Constraining a Generative Word Alignment Model with Discriminative Output

Chooi-Ling GOH†(a), Taro WATANABE†, Nonmembers, Hirofumi YAMAMOTO†, and Eiichiro SUMITA†, Members

SUMMARY We present a method to constrain a statistical generative word alignment model with the output from a discriminative model. The discriminative model is trained using a small set of hand-aligned data that ensures higher precision in alignment. On the other hand, the generative model improves the recall of alignment. By combining these two models, the alignment output becomes more suitable for use in developing a translation model for a phrase-based statistical machine translation (SMT) system. Our experimental results show that the joint alignment model improves the translation performance. The improvement in average of BLEU and METEOR scores is around 1.0–3.9 points.

key words: word alignment, discriminative model, generative model, hybrid, SMT

1. Introduction

From our previous research on word alignment, we realized that a generative model provides higher recall but lower precision, while a discriminative model provides higher precision but lower recall [1]. In the case of the discriminative model, better precision can be obtained either by using a better feature set or larger training data; however, the recall is lowered quite considerably if the training data is small. In other words, there exist many null links among the sentence pair. The existence of null links will cause problems in the creation of translation phrases for a translation model. This is because it will be possible to freely join unattached words with any neighboring words to form new phrases that are not desired. In this case, the combination of phrases increases and the phrase table generated becomes too large. Furthermore, the translation probability for each phrase will not be calculated correctly. As result, during decoding, the correct target translation phrases cannot be retrieved even when the source phrases exist in the phrase table.

In this study, we propose to maintain the high precision of a discriminative model, and to improve the recall by combining it with a generative model. First, a discriminative model using Conditional Random Fields (CRF) is trained with a small set of training data. Then, this model is used to align a large parallel corpus which will be used to train a translation model. The remaining words that are not aligned by CRF will be covered by a generative model (Model 1 and Hidden Markov Model (HMM)). Doing this will reduce the number of null links and consequently, the size of the phrase table generated will become smaller and more precise. Our experiment results on a phrase-based statistical machine translation (SMT) system show that the joint alignment model improves the translation performance. The improvement in the average of BLEU and METEOR scores is around 1.0–3.9 points.

2. Related Work

With the increase in available labeled data, recent research has investigated supervised word alignment using discriminative models [2]–[6]. Discriminative models allow the introduction of various features, either lexically, syntactically, or statistically during training. Previous results have shown that discriminative models outperformed generative models in terms of precision with comparable recall. However, supervised models need some manually aligned data for the training. Moreover, the accuracy may be too low if the training data is small; at the same time, it is extremely time- and resource-consuming to accumulate a sufficiently large amount of aligned data. While the labeled data is not sufficient, the alignment estimated from the generative models is used as part of the features in the discriminative models.

Some research has also been conducted on semi-supervised word alignment models [7]–[9]. These models used a small amount of hand-aligned data to estimate the parameters used in a generative model. The semi-supervised models helped to boost up the performance of the unsupervised models. In such cases, only a small amount of hand-aligned data is necessary and it is more reasonable to obtain. Our proposed method is most similar to [9] which allows the discriminative predictions to influence the generative model.

In [10], they tried to introduce some prior knowledge such as heuristics or linguistics features in statistical generative word alignments models. A constraint value is calculated using either the entropy principle or the bilingual latent semantic analysis. This value is an indicator of how strong the word pair should be considered as a candidate that serves as some soft constraints. In [11], a log linear model was proposed which applied the feature functions of IBM translation Model 3, part-of-speech tags and bilingual dictionaries. These researches have proved that providing some prior knowledge to a statistical alignment model could help improving the alignment accuracy. In our approach, instead of...
using the hand-aligned data to estimate the parameters in a generative model, we train a discriminative model with this training data and use the output of this model as a constraint or prior knowledge in a generative model. Since a discriminative model can ensure high precision, we can rely on its output and employ them as hard constraints in the generative model.

A related work can also be found in [12]. They tried to combine the output of a few alignment models using the maximum entropy (ME) approach. Their method uses ME to decide whether to include/exclude a particular alignment link based on some feature functions that are extracted from the input alignments and linguistic features of the words. Since their study involves supervised learning, they also needed a small set of annotated data for the learning. Our method differs from theirs with regard to the selection of alignment links from different models, because in our study, we add or delete alignment links using a generative model in addition to a high-precision model.

3. Proposed Method

We propose to combine a generative model with a discriminative model by constraining the generative model with the output from the discriminative model. The output from the discriminative model serves as some prior knowledge to the generative model. We use IBM Model 1 and an HMM as the generative models and a CRF model as the discriminative model. Although a bitext is sufficient to train the generative model, a small amount of hand-aligned data is needed to train the discriminative model. A more detailed description is provided below.

3.1 Model 1 and HMM

Here, we apply only IBM Model 1 and an HMM model[13,14] that include the dependence of word classes. The training is similar to the one in [15], where alignment agreement is found between the models from two translation directions. Both models can be generalized in the following equations:

\[ p(e|f) = \sum_a p(a, f|e) \]  

where \( e = (e_1, \ldots, e_t) \) is the target sentence and \( f = (f_1, \ldots, f_t) \) is the source sentence.

\[ p(a, f|e) = \prod_{j=1}^t p_d(a_j|a_{j-1}, j, F(f_j), E(e_{a_j})), E(e_{a_{j-1}}) p(f_j|a_{j-1}) \]  

where \( j- \) is the position of the last non-null-aligned source word before position \( j \) and \( F(\cdot) \) and \( E(\cdot) \) are the word classes. Further, the distortion model is as below:

\[ p_d(a_j | a_{j-1} = i, F, E) = p_0 \]

\[ p_d(a_j | i' = a_{j-1} = i, F, E) = (1 - p_0) \cdot \frac{1}{c(i' - i, F, E)} \]  

(Model 1)  

\[ \text{HMM} \]

where \( p_0 \) is the null-word probability and \( c(\cdot) \) is the distortion parameters. The null-word probability is set to a constant of \( p_0 = 0.05 \).

Then, an EM algorithm is used to maximize an objective function that incorporates both data likelihood and a measure of agreement between models from two directions [15].

\[
\max \sum_{e,f} \left[ \log p(e \rightarrow f) + \log p(f \rightarrow e) \right] + \log \sum_a p(a|e \rightarrow f)p(a|f \rightarrow e) 
\]

In the E-step, the agreed lexical count is computed for \( e, f \):

\[
c(f, e; f, e) = \sum_a p(a|f, e) \sum_{i,j} \delta(f_i, f_j) \delta(e_i, e_i) \]  

(3)

where \( \delta(\cdot) \) is a Kronecker delta function which is 1 if the arguments are equal, and 0 otherwise. These collected counts are marginalized in each direction for the M-step. We name this generative model as the NICT™ model.

3.2 CRF Model

In the second model, we apply CRF for discriminative training. Word alignment is treated as a sequential labeling problem. Each pair of words is assigned with some features and labeled as either “align” or “not aligned”.

A linear-chain CRF with parameters \( \Lambda = \{\lambda_1, \ldots, \lambda_K\} \) defines a conditional probability for a label sequence \( y = y_1 \ldots y_T \) given an input sequence \( x = x_1 \ldots x_T \) to be:

\[
P_{\Lambda}(y|x) = \frac{1}{Z_x} \exp \left( \sum_{t=1}^T \sum_{k=1}^K \lambda_k f_k(y_{t-1}, y_t, x) \right)
\]

where \( Z_x \) is the normalization factor that makes the probability of all state sequences sum to one; \( f_k(y_{t-1}, y_t, x) \) is a feature function, and \( \lambda_k \) is a learned weight associated with feature \( f_k \). We use a public training tool CRF++\(^{11}\), that is easy and fast for training and decoding.

The CRF alignment model is similar to the one proposed in [1]. In order to train the CRF model, we must prepare a feature set. The features are chosen such that they will provide certain clues for the alignments. CRF allows the use of arbitrary and overlapping features. Hence, we are free to introduce any possible features such as syntactical, lexical, and contextual features.

3.2.1 Dice Coefficient

The first feature is the Dice coefficient, which is an estimation of the closeness of two words. The word association is calculated using sentence aligned corpus.

\(^{11}\)http://crfpp.sourceforge.net/

---

GOH et al.: CONSTRAINING A GENERATIVE WORD ALIGNMENT MODEL WITH DISCRIMINATIVE OUTPUT

---

1977
Here $C_E$ and $C_F$ represent the number of occurrences of the words $e$ and $f$ in the corpus while $C_{EF}$ represents the number of co-occurrences. A high (low) value indicates that the word pair is closely (loosely) related to each other.

### 3.2.2 Bilingual Dictionary

The second measurement parameter for the two words can be a bilingual dictionary. If the pair of words exists in the same entry in the dictionary, there is a high possibility that they can be aligned together. However, many words belonging to one language are not always translated into one single word in the other language. A word in a source language can be translated into a compound word in the other language and vice versa. This is especially true for translations between languages that are fairly different syntactically.

Therefore, the similarity between the two words is calculated as follows:

$$\text{Bi-dic} = \text{Sim}(e, E) = \max(Sim(e, e_i) = \frac{1}{|e_i|} \text{ if } e \in e_i \text{ and } e_i \in E \text{ else } 0)$$

Assume that the word pair that we consider for alignment is $(c, e)$. Then, we search for the translation for $c$ in the dictionary. There may exist multiple translations for $c$, i.e., $E$. We compare $e$ and $E$ as given in the equation above. For each translation $e_i$ in $E$, if there is a one-to-one match, that is, if $e = e_i$, then the score is 1; else, the score is $1/N$ if word $e$ exists in $e_i$ where $N$ is the number of words in the translation $e_i$; else, the score is 0. If the word $e$ matches a few translations, we only take the maximum value.

### 3.2.3 Relative Sentence Position

$$\text{Relpos} = \text{abs} \left( \frac{a_i}{|e|} - \frac{t}{|f|} \right)$$

where $a_i$ is the position of the aligned source word in $e$, and $t$ is the position of the target word in $f$.

The relative sentence position allows the model to learn the preferences for aligning words that are close to the alignment matrix diagonal. This feature is useful when source and target sentences share similar grammar structures.

### 3.2.4 POS Tags

In order to reduce the sparseness of the lexical words, part-of-speech tags for both languages are used as features. For languages without word boundaries like Chinese and Japanese, word segmentation must be carried out before POS tagging could take place.

### 3.2.5 Lemmatization

Some languages are inflectional, like English and Japanese, while some are not, like Chinese. A different form of words in inflectional language can be aligned to the same form in non-inflectional language. In order to reduce such sparsity, the lemma of the inflectional word is used. This is not necessary for non-inflectional language. With the matching of word lemmas, we hope that the alignment can be enhanced even further.

### 3.2.6 Contextual Features

In this study, we introduce a new set of contextual features that allow our learning to consider the competition between the adjacent words. Since our learning method is similar to a sequential labeling problem, the contexts can be the words and POS tags before and after current word pairs. Both source and target contexts are added as the features.

### 3.2.7 Multigram Features

Finally, the feature set consists of the following: Dice coefficient (Dice), relative position (relpos), bilingual dictionary similarity measure (Bi-dic), source word (S-word), target word (T-word), source lemma (S-lemma), target lemma (T-lemma), source part-of-speech tag (S-POS) and target part-of-speech tag (T-POS). For the context feature, we use the lemmas and part-of-speech tags of the words before and after the current word pair (S-lemma-1, S-POS-1, T-lemma-1, T-POS-1, S-lemma+1, S-POS+1, T-lemma+1, T-POS+1). To build these features, we will need a morphological analyzer to obtain the POS tags and lemmas of the words for each source and target language. Besides, a large parallel corpus and a bilingual dictionary are needed to calculate the Dice coefficient and dictionary similarity measure.

The combination of features that we applied in this implementation is listed below:

- unigram features (S-word, T-word, relpos, Dice, Bi-dic, S-lemma, S-POS, T-lemma, T-POS)
- bigram features (S-lemma/S-lemma, S-lemma/S-POS, T-lemma/T-POS, S-POS/T-POS)
- trigram features (S-lemma/T-lemma/relpos, S-word/T-lemma/Dice, S-lemma/T-lemma/Bi-dic)
- contextual features
  - unigram (S-lemma-1, S-POS-1, T-lemma-1, T-POS-1, S-lemma+1, S-POS+1, T-lemma+1, T-POS+1)
  - bigram (S-lemma-1/S-lemma, S-POS-1/S-POS, T-lemma-1/T-lemma, T-POS-1/T-POS, S-lemma/S-lemma+1, S-POS/S-POS+1, T-lemma/T-lemma+1, T-POS/T-POS+1)

### 3.3 Combination Model

Figure 1 provides an overview of the combination of two word alignment models. First we train a CRF alignment model using a small amount of training data as described in Sect.3.2. Then, we use this CRF model to align a large parallel corpus, which is used to develop the translation model. Next, the large parallel corpus, together with the
alignment output from the CRF model, is used in the generative model, as described in Sect. 3.1. Here, the alignment of the generative model is constrained by the output of the CRF model. The alignment points extracted from the CRF model ($A_{CRF}$) is used to constrain the E-step (Eq. 3) in the generative model as below.

$$c(f, e; f', e') = \sum_a p(a | f', e') \sum_{i,j} \delta(f, f_i) \delta(e, e_i) \times C(i, j)$$

(4)

where $C(i, j)$ is a constraint function calculated as below:

$$C(i, j) = \begin{cases} 0.0 & \text{if } (i, j) \in A_{CRF} \text{ and } (i, j) \notin A_{CRF} \\ 1.0 & \text{otherwise} \end{cases}$$

If there is a point $(i, j)$ where alignment has already been done by CRF model, then new alignments will not be considered again, meaning only $C(i, j) = 1.0$ but all others $C(i, j) = 0.0$. If there is no such a point $(i, j)$, then all $C(i, j) = 1.0$. Alignments will only be done on the area where $C(i, j) = 1.0$ in the generative model.

Figure 2 shows an illustration of the alignment process. On the left-hand side, the “x” $(x_i, x_j)$ indicates some alignment points produced by the CRF model ($x_i, x_j \in A_{CRF}$). There exist some null links or words that have not been aligned. These null links form the gray areas. The generative model is allowed to align only in these areas. We carry out the alignment for both translation directions and finally combine the output of both directions using the grow-diag-final-and heuristics method [14] which is a convention used in GIZA++.

As shown in the right-hand side of Fig. 2, the generative model may add more alignment points; however, it can also remove some conflict alignment points. The changes are marked in gray color. Finally, the Moses† toolkit is used to build a translation model on the basis of this alignment output.

### 3.4 Training Data Selection Algorithm

Currently we have large Chinese-English and Japanese-English manually aligned corpora at our disposal. We would like to select only a small set, says 1 K, from each of them for use in the training of our CRF model. However, the method of selection of a representative set is a problem. We would like to obtain a set that ensures optimal training of the CRF model. In other words, we think that the greater the diversity of the data, the more effective the model can

†http://www.statmt.org/moses/
be. Therefore, we want to include as many different word pairs as possible in the selection.

The following algorithm shows the steps for selection. The \( \text{count} \) is the size of the small dataset desired.

1. From all pairs, select only candidates where length \( \geq \text{minlen} \) and \( \leq \text{maxlen} \).
2. Sort candidates by length in reverse order.
3. Set threshold = predefined and total = 0.
4. While total < count,
   a. get the next candidate;
   b. if the candidate has less than the threshold of similar words that exist in previously selected candidates, then include this candidate and increase count by one, else discard.
5. If no more candidates and total < count, then increase threshold by 10% and repeat step 4)

First, the candidates are limited to sentence pairs that have a certain range of sentence length. Too short sentences do not provide adequate information, while too long sentences create complication in alignment. However, longer sentences are preferred hereafter. A word list is built on the basis of the sentence pairs that have been selected. If a new sentence pair has more than a predefined percentage (threshold) of similar words in the word list, meaning most of the words are not new to the selected set anymore, then this sentence pair will not be included in the selection. If not, then this new sentence pair is recorded and the new words are added into the word list. The same process is repeated until there are enough count of sentence pairs. If at the end, the number of sentence pairs in the selection is inadequate, the threshold will be decreased and the same process will be repeated again until there are enough sentence pairs.

The disadvantage of this method is that only the differences of words in the sentence pairs are considered but not the syntactic structure of the sentences. Therefore, the coverage may not be sufficiently thorough. In the future, we would like to research a more intelligent method of sentence pair selection. Such a selection will prove to be useful in the future if we would like to build a new set of small hand-aligned data for a different language pair.

4. Experimental Results

We conducted our experiments with 160 K Chinese-English and Japanese-English translation pairs taken from the Basic Travel Expression Corpus (BTEC) [16]. First, 35 K Chinese-English† and 10 K Japanese-English sentence pairs were manually aligned following the guidelines in [17]. These hand-aligned corpora were used for the training of the CRF model. Next, part of the IWSLT†† evaluation campaign corpora (training, development and testing) were also used in our experiment.

4.1 Alignment Results

As compared to [1], for the implementation of the CRF model in this study, we used the 160 K parallel corpus to calculate the Dice and applied a bigger bilingual dictionary for the similarity measurement. Our Chinese-English dictionary has 617,543 entries, while the Japanese-English dictionary has 1,992,958 entries. For part-of-speech tagging and lemmatization, TreeTagger††† is used for English; ChaSen†††† for Japanese; an in-house Chinese segmenter and POS tagger, for Chinese. Since in Chinese there are no morphological changes in the word form, the lemma would be the same as the base form.

We built two CRF alignment models for each language pair: a full model with all training data (35 K and 10 K, respectively) and a small model with only 1 K training data. The 1 K training data for each language pair is selected based on the algorithm proposed in Sect. 3.4. We started with the predefined threshold as 80% and increased it by 10% for each iteration. The minimum sentence length is 5 words and the maximum length is 30 words. The test data is another 1 K dataset, a portion extracted from the full training data randomly. In other words, in relation to the full model, it is an almost closed test; however in relation to the small model, it is an almost open test.

GIZA++ is trained using the default settings (Model 4, \( 1^5H^33^4 \)), and the output of both directions are combined using the grow-diag-final-and heuristics method. Our generative alignment with agreement model ran 5 iterations for Model 1 and 5 iterations for HMM (named NICT TM 15). For the training of GIZA++ and NICT TM 15, the same 35 K training data was used for Chinese-English language pair, but an additional of 20 K of training data was added to the 10 K training data for Japanese-English language pair, making a total of 30 K. This is because, as we know, an unsupervised generative model would not do well if the training data is too small.

Table 1 shows the alignment results, evaluated by conventional precision, recall, and F-measure, as given in the

| Model          | Precision | Recall | F-measure |
|----------------|-----------|--------|-----------|
| Chinese-English (35 K) |           |        |           |
| CRF[1]         | 89.84     | 79.71  | 84.59     |
| CRF 35 K       | 97.60     | 95.26  | 96.42     |
| CRF 1 K        | 91.84     | 66.84  | 77.37     |
| GIZA++         | 76.49     | 79.36  | 77.90     |
| NICTTM         | 76.30     | 76.11  | 76.21     |
| Joint (1 K)    | 83.27     | 80.57  | 81.90     |
| Japanese-English (30 K) |           |        |           |
| CRF[1]         | 82.05     | 72.41  | 76.93     |
| CRF 10 K       | 96.67     | 93.67  | 95.14     |
| CRF 1 K        | 85.83     | 65.14  | 74.07     |
| GIZA++         | 57.65     | 61.70  | 59.61     |
| NICTTM         | 59.74     | 60.84  | 60.29     |
| Joint (1 K)    | 68.45     | 67.85  | 68.15     |

†With the co-operation of Institute of Computing Technology, Chinese Academy of Sciences

††International Workshop on Spoken Language Translation

†††http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

††††http://chasen-legacy.sourceforge.jp/
equations below; here, $A$ represents the gold-standard alignments; $S$, the output alignments; and $A \cap S$, the correct alignments.

$$\text{precision} = \frac{|A \cap S|}{|S|} \quad \text{recall} = \frac{|A \cap S|}{|A|}$$

$$F - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

In the closed test, accuracy is high since there is no unseen data in the test. The open test results reported in [1] are included for reference. In the open test using only 1K training data, although precision remains rather high, recall shows considerable deterioration. Thus, this creates room for improvement through the adding of links using the generative model. For Chinese-English language pair, the recall is high using the generative models, but the precision is low. Our proposed joint model has improved on the recall although the precision has become lower. For Japanese-English language pair, the generative models do not do any better than the CRF model for both recall and precision, but the joint model has improved the recall with a lower precision. In general, the joint model could improve the results for both recall and precision as compared to generative models. Although the joint model has lower precision compared to CRF model only, it creates less null links to alignment which is actually better for creating the phrase translation model in an SMT system.

4.2 Translation Results

We built the translation models using a small, 20 K sentences (IWSLT 2008 train data), and large amount of training data, 160 K sentences (BTEC), for Chinese-English language pairs; similarly, we used 20 K (IWSLT 2004 train data) and 160 K (BTEC) for Japanese-English language pairs. A separate development data set (506 sentence pairs with 16 references, IWSLT 2005 test data) is used for the optimization with minimum error rate trainer (MERT), which is identical for all our experiments. For testing, 507 sentence pairs for Chinese-English (IWSLT 2008 test data) and 500 sentence pairs for Japanese-English (IWSLT 2004 test data), all with 16 references, are used. The average sentence length is 7–10 words for all languages. Table 2 shows the

| Table 2 | Data sets. |
|---------|------------|
|          | Chinese-English | Japanese-English |
| Size     | Large | Small | Large | Small |
| Training | 160 K | 20 K   | 160 K | 20 K   |
| Development | 506 | 506 |
| Test     | 507   | 500 |

| Table 3 | Chinese-English translation results. |
|---------|-------------------------------------|
| Model   | BLEU | METEOR | B+M / 2 | Align points | Phrase table |
|---------|------|--------|---------|--------------|--------------|
| GIZA++ Model 4 | 44.13 | 60.08  | 52.10  | 211,866 | 352,779 |
| NICTTM TB $^3$  | 43.97 | 58.86  | 51.41  | 184,681 | 494,588 |
| CRF 35 K       | 45.47 | 59.77  | 52.62  | 190,823 | 496,653 |
| CRF 35 K + NICTTM | 46.17 | **60.88** | 53.52  | 189,070 | 413,089 |
| CRF 1 K        | 46.14 | 59.14  | 52.64  | 143,110 | 1,085,588 |
| CRF 1 K + NICTTM | **46.70** | 60.69  | **53.69** | 185,424 | 1,058,588 |

| Table 4 | Japanese-English translation results. |
|---------|-------------------------------------|
| Model   | BLEU | METEOR | B+M / 2 | Align points | Phrase table |
|---------|------|--------|---------|--------------|--------------|
| GIZA++ Model 4 | 49.44 | 60.75  | 55.09  | 237,408 | 388,949 |
| NICTTM TB $^3$  | 51.73 | 62.20  | 56.96  | 223,646 | 358,299 |
| CRF 10 K       | 51.62 | 63.34  | 57.48  | 235,385 | 3,704,005 |
| CRF 10 K + NICTTM | **52.08** | **65.85** | **58.96** | 221,363 | 324,069 |
| CRF 1 K        | 49.28 | 60.78  | 55.03  | 155,577 | 1,374,013 |
| CRF 1 K + NICTTM | 51.22 | 62.63  | 56.92  | 218,730 | 1,702,591 |

4.3 Results

We report the results on four Chinese-English and four Japanese-English language pairs, using 20 K sentences (IWSLT 2008 train data), and large amount of training data, 160 K sentences (BTEC), for Chinese-English language pairs; similarly, we used 20 K (IWSLT 2004 train data) and 160 K (BTEC) for Japanese-English language pairs. A separate development data set (506 sentence pairs with 16 references, IWSLT 2005 test data) is used for optimization with minimum error rate trainer (MERT), which is identical for all our experiments. For testing, 507 sentence pairs for Chinese-English (IWSLT 2008 test data) and 500 sentence pairs for Japanese-English (IWSLT 2004 test data), all with 16 references, are used. The average sentence length is 7–10 words for all languages. Table 2 shows the
summary of these data sets.

GIZA++ and NICTTM are trained using the same settings as in the alignment experiment, namely $1^5H^33^4S^3$ and $1^5H^3$. The CRF model is trained using 35 K/1 K and 10 K/1 K training data for Chinese and Japanese, respectively. These training datasets are the subsets of the whole 160 K training datasets, but not in the smaller 20 K datasets. The SMT decoder is an in-house standard phrase-based decoder, CleopATRa. The results are evaluated automatically using the BLEU score, a geometric mean of n-gram precision with respect to N reference translations, and METEOR score, which calculates unigram overlaps between translations and reference texts using various levels of matches (exact, stem, synonym), and their average. Table 3 and Table 4 show the translation results for Chinese-English and Japanese-English translation directions, respectively.

We can observe that when only the CRF alignment model is used, the number of alignment points is very low even with the full training data, and this causes the phrase table to be 2–3 times larger than that of the GIZA++ and NICTTM models. This is even more obvious when the training data for CRF model is small (1 K). Although the accuracy of translation is just slightly lower or equivalent, the decoding time increased tremendously as the phrase table is large. By constraining the NICTTM model with the CRF model output, the translation accuracy increased by around 0.97–3.87 points as compared to the GIZA++ model. The improvement is more significant in the case of Japanese-English translation.

From the results, we can also conclude that if the CRF model has performed very well in terms of alignment with a large amount of training data (35 K or 10 K), which implies that it leaves only very little space for the NICTTM, the improvement may not as good when using a small amount of training data (1 K). In the joint alignment model, the CRF model serves as a hard constraint to quicken the alignment process in the generative model with respect to its high-precision output. Moreover, the CRF model trained with a small amount of training data cannot be satisfactorily used as a complete alignment model because too many null links remain. A joint model is a perfect solution as it takes advantages of the two models and eliminates their disadvantages.

5. Analysis

Table 5 shows some examples of translation outputs for Japanese-English translation pair. In these cases, the CRF+NICTTM joint alignment model has given better translation output due to the better quality of word alignment results.

6. Conclusions and Future Work

We have presented a simple combination method to introduce a discriminative model into a generative model in order to retain the high precision of the former and the recall of the latter. We incorporated some linguistic features either lexically or syntactically into the CRF model in order to improve the accuracy of word alignment. For the generative model, we deployed the alignment by agreement on both directions with Model 1 and an HMM, which also improves word alignment. Our combination of the advantages of the two-stage models has not only further improved the accuracy of alignment but also made the alignment more suitable for use in building the translation model used in a phrase-based SMT system. Our experimental results proved that the proposed method improved the performance of translation by up to 1.0–3.9 points (the average of BLEU and METEOR scores) as compared to using GIZA++. Since our method needs only a small amount of hand-aligned data, it is feasible to apply it to any other language pair. Future work includes the selection of 1 K parallel text used for manual alignment that is more representative for training the CRF model and the feature set used for the CRF model. Furthermore, the confidence measure from the CRF model may also be useful in the generative model for deciding whether to include or exclude a certain link.
Acknowledgements

This work is partly supported by the Grant-in-Aid for Scientific Research (C) and the Special Coordination Funds for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] C.L. Goh and E. Sumita, “A feature-rich supervised word alignment model for phrase-based statistical machine translation,” International Journal of Asian Language Processing, vol.19, no.3, pp.109–125, 2009.
[2] P. Blunsom and T. Cohn, “Discriminative word alignment with conditional random fields,” Proc. COLING/ACL, pp.65–72, 2006.
[3] S. Lacoste-Julien, D. Klein, B. Taskar, and M. Jordan, “Word alignment via quadratic assignment,” Proc. HLT/NAACL, pp.112–119, 2006.
[4] R.C. Moore, “A discriminative framework for bilingual word alignment,” Proc. HLT/EMNLP, pp.81–88, 2005.
[5] B. Taskar, S. Lacoste-Julien, and D. Klein, “A discriminative matching approach to word alignment,” Proc. HLT/EMNLP, pp.73–80, 2005.
[6] A. Ittycheriah and S. Roukos, “A maximum entropy word aligner for Arabic-English machine translation,” Proc. HLT/EMNLP, pp.89–96, 2005.
[7] H. Wu, H. Wang, and Z. Liu, “Boosting statistical word alignment using labeled and unlabeled data,” Proc. COLING/ACL Poster Session, pp.913–920, 2006.
[8] A. Fraser and D. Marcu, “Semi-supervised training for statistical word alignment,” Proc. COLING/ACL, pp.769–776, 2006.
[9] A. Fraser and D. Marcu, “Getting the structure right for word alignment: LEAF,” Proc. 2007 Joint Conference on EMNLP and CoNLL, pp.51–60, 2007.
[10] Y. Deng and Y. Gao, “Guiding statistical word alignment models with prior knowledge,” Proc. ACL, pp.1–8, 2007.
[11] Y. Liu, Q. Liu, and S. Lin, “Log-linear models for word alignment,” Proc. ACL, pp.459–466, 2005.
[12] N.F. Ayan and B.J. Dorr, “A maximum entropy approach to combining word alignments,” Proc. HLT/NAACL, pp.96–103, 2006.
[13] P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer, “The mathematics of statistical machine translation: Parameter estimation,” Computational Linguistics, vol.19, no.2, pp.263–311, 1993.
[14] F.J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” Computational Linguistics, vol.29, no.1, pp.19–52, 2003.
[15] P. Liang, B. Taskar, and D. Klein, “Alignment by agreement,” Proc. HLT/NAACL, pp.104–111, 2006.
[16] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, “Comparative study on corpora for speech translation,” IEEE Trans. Audio, Speech and Language, vol.14, no.5, pp.1674–1682, 2006.
[17] H. Zhao, Q. Liu, R. Zhang, Y. Lü, E. Sumita, and C.L. Goh, “A guideline for Chinese-English word alignment,” J. Chinese Information Processing, vol.23, no.3, pp.65–87, 2009.

Chooi-Ling Goh received the M.E. and Ph.D. in Information Science from Nara Institute of Science and Technology, Japan, in 2003 and 2006, respectively. She is currently an expert researcher at National Institute of Information and Communications Technology, Japan. She is a member of the ANLP. Her main research interests include machine translation and morphological analysis.

Taro Watanabe received the B.E. and M.E. degrees in information science from Kyoto Univ., Kyoto, Japan in 1994 and 1997, respectively, and obtained the Master of Science degree in language and information technologies from the School of Computer Science, Carnegie Mellon University in 2000. In 2004, he received the Ph.D. in informatics from Kyoto Univ., Kyoto, Japan. Dr. Watanabe is a researcher of National Institute of Information and Communications Technology. His research interests include natural language processing, machine learning and statistical machine translation.

Hirofumi Yamamoto received the M.S. degree in agriculture from the Tokyo University in 1981 and the Ph.D. degree in global information and telecommunication from the Waseda University in 2004. Dr. Yamamoto is currently a professor at Kinki University School of Science and Engineering Dept. Informatics, short-term researcher at National Institute of Information and Communications Technology, and cooperate researcher at ATR. His research interests include speech recognition and machine translation. He is a member of the IEEE, the ASJ and the ANLP.

Eiichiro Sumita received the M.S. degree in computer science from the University of Electro-Communications in 1982 and the Ph.D. degree in engineering from Kyoto University in 1999. Dr. Sumita is the group leader of NICT/MASTAR Project/Language Translation Group, and the visiting professor of Kobe University. His research interests include machine translation and e-Learning.