{a};\mathbf{f},\mathbf{e}) := p_{ef}(\mathbf{a}|\mathbf{f},\mathbf{e};\theta_{ef}) \cdot p_{fe}(\mathbf{a}|\mathbf{f},\mathbf{e};\theta_{fe}) \quad (6)$$

In the M-step, we update the parameter θ:

$$
\begin{aligned}
\theta' := & \operatorname*{argmax}_{\theta} \sum_{\mathbf{a},\mathbf{f},\mathbf{e}} q(\mathbf{a};\mathbf{f},\mathbf{e}) \log p_{ef}(\mathbf{a},\mathbf{f},\mathbf{e};\theta_{ef}) \\
& + \sum_{\mathbf{a},\mathbf{f},\mathbf{e}} q(\mathbf{a};\mathbf{f},\mathbf{e}) \log p_{fe}(\mathbf{a},\mathbf{f},\mathbf{e};\theta_{fe}) \\
= & \operatorname*{argmax}_{\theta} \sum_{\mathbf{a},\mathbf{f},\mathbf{e}} q(\mathbf{a};\mathbf{f},\mathbf{e}) \log p(\mathbf{e}) \cdot p_{ef}(\mathbf{a},\mathbf{f}|\mathbf{e};\theta_{ef}) \\
& + \sum_{\mathbf{a},\mathbf{f},\mathbf{e}} q(\mathbf{a};\mathbf{f},\mathbf{e}) \log p(\mathbf{f}) \cdot p_{fe}(\mathbf{a},\mathbf{e}|\mathbf{f};\theta_{fe})
\end{aligned}
$$

(7)

Note that $p(\mathbf{e})$ and $p(\mathbf{f})$ have no effect on maximization, and $p_{ef}(\mathbf{a},\mathbf{f}|\mathbf{e};\theta_{ef})$ and $p_{fe}(\mathbf{a},\mathbf{e}|\mathbf{f};\theta_{fe})$ appeared in Equation 1 or Equation 4.

In the following sections, we decompose the lexicon probability and alignment probability.

2.2 Phrase Translation Probability

Suppose $\mathbf{f}$ consists of N phrases $F_1, F_2, ..., F_N$, and $\mathbf{e}$ consists of M phrases $E_1, E_2, ..., E_M$. The alignment mapping $\mathbf{a}$ consists of associations $j \to i = A_j$ from source phrase j to target phrase $i = A_j$.

We consider *phrase translation probability* $p(F_j|E_i)$ instead of word translation probability. There is one restriction: that phrases composed of more than one word cannot be aligned to NULL. Only a single word can be aligned to NULL.

We denote a phrase which the word f_j belongs to as $F_{s(j)}$, and a phrase which the word e_i belongs to as $E_{t(i)}$. With these notations, we refine Equation 2 as follows:

$$p(\mathbf{f}|\mathbf{e},\mathbf{a}) = \prod_{j=1}^{J} p(F_{s(j)}|E_{A_{s(j)}}) \quad (8)$$

Suppose phrase F_j and E_i are aligned where the number of words in F_j is denoted by $|F_j|$ and that number in E_i is $|E_i|$, the probability mass related to this alignment in Equation 8 is as follows:

$$p(F_j|E_i)^{|F_j|} \cdot p(E_i|F_j)^{|E_i|} \quad (9)$$

We call this probability for the link between F_j and E_i *phrase alignment probability*. The upper part of Table 1 shows phrase alignment probabilities for the alignment in Figure 1.

2.3 Dependency Relation Probability

The reordering model in the IBM Models is defined on the relative position between an alignment and its previous alignment, as shown in Equation 3. Our model, on the other hand, considers dependencies of words instead of positional relations.

We start with a dependency relation where f_c depends on f_p in the source sentence. In a possible alignment, f_c belongs to $F_{s(c)}$, f_p belongs to $F_{s(p)}$, and $F_{s(c)}$ depends on $F_{s(p)}$. In this situation, we consider the relation between $E_{A_{s(p)}}$ and $E_{A_{s(c)}}$. Even if two languages have different word order, their dependency structures are similar in many cases, and $E_{A_{s(c)}}$ tends to depend on $E_{A_{s(p)}}$. Our model takes this tendency into consideration. In order to denote the relationship between phrases, we introduce $rel(E_{A_{s(p)}}, E_{A_{s(c)}})$. This is defined as the path from $E_{A_{s(p)}}$ to $E_{A_{s(c)}}$. It is represented by applying the notations below:

- 'c' if going down to the *c*hild node
- 'p' if going down to the *p*arent node

For example, in Figure 1, the path from "for" to "photodetector" is 'c', from "the" to "for" is 'p;p' because it travels across two nodes. All the phrases are considered as a single node, so the path from "photogate" to "the" is 'p;c;c;c' with the alignment in Figure 1.

We refine Equation 3 using rel as follows:

$$p(\mathbf{a}|\mathbf{e}) = \prod_{(p,c) \in D_{s\text{-}pc}} p_t(rel(E_{A_{s(p)}}, E_{A_{s(c)}})|\text{pc})$$

(10)

where $D_{s\text{-}pc}$ denotes a set of parent-child word pairs in the source sentence. We call $p_t(rel(E_{A_{s(p)}}, E_{A_{s(c)}})|\text{pc})$ target side *dependency relation probability*. p_t is a kind of tree-based reordering model.

12

Table 1: A probability calculation example.

Source	Target	Phrase alignment probability
	photodetector	$p(\quad |\text{photodetector})^3 \cdot p(\text{photodetector}|\quad)$
	for	$p(\quad |\text{for})^2 \cdot p(\text{for}|\quad)$
	photogate	$p(\quad |\text{a photogate})^2 \cdot p(\text{a photogate}|\quad)^2$
	is used	$p(\quad |\text{is used})^2 \cdot p(\text{is used}|\quad)^2$
NULL	the	$p(\text{the}|\text{NULL})$

Source		Target dependency	Target		Source dependency
c	p	relation probability	c	p	relation probability
		$p_t(\text{SAME}|\text{pc})$	A	photogate	$p_s(\text{SAME}|\text{pc})$
		$p_t(\text{SAME}|\text{pc})$	photogate	is	$p_s(\text{c}|\text{pc})$
		$p_t(\text{c}|\text{pc})$	used	is	$p_s(\text{SAME}|\text{pc})$
		$p_t(\text{SAME}|\text{pc})$	for	used	$p_s(\text{c}|\text{pc})$
		$p_t(\text{c}|\text{pc})$	the	photodetector	$p_s(\text{NULL\_c}|\text{pc})$
		$p_t(\text{SAME}|\text{pc})$	photodetector	for	$p_s(\text{c}|\text{pc})$
		$p_t(\text{c}|\text{pc})$			
		$p_t(\text{SAME}|\text{pc})$			

There are some special cases for rel. When $F_{s(c)}$ and $F_{s(p)}$ are the same, that is, f_c and f_p belong to the same phrase, rel is represented as 'SAME'. When f_p is aligned to NULL, f_c is aligned to NULL, and both of them are aligned to NULL, rel is represented as 'NULL_p', 'NULL_c', and 'NULL_b', respectively. The lower part of Table 1 shows dependency relation probabilities corresponding to Figure 1.

Actually, we extend the dependency relation probability to consider a wider relation, i.e, the grandparent-child relation, as follows:

$$p(\mathbf{a}|\mathbf{e}) = \prod_{(p,c) \in D_{s\text{-}pc}} p_t(rel(E_{A_{s(p)}}, E_{A_{s(c)}})|\text{pc}) \cdot$$
$$\prod_{(g,c) \in D_{s\text{-}gc}} p_t(rel(E_{A_{s(g)}}, E_{A_{s(c)}})|\text{gc}) \quad (11)$$

where $D_{s\text{-}gc}$ denotes a set of grandparent-child word pairs in the source sentence.

3 Model Training

Our model is trained in two steps. In Step 1, word translation probability is estimated. Then, in Step 2, possible phrases are acquired, and both phrase translation probability and dependency relation probability are estimated. In both steps, parameter estimation is done with the EM algorithm.

3.1 Step 1

In Step 1, word translation probability in each direction is estimated independently. This is done in exactly the same way as in IBM Model 1.

In this process, the alignment unit is a word. When we consider $\mathbf{f}$-to-$\mathbf{e}$ alignment, each word on the source side f_j can correspond to a word on the target side e_i or a NULL word, independently of other source words. The probability of one possible alignment $\mathbf{a}$ is calculated as follows:

$$p(\mathbf{a}, \mathbf{f}|\mathbf{e}) = \prod_{j=1}^{J} p(f_j|e_{a_j}) \quad (12)$$

By considering all possible alignments, $p(\mathbf{f}|\mathbf{e})$ is calculated as:

$$p(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} p(\mathbf{a}, \mathbf{f}|\mathbf{e}) \quad (13)$$

As initial parameters of $p(f|e)$, we use uniform probabilities. Then, after calculating Equation 12 and 13, we give the fractional count $\frac{p(\mathbf{a}, \mathbf{f}|\mathbf{e})}{p(\mathbf{f}|\mathbf{e})}$ to all word alignments in $\mathbf{a}$, and we estimate $p(f|e)$ by MLE. We perform this estimation iteratively.

The inverse model $\mathbf{e}$-to-$\mathbf{f}$ can be calculated in the same manner.

3.2 Step 2

Both phrase translation probability and dependency relation probability are estimated, and one undirected alignment is found using the $\mathbf{e}$-to-$\mathbf{f}$ and $\mathbf{f}$-to-$\mathbf{e}$ probabilities simultaneously in this step. In contrast to Step 1, it is impossible to enumerate all the possible alignments. To find the best alignment, we first create an initial alignment based on phrase translation probability only, and then gradually revise it

13

by considering the dependency relation probability with a hill-climbing algorithm.

The initial parameters of Step 2 are calculated as follows. The dependency relation probability is calculated using the final alignment result of Step 1, and we use the word translation probability estimated in Step 1 as the initial phrase translation probability.

3.2.1 Initial Alignment

We first create an initial alignment based on the phrase translation probability without considering the dependency relation probabilities.

For all the combinations of possible phrases (including NULL), phrase alignment probabilities are calculated (equation 9). Correspondences are adopted one by one in descending order of geometric mean of the phrase alignment probabilities. All the words should be aligned only once, that is, the correspondences are adopted exclusively. Generation of possible phrases is explained in Section 3.2.3.

3.2.2 Hill-climbing

To find better alignments, the initial alignment is gradually revised with a hill-climbing algorithm. We use four kinds of revising operations:

Swap: Focusing on any two correspondences, the partners are swapped. In the first step in Figure2, the correspondences " ↔ photogate" and " ↔ photodetector" are swapped to " ↔ photodetector" and " ↔ photogate".

Extend: Focusing on one correspondence, the source or target phrase is extended to include its neighboring (parent or child) NULL-aligned word.

Add: A new correspondence is added between a source word and a target word both of which are aligned to NULL.

Reject: A correspondence is rejected and the source and target phrase are aligned to NULL.

Figure 2 shows an illustrative example of hill climbing. The alignment is revised only if the alignment probability gets increased. It is repeated until no operation can improve the alignment probability, and the final state is the best approximate alignment. As a by-product of hill-climbing, pseudo n-best alignment can be acquired. It is used in collecting fractional counts.

3.2.3 Phrase Generation

If there is a word which is aligned to NULL in the best approximate alignment, a new possible phrase is generated by merging the word into a neighboring phrase which is not aligned to NULL. In the last alignment result in Figure 2, for example, " " is treated as being included in the correspondence between " " and "photodetector" and the correspondence between " " and "for". As a result, we consider the correspondence between "

" and "photodetector" and the correspondence between " " and "for" existing in parallel sentences. The new possible phrase is taken into consideration from the next iteration.

3.2.4 Model Estimation

Collecting all the alignment results, we estimate phrase alignment probabilities and dependency relation probabilities.

One way of estimating parameters of phrase alignment probabilities is using the following equations:

$$p(F_j|E_i) = \frac{C(F_j, E_i)}{\sum_k C(F_k, E_i)}$$

$$p(E_i|F_j) = \frac{C(F_j, E_i)}{\sum_k C(E_k, F_j)} \quad (14)$$

where $C(F_j, E_i)$ is a frequency of F_j and E_i is aligned.

However, if we use this in our model, the phrase translation probability of the new possible phrase can become extremely high (often it becomes 1). To avoid this problem, we use the equations below for the estimation of phrase translation probability in place of Equation 14:

$$p(F_j|E_i) = \frac{C(F_j, E_i)}{C(E_i)}, p(E_i|F_j) = \frac{C(F_j, E_i)}{C(F_j)} \quad (15)$$

$C(E_i)$ is the frequency of the phrase E_i in the training corpus which can be pre-counted. This definition can resolve the problem where the phrase translation probability of the new possible phrase becomes too high.

As for the NULL, we use Equation 14 because we cannot pre-count the frequency of NULL.

Using the estimated phrase alignment probabilities and dependency relation probabilities, we go back to the initial alignment described in Section 3.2.1 iteratively.
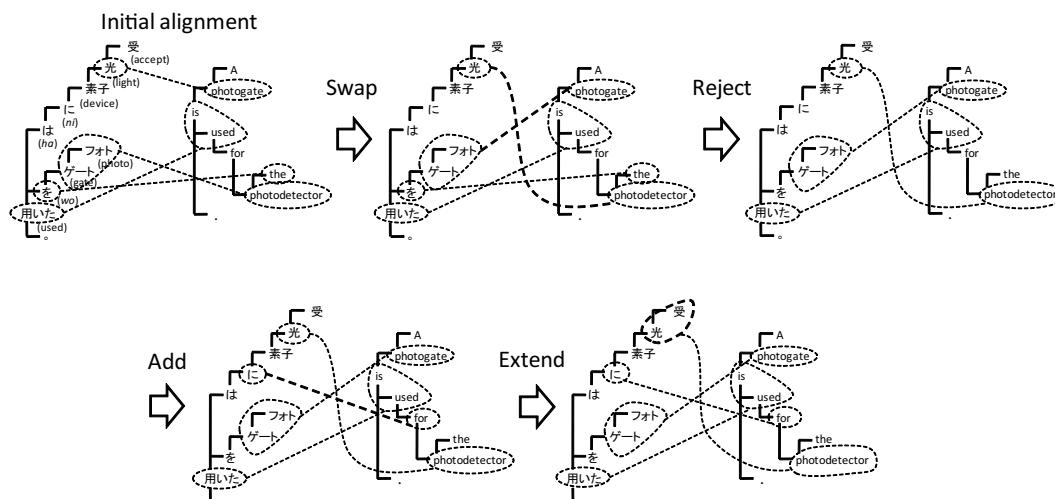
Figure 2: An example of hill-climbing.

4 Experimental Results

We conducted alignment experiments. A JST[1] Japanese-English paper abstract corpus consisting of 1M parallel sentences was used for the model training. This corpus was constructed from a 2M Japanese-English paper abstract corpus by NICT[2] using the method of Uchiyama and Isahara (2007). As gold-standard data, we used 475 sentence pairs which were annotated by hand. The annotations were only sure (S) alignments (there were no possible (P) alignments) (Och and Ney, 2003). The unit of evaluation was word-base for both Japanese and English. We used precision, recall, and F-measure as evaluation criteria.

We conducted two experiments to reveal 1) the contribution of our proposed model compared to the existing models, and 2) the effectiveness of using dependency tree structure and phrases, which are larger alignment units than words. Trainings were run on the original forms of words for both the proposed model and the models used for comparison.

4.1 Comparison with Word Sequential Model

For comparison, we used GIZA++ (Och and Ney, 2003) which implements the prominent sequential word-base statistical alignment model of IBM Models. We conducted word alignment bidirectionally with its default parameters and merged them using three types of symmetrization heuristics (Koehn et al., 2003). The results are shown in Table 2.

The result of 'Step 1' uses parameters estimated after 5 iterations of Step 1. The alignment is obtained by the method of initial alignment shown in Section 3.2.1. In 'Step 2-1', the phrase translation probabilities are the same as those in 'Step 1'. In addition, dependency relation probabilities estimated from the 'Step 1' alignment result are used. By comparing 'Step 1' and 'Step 2-1', we can see the effectiveness of dependency relation probability. We performed 5 iterations for Step 2 and calculated the alignment accuracy each time. As a result, the proposed model could achieve a higher F-measure by 1.7 points compared to the sequential model. 'Intersection' achieved best Precision, but its Recall is quite low. 'grow-diag-final-and' achieved best Recall, but its Precision is lower than our best result where the Recall is almost same. Thus, we can say our result is better than sequential word alignment models.

4.2 Effectiveness of Dependency Trees and Phrases

To confirm the effectiveness of dependency trees and phrases, we conducted alignment experiments on the following four conditions:

- Using both dependency trees and phrases (referred to as 'proposed').
- Using dependency trees only.
- Using phrases only.
- Not using dependency trees or phrases (referred to as 'none')

For the conditions which do not use dependency trees, we used positional relations of a sentence as

Table 2: Results of alignment experiment.

	Precision	Recall	F
Step 1	77.55	33.92	47.20
Step 2-1	83.46	40.03	54.11
Step 2-2	87.74	45.37	59.81
Step 2-3	87.62	48.92	62.79
Step 2-4	86.87	50.42	63.81
Step 2-5	85.90	50.75	63.80
Step 2-6	85.54	51.00	**63.90**
Step 2-7	85.18	50.87	63.70
Step 2-8	84.66	50.75	63.46
intersection	**90.34**	34.28	49.71
grow-final-and	81.32	48.85	61.04
grow-diag-final-and	79.39	**51.15**	62.22

Table 3: Effectiveness of dependency trees and phrases (results after 5 iterations in Step 2.)

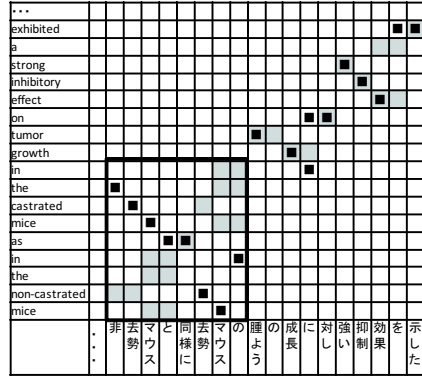|	Precision	Recall	F
proposed	85.54	**51.00**	**63.90**
dependency tree only	**89.77**	39.47	54.83
phrase only	84.41	47.33	60.65
none	85.07	38.06	52.59

Figure 3: An alignment example of the word sequential model (grow-diag-final-and).
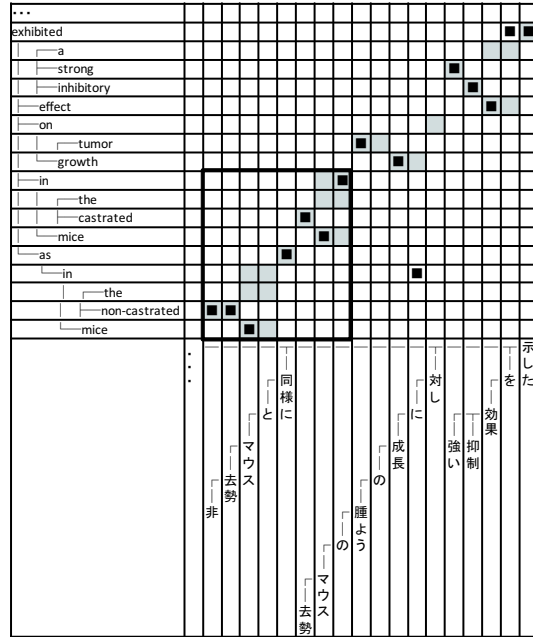
Figure 4: An alignment example of the proposed model.

a sequence of words instead of dependency tree relations. The results are shown in Table 3. All the results are the alignment accuracy after 5 iterations of Step 2.

5 Discussion

Table 2 shows that our proposed model could achieve reasonably high accuracy of alignment, and is better than sequential word-base models. As an example, alignment results of a word sequential model are shown in Figure 3. The gray colored cells are the gold-standard alignments, and the black boxes are the outputs of the sequential model. The model failed to resolve the correspondence ambiguities between " (*not*) (*castrated*) (*mice*)", and " "; and "non-castrated mice", and "castrated mice" respectively. This is because these words are placed close to each other and are also close to the correspondence " ↔ as" which can be a clue to the word order. Using the tree structure in Figure 4, these words were correctly aligned. This is because in the English tree, the phrase "castrated mice" does not depend on "as", and "non-castrated mice" does. Similarly in the Japanese tree, " " depends on " " and " " does not.

As mentioned in Section 1, sequential statistical

methods, which regard a sentence as a sequence of words, work well for language pairs that are not too different in their language structure. Japanese and English have significantly different structures. One of the issues is that Japanese sentences have a SOV word order, but in English, the word order is SVO, so the dependency relations are often turned over. For language pairs such as Japanese and English, deeper sentence analysis using NLP resources is necessary and useful. Our method is therefore suitable for such language pairs.

As another example of an alignment failure by the sequential model, Figure 5 shows the phrase correspondence " ↔ photodetector", which was correctly found as shown in Figure 6. The pro-

	受	光	素子	に	は	フォト	ゲート	を	用いた
A									
photogate						■	■		
is									■
used									■
for								■	
the							■		
photodetector	■								

Figure 5: An unsuccessful example of phrase detection in the sequential model (grow-diag-final-and).

	受	光	素子	に	は	フォト	ゲート	を	用いた
┌─A						■	■		
┌─photogate						■	■		
is								■	■
└─used								■	■
└─for						■	■		
┌─the									
└─photodetector	■	■	■						

Figure 6: An example of phrase detection in the proposed model.

posed method of generating possible phrases during iterations works well and improves alignment.

From the result of our second experiment, we can see the following points:

1. Phrasal alignment improves the recall, but lowers the precision.
2. By using dependency trees, precision can be improved.
3. We can find a balance point by using both phrasal alignment and dependency trees.

The causes of alignment errors in our model can be summarized into categories. The biggest one is parsing errors. Since our model is highly dependent on the parsing result, the alignments would easily turn out wrong if the parsing result was incorrect.

Sometimes the hill-climbing algorithm could not revise the initial alignment. Most of these cases would happen when one word occurred several times on one side, but some of those occurrences were omitted on the other side. Let's suppose there are two identical words on the source side, but the target side has only one corresponding word. Initial alignment is created without considering the dependencies at all, so it cannot judge which source word should be aligned to the corresponding target word. In this case, the best alignment searching sometimes gets the local solution. This problem could be resolved by considering local dependencies for ambiguous words.

One difficulty is how to handle function words. Function words often do not have exactly corresponding words in the opposite language. Japanese case markers such as "(ha)", "(ga)" (subjective case), "(wo)" (objective case) and so on, and English articles are typical examples of words, that do not have corresponding parts. There is a difference between alignment criteria for function words of gold-standard and our outputs, and it is somewhat difficult to improve alignment accuracy.

6 Conclusion

In this paper, we have proposed a linguistically-motivated probabilistic phrase alignment model based on dependency tree structures. The model incorporates the tree-based reordering model. Experimental results show that the word sequential model does not work well for linguistically different language pairs, and this can be resolved by using syntactic information. We have conducted the experiments only on Japanese-English corpora. To firmly support our claim that syntactic information is important, it is necessary to do more investigation on other language pairs.

Most frequent alignment errors are derived from parsing errors. Because our method depends heavily on structural information, parsing errors easily make the alignment accuracy worse. Although the parsing accuracy is high in general for both Japanese and English, it sometimes outputs incorrect dependency structures because technical or unknown words often appears in scientific papers. This problem could be resolved by introducing parsing probabilities into our model using parsing tools which can output n-best parsing with their parsing probabilities. This will not only improve the alignment accuracy, it will allow revision of the parsing result. Moreover, we need to investigate the contribution of our alignment result to the translation quality.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Association for Computational Linguistics*, 19(2):263–312.

Colin Cherry and Dekang Lin. 2003. A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics*, pages 88–95.

Brooke Cowan, Ivona Kučerová, and Michael Collins. 2006. A discriminative model for tree-to-tree translation. In *Proceedings of the 2006 Conference on EMNLP*, pages 232–241, Sydney, Australia, July. Association for Computational Linguistics.

Daniel Gildea. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting on ACL*, pages 80–87.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 176–183, New York City, USA, June. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL 2003: Main Proceedings*, pages 127–133.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of The International Workshop on Sharable Natural Language*, pages 22–28.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA, June. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Arul Menezes and Stephen D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL) Workshop on Data-Driven Machine Translation*, pages 39–46.

Toshiaki Nakazawa and Sadao Kurohashi. 2008. Linguistically-motivated tree-based probabilistic phrase alignment. In *In Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA2008)*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Association for Computational Linguistics*, 29(1):19–51.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271–279.

Masao Utiyama and Hitoshi Isahara. 2007. A japanese-english patent parallel corpus. In *MT summit XI*, pages 475–482.

Hideo Watanabe, Sadao Kurohashi, and Eiji Aramaki. 2000. Finding structural correspondences from bilingual parsed corpus for corpus-based translation. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 906–912.

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