Bayesian Subtree Alignment Model based on Dependency Trees

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

Word sequential alignment models work well for similar language pairs, but they are quite inadequate for distant language pairs. It is difficult to align words or phrases of distant languages with high accuracy without structural information of the sentences. In this paper, we pro- pose a Bayesian subtree alignment model that incorporates dependency relations be- tween subtrees in dependency tree struc- tures on both sides. The dependency re- lation model is a kind of tree-based re- ordering model, and can handle non-local reorderings, which sequential word-based models often cannot handle properly. The model is also capable of handling multi-level structures, making it possible to find many-to-many correspondences automatically without any heuristic rules. The size of the structures is controlled by non-parametric Bayesian priors. Experimental alignment results show that our model achieves 3.5 points better alignment error rate for English-Japanese than the word sequential alignment model, thereby verifying that the use of dependency information is effective for structurally different language pairs.

1 Introduction

Alignment accuracy is crucial for providing high quality corpus-based machine translation systems because translation knowledge is acquired from an aligned training corpus. For similar language pairs, alignment accuracy is high, and the state-of-the-art word alignment tool GIZA++ has a smaller than 10% alignment error rate (AER) for French-English. GIZA++ is an implementation of the alignment models called the IBM models (Brown et al., 1993), which handle sentences as sequences of words, usually followed by some heuristic symmetrization rules to combine the alignment results in both directions. Since the accuracy is to some extent good for some language pairs, many researchers focus not on alignment, but on translation with more linguistic information incorporated in the models. Phrase-based SMT (Koehn et al., 2003) uses units larger than words, whereas Hiero (Chiang, 2005) used a kind of sentence structure. Various other studies tried incorporating additional linguistic knowledge such as syntactic trees (Chiang, 2010), dependency trees (Menezes and Quirk, 2008), or packed-forests (Tu et al., 2010) in their translation models. Note that all these works are based on the word sequential alignment models.

However, for distant language pairs such as English-Japanese or Chinese-Japanese, the word sequential model is quite inadequate (about 20 to 30 % AER), and therefore it is important to improve the alignment accuracy itself. The differences between languages can be seen in Figure 1, which shows an example of English-Japanese. The word or phrase order is quite different for these languages. Another important point is that there are often many-to-one or many-to-many correspondences. For example, the Japanese noun phrase "受 光 素子" is composed of three words, whereas the corresponding English phrase consists of only one word "photodetector", and the English function word "for" corresponds to two Japanese function words "に は". In addition, there are basically no counterparts for the English articles (a, an, the). Figure 2 shows the alignment results from bi-directional GIZA++ together with a combination heuristic called grow-diag-final-and[1] for the same sentence pair given in Figure 1. The system failed to align some words in the Japanese noun

[1]This is trained on the same corpus used in Section 4.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 794–802,
Chiang Mai, Thailand, November 8 – 13, 2011.

phrase, and incorrectly aligned "the ↔ は ". The word sequential model is prone to many such errors even for short simple sentences of a distant language pair.

Even if the word order differs greatly between languages, phrase dependencies tend to hold between languages. This is also true in Figure 1. Therefore, incorporating dependency analysis into the alignment model is useful for distant language pairs. Cherry and Lin (2003) proposed a model that uses a source side dependency tree structure and constructs a discriminative model. However, the drawback is that the alignment unit is the word, and thus, it can only find one-to-one alignments. The capability of generating many-to-many correspondences is also important because one or more words often correspond to more than one word on the other side.

Nakazawa and Kurohashi (2009) also proposed a model focusing on dependency relations. They modeled phrase dependency relations in dependency trees on both sides. The model is also capable of estimating many-to-many correspondences automatically without any heuristics through maximum likelihood estimation. One serious drawback of their model is that it tends to acquire incorrect larger subtrees. For models that can handle multiple levels (or sizes) of structures, larger structures always defeat smaller ones in maximum likelihood estimation, and the best solution is to align one sentence as a structure with the other for all sentence pairs. Although Denero et al. (2008) solved this degeneracy by placing a Dirichlet process prior over the parameters that can control the size of phrases properly, their Bayesian model again only handles sentences as sequences of words. In this paper, we take advantage of the two studies by Nakazawa et al. and Denero et al., and propose a Bayesian subtree alignment model based on dependency trees to improve alignment accuracy for distant language pairs.

2 Dependency Tree-based Alignment Model

Our model is an extension of the one proposed by Denero et al. (2008). Two main drawbacks of the previous model are the lack of structural information and a naive distortion model. For similar language pairs such as French-English (Marcu and Wong, 2002) or Spanish-English (DeNero et al., 2008), even a simple model that handles sentences as a sequence of words works adequately. This does not hold for distant language pairs such as English-Japanese or Chinese-Japanese, in which word orders differ greatly. We incorporate dependency relations of words into the alignment model and define the reorderings on the word dependency trees. Figure 1 shows an example of the dependency trees for Japanese and English.

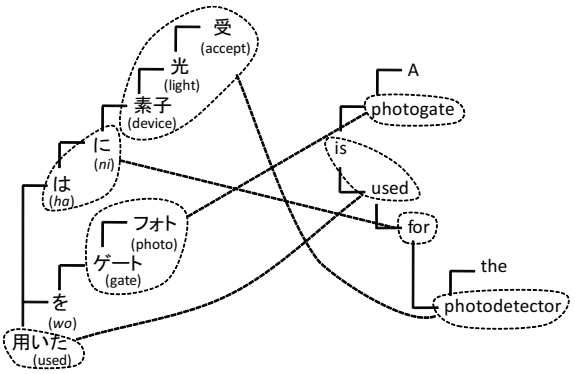

Figure 1: Example of dependency trees and alignment of subtrees. The root of the tree is placed at the extreme left and words are placed from top to bottom.

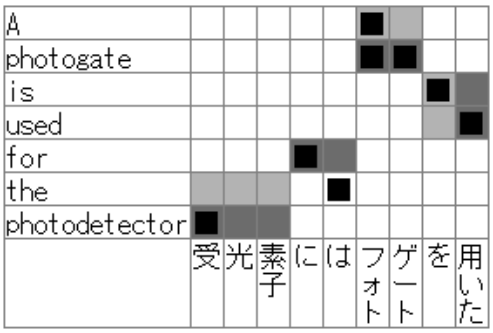

Figure 2: Alignment results from bi-directional GIZA++. Black boxes depict the system output, while dark (Sure) and light (Possible) gray cells denote gold-standard alignments.

2.1 Generative Story Description

Similar to the previous works (Marcu and Wong, 2002; DeNero et al., 2008), we first describe the generative story for the joint alignment model.

1. Generate ℓ concepts from which subtree pairs are generated independently.

2. Combine the subtrees in each language so as to create parallel sentences.

Here, subtrees are equivalent to phrases in the previous works. One subtree in a concept can be NULL, which represents an unaligned subtree. We restrict the unaligned subtrees to be composed of exactly one word, because of our model simplicity (NULL-alignment restriction).

The number of concepts ℓ is parameterized using a geometric distribution:

$$P(\ell) = p_c \cdot (1 - p_c)^{\ell - 1}. \quad (1)$$

Each concept c_i generates a subtree pair $\langle e_i, f_i \rangle$ from an unknown distribution θ_T, and then they are combined in each language. We denote the combinations of subtrees in English as $D_E = \{(j \rightarrow k)\}$, where $(j \rightarrow k)$ denotes that subtree e_j depends on subtree e_k, and in the foreign language as D_F. D refers to D_E and D_F as a whole.

With these notations, the joint probability for a sentence pair is defined as:

$$P(\{\langle e, f \rangle\}, D) = P(\ell) \cdot P(D|\{\langle e, f \rangle\}) \cdot \prod_{\langle e, f \rangle} \theta_T(\langle e, f \rangle). \quad (2)$$

2.2 Subtree Generation

When generating subtrees, we first decide whether to generate an unaligned subtree (with probability p_ϕ) or an aligned subtree pair (with probability $1 - p_\phi$). DeNero et al. (2008) used $p_\phi = 10^{-10}$ to strongly discourage NULL alignment, but this is not reasonable for some language pairs. Taking Japanese and English as an example, English determiners (a, an, the) and Japanese case markers (*ha*, *ga*, *wo*, etc.) rarely have counterparts. In addition, if the corpus is less clean and sentence pairs often contain a different amount of information, the strict restriction may lead to alignment errors. Therefore, we use $p_\phi = 0.33$.

Aligned subtree pairs are generated from an unknown probability distribution θ_A, which obeys the Dirichlet process (DP):

$$\theta_A(\langle e, f \rangle) \sim DP(M_A, \alpha_A), \quad (3)$$

where M_A is the base distribution and α_A is a concentration parameter. The base distribution is defined as:

$$\begin{aligned} M_A(\langle e, f \rangle) &= [P_f(f) P_{WA}(e|f) \cdot P_e(e) P_{WA}(f|e)]^{\frac{1}{2}} \\ P_f(f) &= p_t \cdot (1 - p_t)^{|f|-1} \cdot \left(\tfrac{1}{n_f}\right)^{|f|} \\ P_e(e) &= p_t \cdot (1 - p_t)^{|e|-1} \cdot \left(\tfrac{1}{n_e}\right)^{|e|}, \end{aligned} \quad (4)$$

where P_{WA} is the IBM model1 likelihood (Brown et al., 1993), and n_f and n_e are the numbers of word types in each language. θ_A gives a non-zero weight to aligned subtree pairs only.

Unaligned subtrees are generated from another unknown probability distribution θ_N:

$$\begin{aligned} \theta_N(\langle e, f \rangle) &\sim DP(M_N, \alpha_N) \\ M_N(\langle e, f \rangle) &= \begin{cases} P_{WA}(e|\text{NULL}) & \text{if } f = \text{NULL} \\ P_{WA}(f|\text{NULL}) & \text{if } e = \text{NULL} \end{cases}. \end{aligned} \quad (5)$$

θ_N gives a non-zero weight to unaligned subtrees only. Note that unaligned subtrees are always composed of only one word in our model. Finally, θ_T can be decomposed as:

$$\theta_T(\langle e, f \rangle) = p_\phi \theta_N(\langle e, f \rangle) + (1 - p_\phi) \theta_A(\langle e, f \rangle). \quad (6)$$

2.3 Dependency Relation Probability

Instead of the naive reordering model in the previous work, our model considers dependency relations between subtrees and assigns a weight to each relation. Suppose subtree f_j depends on subtree f_k (parent subtree), which means $(j \rightarrow k) \in D_F$, and both f_j and f_k are aligned subtrees. Their counterparts, e_j and e_k respectively, are somewhere on the dependency tree of the other side. We can assume that e_j tends to depend on e_k because the dependencies between concepts hold across languages. The dependency relation probability reflects this tendency.

Formally, we extract a tuple $(N(f_j), rel(f_j, f_{j'}))$ for subtree f_j, and assign the dependency relation probability to that tuple. For unaligned subtrees, the dependency relation probability is not taken into consideration. If the parent subtree is an unaligned subtree, we ascend the dependency tree to the root node until an aligned subtree is found. We call the nearest aligned subtree a *pseudo parent*. The pseudo parent for subtree f_j is denoted as $f_{j'}$, and the number of unaligned subtrees from f_j to $f_{j'}$ is denoted as $N(f_j)$. We consider an imaginary root node as a pseudo parent for the root subtree. For example, the parent subtree of "フォト ゲート (photogate)" is "を (accusative)" which is unaligned in Figure 1. The pseudo parent is "用いた (used)" and the number of unaligned subtrees $N = 1$. Japanese function words are often unaligned, similar to this example, but the dependency relations between subtrees stepping over the function words are assumed to hold on

the other side. Therefore we introduce a pseudo parent to capture the relations.

Function $rel(f_j, f_{j'}))$ returns a dependency relation between the counterparts of the two arguments. Note that the counterparts of f_j and $f_{j'}$ are e_j and $e_{j'}$, respectively. We express a dependency relation as the shortest path from one subtree to another. For simplicity, we indicate the path with a pair of non-negative integers, where the first is the number of steps going up (Up) the dependency tree and the other is the number going down ($Down$). It also requires one additional step for going through unaligned subtrees. For example, in Figure 1, traveling from "A photogate" to "photodetector" requires 1 step going up (to reach "is used") and 2 steps going down (via "for"), so the dependency relation is $(Up, Down) = (1, 2)$. Consequently, the tuple is represented as a triplet of non-negative integers $R_f = (N, Up, Down)$.

The dependency relation probabilities for the foreign language side are drawn from an unknown probability distribution θ_{fe} and for the English side from θ_{ef}, with both obeying the DP:

$$\begin{aligned} \theta_{fe}(R_f) &\sim DP(M_{fe}, \alpha_{fe}) \\ M_{fe}(R_f) &= p_{fe} \cdot (1 - p_{fe})^{N+Up+Down-1} \\ \theta_{ef}(R_e) &\sim DP(M_{ef}, \alpha_{ef}) \\ M_{ef}(R_e) &= p_{ef} \cdot (1 - p_{ef})^{N+Up+Down-1}. \end{aligned} \quad (7)$$

Using the notations and definitions above, the dependency tree-based reordering model $P(D|\{\langle e, f\rangle\})$ is decomposed as:

$$P(D|\{\langle e, f\rangle\}) = \prod_{\langle e,f\rangle} \theta_{fe}(R_f) \cdot \theta_{ef}(R_e). \quad (8)$$

3 Model Training

We train the model by means of a collapsed Gibbs sampling, which has been used in some recent NLP works (DeNero et al., 2008). In a Gibbs sampling, we first need to initialize the states of the training data, such as the boundaries between subtrees and their alignments, and also initialize the latent variables according to the initial states of the data. Starting with the initial state, we generate many samples in order from the last state by changing a small local point. Normalizing the counts in the samples yields the parameter estimations.

3.1 Initialization

We initialize the states of the training data by heuristically merging bi-directional alignment results of the standard word alignment tool GIZA++. Many machine translation studies use heuristics to combine the two alignment results, one of which is called grow-diag-final-and (Koehn et al., 2007). Our heuristic is similar to this, but the difference is that we combine the two results based on dependency trees, and not on word sequences. The initialization is carried out by the following steps:

1. Take the intersection of the two results.
2. Add alignment points connected to at least one accepted point in terms of the dependency tree (corresponds to grow-diag).
3. Add alignment points between two unaligned words (corresponds to final-and).

Initial boundaries of subtrees and their alignments, and also the counts of subtree pairs and dependency relations are thus acquired.

3.2 Sampling Operators

Our sampler uses three operators repeatedly to generate samples. The operators are illustrated in Figure 3. A solid circle represents a single word, while a subtree is depicted as the part surrounded by a dotted line. Alignment links between subtrees are represented by broken lines.

Each application of an operator generates one new sample, and of course we could use all the generated samples. However, successive samples are almost the same, except for one local part. It is no use keeping all the samples, so we keep only one sample, which is the final outcome after applying all the operators to all the possible points in all the sentence pairs in the training corpus.

SWAP

The SWAP operator exchanges the counterparts of two subtrees with each other. Boundaries between subtrees and the number of aligned subtrees in the sentence pair are fixed. There are two cases for SWAP:

1. Both of the subtrees are aligned (SWAP-1).
2. One of the two subtrees is unaligned and the other is composed of only one word (SWAP-2).

An illustration of the result of the SWAP operator is shown at the top of Figure 3.

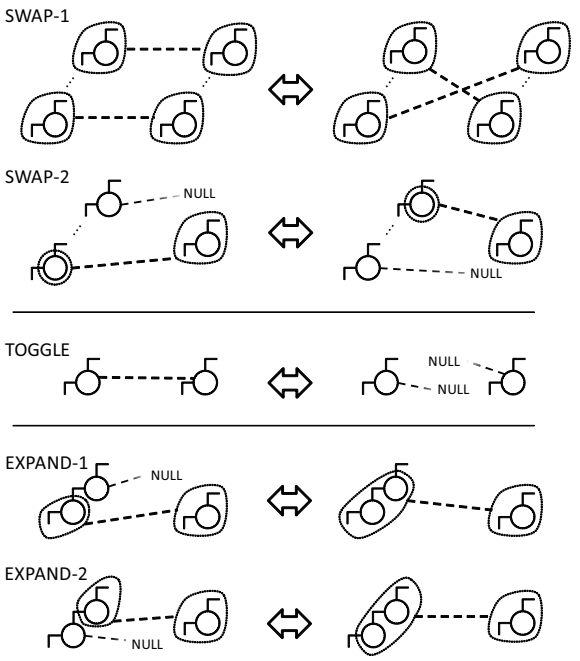

Figure 3: Illustration of the operators.

TOGGLE

The TOGGLE operator adds or removes an alignment. If f_j and e_k are both unaligned subtrees, TOGGLE links these two subtrees. Alternatively, if f_j and e_j are aligned, TOGGLE cuts the link and makes each of the subtrees unaligned. Because of the NULL-alignment restriction, TOGGLE does nothing if f_j or e_j is composed of more than one word. Boundaries between subtrees are fixed. An illustration of the TOGGLE operation is shown in the middle of Figure 3.

EXPAND

The EXPAND operator expands or contracts an aligned subtree. If an unaligned subtree is next to an aligned one, EXPAND tries to merge the unaligned and aligned subtrees. It also tries to exclude a marginal node from a subtree, and to make the excluded node unaligned. There are two cases for EXPAND:

1. A node is added to a subtree as a new leaf node, or a leaf node of a subtree is excluded (EXPAND-1).

2. A node is added to a subtree as the new root node, or the root node of a subtree is excluded (EXPAND-2).

EXPAND-1 does not have any restrictions on its operation. However, for EXPAND-2, if the root node has more than one child node inside the subtree, it cannot exclude the root node, because the subtree will be divided into two subtrees by the exclusion, and it is impossible to return to the previous state. Operators in a Gibbs sampler must be able to return the same status by immediately reapplying the same operator to the same point. An illustration of the EXPAND operation is shown in Figure 3.

3.3 Computational Complexity and Parallel Sampling

The distortion model of the previous work does not consider any relations to neighboring phrases, so any operation does not affect the distortions of neighboring phrases. On the contrary, our proposed model considers the relations, and this leads to an increase in the computational complexity for one operation. For example, swapping two aligned subtree pairs requires re-calculation of the dependency relation probabilities for not only the four focused subtrees, but also subtrees whose pseudo parent is one of the focused subtrees. This is the same for the TOGGLE operation.

To make matters worse, the EXPAND operator requires much more. All the subtrees that traverse the locally changed subtree, 1) in finding a pseudo parent, and 2) in finding the dependency relations, need re-calculation of the dependency relation probabilities, because the number of steps (N, Up and $Down$) will change.

This increase in computational complexity makes the training time much longer, so that it is impossible to train using a single CPU. To alleviate this problem somewhat, we divide the training data into several parts and execute sampling in parallel. The overall flow of model training is summarized as follows:

1. Initialize the training corpus and current counts of subtree pairs and dependency relations, and divide the corpus into several sections.

2. Start sampling in parallel using the same current counts, and generate one sample from each section. A sample is the final state of the section.

3. Gather samples and update the current counts by counting subtree pairs and dependency relations in the samples.

4. Go back to step 2.

4 Alignment Experiments

We conducted alignment experiments on English-Japanese and Chinese-Japanese corpora to show the effectiveness of the proposed model.

4.1 Settings

For English-Japanese, the JST[2] paper abstract corpus was used. This corpus was created by NICT[3] from JST's 2M English-Japanese paper abstract corpus using the method of Utiyama and Isahara (2007). For Chinese-Japanese, we used the paper abstract corpus provided by JST and NICT. This corpus was developed during a project in Japan called the "Development and Research of Chinese-Japanese Natural Language Processing Technology". The statistics of these corpora are shown in Table 4.1. Unfortunately, these two corpora are not freely available now, but they will become available to everyone in near future.

As gold-standard data, we used 479 sentence pairs of English-Japanese and 510 sentence pairs of Chinese-Japanese. These were annotated by hand using two types of annotations: sure (S) alignments and possible (P) alignments (Och and Ney, 2003). The unit of evaluation was the word for all the languages. We used precision, recall, and alignment error rate (AER) as evaluation criteria. All the experiments were run on the original forms of words. The hyper parameters for our model used in the experiments are summarized in Table 4.1.

Japanese sentences were converted into dependency structures using the morphological analyzer JUMAN (Kurohashi et al., 1994), and the dependency analyzer KNP (Kawahara and Kurohashi, 2006). For English sentences, Charniak's nlparser was used to convert them into phrase structures (Charniak and Johnson, 2005), and then they were transformed into dependency structures by rules defining head words for phrases. Chinese sentences were converted into dependency trees using the word segmentation and POS-tagging tool by Canasai et al. (2009) and the dependency analyzer CNP (Chen et al., 2008).

For comparison, we used GIZA++ (Och and Ney, 2003) which implements the prominent sequential word-based statistical alignment model

[2]http://www.jst.go.jp/

[3]http://www.nict.go.jp/

	En-Ja		Zh-Ja	
	En	Ja	Zh	Ja
# of sentences	996K		680K	
# of words	24.7M	27.5M	18.8M	22.3M
ave. sent. length	24.8	27.6	27.7	32.9

Table 1: Statistics of the training corpus.

	α_A	α_N	p_t	α_{fe}, α_{ef}	p_{fe}, p_{ef}
En-Ja	100	100	0.8	100	0.5
Zh-Ja	10	100	0.8	100	0.5

Table 2: Hyper parameters used in experiments.

of the IBM Models. We conducted word alignment bidirectionally with the default parameters and merged them using the grow-diag-final-and heuristic (Koehn et al., 2003). Furthermore, we used the BerkeleyAligner[4] (DeNero and Klein, 2007) with default settings for unsupervised training. Experimental results for English-Japanese are shown in Table 4.1, and those for Chinese-Japanese are shown in Table 4.1. The alignment accuracy of the initialization described in Section 3.1 is indicated as "Initialization", while the accuracy after conducting Gibbs sampling is indicated as "Proposed".

4.2 Discussion

For English-Japanese, our proposed model achieved reasonably high alignment accuracy compared with that of GIZA++ and the BerkeleyAligner. Using tree structures combined with the bi-directional alignment results leads to better accuracy than the original sequential heuristic (indicated as Initialization). We also give the alignment accuracy obtained by Nakazawa and Kurohashi (2009) (indicated as Nakazawa+ in Table 4.1, and applicable only to the English-Japanese corpus)[5]. Their model suffered from a degeneracy in acquiring larger phrases caused by maximum likelihood estimation, and this led to low precision and high recall. Compared with their model, our proposed model overcomes the degeneracy and outperforms in terms of both precision and recall. Figure 4 shows an example of the improvement in alignment. GIZA++ aligned function words incorrectly because of the lack of structural information. For example, GIZA++ aligned the English "of" and the Japanese "の"

[4]http://code.google.com/p/berkeleyaligner/

[5]They used 475 sentence pairs instead of 479 sentence pairs of ours. The difference comes from the inconsistency of word segmentations of Japanese, but it is negligibly small.

	Pre.	Rec.	AER
grow-diag-final-and	81.17	62.19	29.25
BerkeleyAligner	85.00	53.82	33.72
Nakazawa+	80.28	63.85	28.67
Initialization	82.39	61.82	28.99
Proposed	**85.93**	**64.71**	**25.73**

Table 3: Results of English-Japanese alignment experiments.

	Pre.	Rec.	AER
grow-diag-final-and	83.77	75.38	20.39
BerkeleyAligner	**88.43**	69.77	21.60
Initialization	84.71	**75.46**	19.90
Proposed	85.52	74.71	**19.89**

Table 4: Results of Chinese-Japanese alignment experiments.

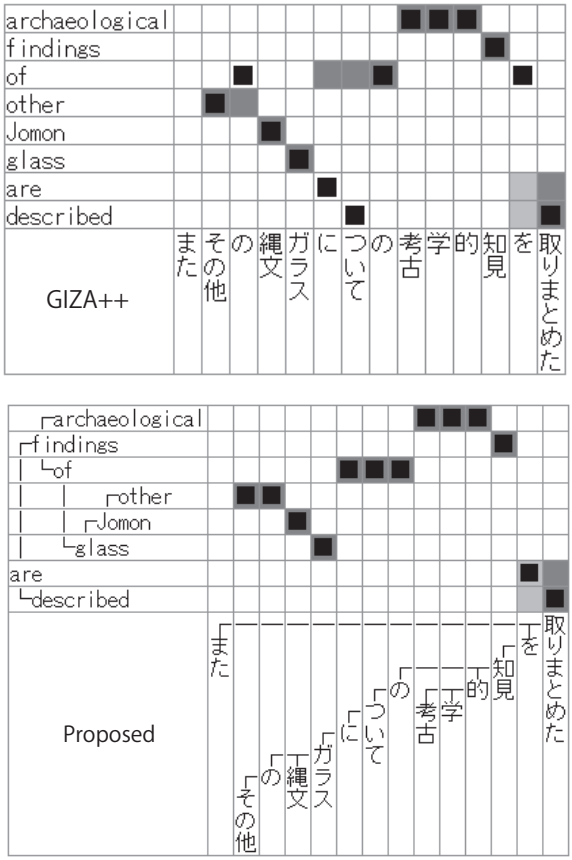

Figure 4: Alignment results from bi-directional GIZA++ (top) and proposed model (bottom).

in the third word. This was incorrectly derived from the combination heuristic: English posterior word "other" and Japanese prior word "その他" are aligned by intersection, and thus, "of ↔ の" is also aligned through the grow-diag heuristic. This could be avoided by introducing dependency trees: the English words "of" and "other" are not contiguous in the dependency tree. In addition, there is no correspondence between "に ついて の ↔ of". This is a rare translation for "of", which is most frequently translated as "の", so GIZA++ aligned "の ↔ of" only. A nonparametric Bayesian model is capable of finding the correct alignment with the support of occurrences of the subtree pair elsewhere in the training corpus.

The reasons for recall being significantly lower than precision in all the models are summarized in the two points. One is the separation between gold-standard criteria and system output. In Figure 4, "are described" and "を 取りまとめた" are aligned in their entirety, but the system found one-to-one alignments, which are also acceptable. The other is that it is sometimes hard to align part of a sentence in a smaller unit for distant language pairs, because the expressions are quite different. In such cases, the system is obliged to align expressions enmasse, which leads to low recall.

For Chinese-Japanese, we failed to improve alignment accuracy greatly. A major cause of this undesirable result could be the accuracy of the Chinese parser. Both the English and Japanese parsers used in the experiments can analyze sentences with over 90% accuracy, whereas the accuracy of the Chinese parser is less than 80% despite it being state-of-the-art in the world (Chen et al., 2008). The parsing accuracy reported in this paper was obtained from an experiment using gold-standard word segmentation and POS-tags. Starting with raw sentences results in about 77.4% accuracy. This information was obtained from communication with the authors. Fundamental NLP technologies in each language must be improved in the long term, and sophisticated models, which use deeper analysis of sentences like our model, should become effective in the near future. One possible short-term solution for the parsing problem is to use the n-best parsing results in the model. Another kind of solution was proposed by Burkett et al. (2010), who described a joint parsing and alignment model that can exchange useful information between the parser and aligner.

However, even in the case of Chinese-Japanese alignment, we achieved higher precision than GIZA++. For translation, precision is more im-

	BLEU
grow-diag-final-and	24.59
Initialization	**25.33**
Proposed	24.50

Table 5: BLEU results for Japanese-to-English translation experiments.

portant than recall. The BerkeleyAligner showed much higher precision than GIZA++ and, in Chinese-Japanese alignment, than our model as well. However, recall was quite low compared with all the other models. Lower recall results in a large translation table because the phrase extraction heuristics 'grow' over each unaligned word.

5 Translation Experiments

We conducted Japanese-to-English translation experiments on the same corpus used in the alignment experiment. We translated 500 paper abstract sentences from the JST corpus. Note that these sentences were not included in the training corpus. We used the state-of-the-art phrase-based SMT toolkit Moses(Koehn et al., 2007) with default options, except for the distortion limit ($6 \rightarrow 20$). It was tuned by MERT using another 500 development sentence pairs.

Table 5 shows the BLEU scores for the translations. Although the difference in alignment accuracy between grow-diag-final-and and Initialization is small, the BLEU score was improved by 0.74 point. This is because syntactic information reduced incorrect alignment points and the quality of translation table becomes better. This result provided evidence that syntactic knowledge is useful for distant language pairs. However, the BLEU score decreased after iterations of our alignment model. The main reason is that the alignment result of our model is not compatible with Phrase-based SMT. Our model often output sequentially discontiguous alignments which are harmful for PSMT to create fine-grained phrase table. We need to use other decoders which is compatible with our alignment model (Nakazawa and Kurohashi, 2010), and we believe that our model leads to better translation quality.

6 Conclusion

In this paper, we proposed a linguistically-motivated nonparametric Bayesian subtree alignment model based on dependency tree structures for distant language pairs. The model incorporates the tree-based reordering model. It also solves the degeneracy of the maximum likelihood estimation for models capable of handling multiple levels of structures by placing a Dirichlet process prior over parameters. Experimental results show that a word sequential model does not work well for distant language pairs, but that this can be addressed by using syntactic information. Our proposed model achieved a lower AER by about 3.5 points compared with GIZA++.

To support allegation that syntactic information is important for distant language pairs, it is necessary to compare our model with the original word sequential study (DeNero et al., 2008), which was consulted often. It is also important to apply our model not only to other distant language pairs, but also to similar language pairs, and to investigate the results. We are planning to use standard data set such as NIST or IWSLT. Also, we could use the n-best parsing results in our model to alleviate the error propagation from parsing, especially for Chinese.

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.

David Burkett, John Blitzer, and Dan Klein. 2010. Joint parsing and alignment with weakly synchronized grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–135, Los Angeles, California, June. Association for Computational Linguistics.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180.

Wenliang Chen, Daisuke Kawahara, Kiyotaka Uchimoto, Yujie Zhang, and Hitoshi Isahara. 2008. Dependency parsing with short dependency relations in unlabeled data. In *In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08)*, pages 88–94.

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.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270.

David Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1443–1452, Uppsala, Sweden, July. Association for Computational Linguistics.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

John DeNero, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 314–323, Honolulu, Hawaii, October. Association for Computational Linguistics.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*.

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An error-driven word-character hybrid model for joint chinese word segmentation and pos tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 513–521, Suntec, Singapore, August. Association for Computational Linguistics.

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.

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

Arul Menezes and Chris Quirk. 2008. Syntactic models for structural word insertion and deletion during translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 734–743, Honolulu, Hawaii, October. Association for Computational Linguistics.

Toshiaki Nakazawa and Sadao Kurohashi. 2009. Statistical phrase alignment model using dependency relation probability. In *In Proceedings of the third Workshop on Syntax and Structure in Statistical Translation (SSST-3)*, pages 10–18, Boulder, Colorado, June.

Toshiaki Nakazawa and Sadao Kurohashi. 2010. Fully syntactic ebmt system of kyoto team in ntcir-8. In *In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-8)*, pages 403–410.

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

Zhaopeng Tu, Yang Liu, Young-Sook Hwang, Qun Liu, and Shouxun Lin. 2010. Dependency forest for statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1092–1100, Beijing, China, August. Coling 2010 Organizing Committee.

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