Word Alignment for Languages with Scarce Resources Using Bilingual Corpora of Other Language Pairs

Haifeng Wang Hua Wu Zhanyi Liu
Toshiba (China) Research and Development Center
5/F., Tower W2, Oriental Plaza, No.1, East Chang An Ave., Dong Cheng District
Beijing, 100738, China
{wanghaifeng, wuhua, liuzhanyi}@rdc.toshiba.com.cn

Abstract

This paper proposes an approach to improve word alignment for languages with scarce resources using bilingual corpora of other language pairs. To perform word alignment between languages L1 and L2, we introduce a third language L3. Although only small amounts of bilingual data are available for the desired language pair L1-L2, large-scale bilingual corpora in L1-L3 and L2-L3 are available. Based on these two additional corpora and with L3 as the pivot language, we build a word alignment model for L1 and L2. This approach can build a word alignment model for two languages even if no bilingual corpus is available in this language pair. In addition, we build another word alignment model for L1 and L2 using the small L1-L2 bilingual corpus. Then we interpolate the above two models to further improve word alignment between L1 and L2. Experimental results indicate a relative error rate reduction of 21.30% as compared with the method only using the small bilingual corpus in L1 and L2.

1 Introduction

Word alignment was first proposed as an intermediate result of statistical machine translation (Brown et al., 1993). Many researchers build alignment links with bilingual corpora (Wu, 1997; Och and Ney, 2003; Cherry and Lin, 2003; Zhang and Gildea, 2005). In order to achieve satisfactory results, all of these methods require a large-scale bilingual corpus for training. When the large-scale bilingual corpus is unavailable, some researchers acquired class-based alignment rules with existing dictionaries to improve word alignment (Ker and Chang, 1997). Wu et al. (2005) used a large-scale bilingual corpus in general domain to improve domain-specific word alignment when only a small-scale domain-specific bilingual corpus is available.

This paper proposes an approach to improve word alignment for languages with scarce resources using bilingual corpora of other language pairs. To perform word alignment between languages L1 and L2, we introduce a third language L3 as the pivot language. Although only small amounts of bilingual data are available for the desired language pair L1-L2, large-scale bilingual corpora in L1-L3 and L2-L3 are available. Using these two additional bilingual corpora, we train two word alignment models for language pairs L1-L3 and L2-L3, respectively. And then, with L3 as a pivot language, we can build a word alignment model for L1 and L2 based on the above two models. Here, we call this model an *induced model*. With this induced model, we perform word alignment between languages L1 and L2 even if no parallel corpus is available for this language pair. In addition, using the small bilingual corpus in L1 and L2, we train another word alignment model for this language pair. Here, we call this model an *original model*. An *interpolated model* can be built by interpolating the induced model and the original model.

As a case study, this paper uses English as the pivot language to improve word alignment between Chinese and Japanese. Experimental results show that the induced model performs better than the original model trained on the small Chinese-Japanese corpus. And the interpolated model further improves the word alignment results, achieving a relative error rate reduction of

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 874–881,
Sydney, July 2006.

21.30% as compared with results produced by the original model.

The remainder of this paper is organized as follows. Section 2 discusses the related work. Section 3 introduces the statistical word alignment models. Section 4 describes the parameter estimation method using bilingual corpora of other language pairs. Section 5 presents the interpolation model. Section 6 reports the experimental results. Finally, we conclude and present the future work in section 7.

2 Related Work

A shared task on word alignment was organized as part of the ACL 2005 Workshop on Building and Using Parallel Texts (Martin et al., 2005). The focus of the task was on languages with scarce resources. Two different subtasks were defined: *Limited resources* and *Unlimited resources*. The former subtask only allows participating systems to use the resources provided. The latter subtask allows participating systems to use any resources in addition to those provided.

For the subtask of unlimited resources, Aswani and Gaizauskas (2005) used a multi-feature approach for many-to-many word alignment on English-Hindi parallel corpora. This approach performed local word grouping on Hindi sentences and used other methods such as dictionary lookup, transliteration similarity, expected English words, and nearest aligned neighbors. Martin et al. (2005) reported that this method resulted in absolute improvements of up to 20% as compared with the case of only using limited resources. Tufis et al. (2005) combined two word aligners: one is based on the limited resources and the other is based on the unlimited resources. The unlimited resource consists of a translation dictionary extracted from the alignment of Romanian and English WordNet. Lopez and Resnik (2005) extended the HMM model by integrating a tree distortion model based on a dependency parser built on the English side of the parallel corpus. The latter two methods produced comparable results with those methods using limited resources. All the above three methods use some language dependent resources such as dictionary, thesaurus, and dependency parser. And some methods, such as transliteration similarity, can only be used for very similar language pairs.

In this paper, besides the limited resources for the given language pair, we make use of large amounts of resources available for other language pairs to address the alignment problem for languages with scarce resources. Our method does not need language-dependent resources or deep linguistic processing. Thus, it is easy to adapt to any language pair where a pivot language and corresponding large-scale bilingual corpora are available.

3 Statistical Word Alignment

According to the IBM models (Brown et al., 1993), the statistical word alignment model can be generally represented as in equation (1).

$$\Pr(\mathbf{a},\mathbf{f}\mid\mathbf{c}) = \frac{\Pr(\mathbf{a},\mathbf{f}\mid\mathbf{c})}{\sum_{\mathbf{a}'}\Pr(\mathbf{a}',\mathbf{f}\mid\mathbf{c})} \tag{1}$$

Where, **c** and **f** represent the source sentence and the target sentence, respectively[1].

In this paper, we use a simplified IBM model 4 (Al-Onaizan et al., 1999), which is shown in equation (2). This version does not take into account word classes in Brown et al. (1993).

$$\begin{aligned}\Pr(\mathbf{a},\mathbf{f}\mid\mathbf{c}) &= \binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0} \cdot \\ &\prod_{i=1}^{l} n(\phi_i \mid c_i) \cdot \prod_{j=1}^{m} t(f_j \mid c_{a_j}) \cdot \\ &\Big(\prod_{j=1,a_j\neq 0}^{m} ([j = h(a_j)] \cdot d_1(j - \odot_{i-1})) + \\ &\prod_{j=1,a_j\neq 0}^{m} ([j \neq h(a_j)] \cdot d_{>1}(j - p(j)))\Big)\end{aligned} \tag{2}$$

l, m are the lengths of the source sentence and the target sentence respectively.

j is the position index of the target word.

a_j is the position of the source word aligned to the j^{th} target word.

ϕ_i is the fertility of c_i.

p_0, p_1 are the fertility probabilities for c_0, and $p_0 + p_1 = 1$.

$t(f_j \mid c_{a_j})$ is the word translation probability.

$n(\phi_i \mid c_i)$ is the fertility probability.

$d_1(j - \odot_{i-1})$ is the distortion probability for the head word of the cept.

$d_{>1}(j - p(j))$ is the distortion probability for the non-head words of the cept.

[1] This paper uses **c** and **f** to represent a Chinese sentence and a Japanese sentence, respectively. And **e** represents an English sentence.

$h(i) = \min_k \{k : i = a_k\}$ is the head of cept *i*.

$p(j) = \max_{k<j} \{k : a_j = a_k\}$.

$\odot_i$ is the center of cept *i*.

During the training process, IBM model 3 is first trained, and then the parameters in model 3 are employed to train model 4. For convenience, we describe model 3 in equation (3). The main difference between model 3 and model 4 lies in the calculation of distortion probability.

$$\Pr(\mathbf{a},\mathbf{f} \mid \mathbf{c}) = \binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0} \cdot \prod_{i=1}^{l} n(\phi_i \mid c_i) \cdot \prod_{i=1}^{l} \phi_i! \cdot \prod_{j=1}^{m} t(f_j \mid c_{a_j}) \cdot \prod_{j=1, a_j \neq 0}^{m} d(j \mid a_j, l, m) \quad (3)$$

4 Parameter Estimation Using Bilingual Corpora of Other Language Pairs

As shown in section 3, the word alignment model mainly has three kinds of parameters that must be specified, including the translation probability, the fertility probability, and the distortion probability. The parameters are usually estimated by using bilingual sentence pairs in the desired languages, namely Chinese and Japanese here. In this section, we describe how to estimate the parameters without using the Chinese-Japanese bilingual corpus. We introduce English as the pivot language, and use the Chinese-English and English-Japanese bilingual corpora to estimate the parameters of the Chinese-Japanese word alignment model. With these two corpora, we first build Chinese-English and English-Japanese word alignment models as described in section 3. Then, based on these two models, we estimate the parameters of Chinese-Japanese word alignment model. The estimated model is named *induced model*.

The following subsections describe the method to estimate the parameters of Chinese-Japanese alignment model. For reversed Japanese-Chinese word alignment, the parameters can be estimated with the same method.

4.1 Translation Probability

Basic Translation Probability

We use the translation probabilities trained with Chinese-English and English-Japanese corpora to estimate the Chinese-Japanese probability as shown in equation (4). In (4), we assume that the translation probability $t_{\mathrm{EJ}}(f_j \mid e_k, c_i)$ is independent of the Chinese word c_i .

$$\begin{aligned} t_{\mathrm{CJ}}(f_j \mid c_i) &= \sum_{e_k} t_{\mathrm{EJ}}(f_j \mid e_k, c_i) \cdot t_{\mathrm{CE}}(e_k \mid c_i) \\ &= \sum_{e_k} t_{\mathrm{EJ}}(f_j \mid e_k) \cdot t_{\mathrm{CE}}(e_k \mid c_i) \end{aligned} \quad (4)$$

Where $t_{\mathrm{CJ}}(f_j \mid c_i)$ is the translation probability for Chinese-Japanese word alignment. $t_{\mathrm{EJ}}(f_j \mid e_k)$ is the translation probability trained using the English-Japanese corpus. $t_{\mathrm{CE}}(e_k \mid c_i)$ is the translation probability trained using the Chinese-English corpus.

Cross-Language Word Similarity

In any language, there are ambiguous words with more than one sense. Thus, some noise may be introduced by the ambiguous English word when we estimate the Chinese-Japanese translation probability using English as the pivot language. For example, the English word "bank" has at least two senses, namely:

bank1 - a financial organization
bank2 - the border of a river

Let us consider the Chinese word:

河岸 - bank2 (the border of a river)

And the Japanese word:

銀行 - bank1 (a financial organization)

In the Chinese-English corpus, there is high probability that the Chinese word "河岸(bank2)" would be translated into the English word "bank". And in the English-Japanese corpus, there is also high probability that the English word "bank" would be translated into the Japanese word "銀行(bank1)".

As a result, when we estimate the translation probability using equation (4), the translation probability of "銀行(bank1)" given "河岸(bank2)" is high. Such a result is not what we expect.

In order to alleviate this problem, we introduce cross-language word similarity to improve translation probability estimation in equation (4). The cross-language word similarity describes how likely a Chinese word is to be translated into a Japanese word with an English word as the pivot. We make use of both the Chinese-English corpus and the English-Japanese corpus to calculate the cross language word similarity between a Chinese word *c* and a Japanese word *f* given an

Input: An English word e, a Chinese word c, and a Japanese word f;
The Chinese-English corpus; The English-Japanese corpus.

(1) Construct Set 1: identify those Chinese-English sentence pairs that include the given Chinese word c and English word e, and put the English sentences in the pairs into Set 1.

(2) Construct Set 2: identify those English-Japanese sentence pairs that include the given English word e and Japanese word f, and put the English sentences in the pairs into Set 2.

(3) Construct the feature vectors V_{CE} and V_{EJ} of the given English word using all other words as context in Set 1 and Set 2, respectively.

$V_{CE} =< (e_1, ct_{11}), (e_2, ct_{12}), ..., (e_n, ct_{1n}) >$

$V_{EJ} =< (e_1, ct_{21}), (e_2, ct_{22}), ..., (e_n, ct_{2n}) >$

Where ct_{ij} is the frequency of the context word e_j. $ct_{ij} = 0$ if e_j does not occur in Set i.

(4) Given the English word e, calculate the cross-language word similarity between the Chinese word c and the Japanese word f as in equation (5)

$$sim(c, f; e) = \cos(V_{CE}, V_{EJ}) = \frac{\sum_j ct_{1j} \cdot ct_{2j}}{\sqrt{\sum_j (ct_{1j})^2} \cdot \sqrt{\sum_j (ct_{2j})^2}} \quad (5)$$

Output: The cross language word similarity $sim(c, f; e)$ of the Chinese word c and the Japanese word f given the English word e

Figure 1. Similarity Calculation

English word *e*. For the ambiguous English word *e*, both the Chinese word *c* and the Japanese word *f* can be translated into *e*. The sense of an instance of the ambiguous English word *e* can be determined by the context in which the instance appears. Thus, the cross-language word similarity between the Chinese word *c* and the Japanese word *f* can be calculated according to the contexts of their English translation *e*. We use the feature vector constructed using the context words in the English sentence to represent the context. So we can calculate the cross-language word similarity using the feature vectors. The detailed algorithm is shown in figure 1. This idea is similar to translation lexicon extraction via a bridge language (Schafer and Yarowsky, 2002).

For example, the Chinese word "河岸" and its English translation "bank" (the border of a river) appears in the following Chinese-English sentence pair:

(a) *他们沿着河岸走回家。*

(b) *They walked home along the river bank.*

The Japanese word "銀行" and its English translation "bank" (a financial organization) appears in the following English-Japanese sentence pair:

(c) *He has plenty of money in the bank.*

(d) *彼は銀行預金が相当ある。*

The context words of the English word "bank" in sentences (b) and (c) are quite different. The difference indicates the cross language word similarity of the Chinese word "河岸" and the Japanese word "銀行" is low. So they tend to have different senses.

Translation Probability Embedded with Cross Language Word Similarity

Based on the cross language word similarity calculation in equation (5), we re-estimate the translation probability as shown in (6). Then we normalize it in equation (7).

The word similarity of the Chinese word "河岸 (bank2)" and the Japanese word "銀行 (bank1)" given the word English word "bank" is low. Thus, using the updated estimation method, the translation probability of "銀行 (bank1)" given "河岸(bank2)" becomes low.

$$t'_{CJ}(f_j \mid c_i) = \sum_{e_k} (t_{EJ}(f_j \mid e_k) \cdot t_{CE}(e_k \mid c_i) \cdot sim(c_i, f_j; e_k)) \quad (6)$$

$$t_{CJ}(f_j \mid c_i) = \frac{t'_{CJ}(f_j \mid c_i)}{\sum_{f'} t'_{CJ}(f' \mid c_i)} \quad (7)$$

4.2 Fertility Probability

The induced fertility probability is calculated as shown in (8). Here, we assume that the probabil-

ity $n_{\mathrm{EJ}}(\phi_i \mid e_k, c_i)$ is independent of the Chinese word c_i.

$$\begin{aligned} & n_{\mathrm{CJ}}(\phi_i \mid c_i) \\ &= \sum_{e_k} n_{\mathrm{EJ}}(\phi_i \mid e_k, c_i) \cdot t_{\mathrm{CE}}(e_k \mid c_i) \\ &= \sum_{e_k} n_{\mathrm{EJ}}(\phi_i \mid e_k) \cdot t_{\mathrm{CE}}(e_k \mid c_i) \end{aligned} \tag{8}$$

Where, $n_{\mathrm{CJ}}(\phi_i \mid c_i)$ is the fertility probability for the Chinese-Japanese alignment. $n_{\mathrm{EJ}}(\phi_i \mid e_k)$ is the trained fertility probability for the English-Japanese alignment.

4.3 Distortion Probability in Model 3

With the English language as a pivot language, we calculate the distortion probability of model 3. For this probability, we introduce two additional parameters: one is the position of English word and the other is the length of English sentence. The distortion probability is estimated as shown in (9).

$$\begin{aligned} & d_{\mathrm{CJ}}(j \mid i, l, m) \\ &= \sum_{k,n} \Pr(j, k, n \mid i, l, m) \\ &= \sum_{k,n} \Pr(j \mid k, n, i, l, m) \cdot \Pr(k, n \mid i, l, m) \\ &= \sum_{k,n} (\Pr(j \mid k, n, i, l, m) \cdot \\ &\qquad \Pr(k \mid n, i, l, m) \cdot \Pr(n \mid i, l, m)) \end{aligned} \tag{9}$$

Where, $d_{\mathrm{CJ}}(j \mid i, l.m)$ is the estimated distortion probability. k is the introduced position of an English word. n is the introduced length of an English sentence.

In the above equation, we assume that the position probability $\Pr(j \mid k, n, i, l, m)$ is independent of the position of the Chinese word and the length of the Chinese sentence. And we assume that the position probability $\Pr(k \mid n, i, l, m)$ is independent of the length of Japanese sentence. Thus, we rewrite these two probabilities as follows.

$$\Pr(j \mid k, n, i, l, m) \approx \Pr(j \mid k, n, m) = d_{\mathrm{EJ}}(j \mid k, n, m)$$

$$\Pr(k \mid i, l, m, n) \approx \Pr(k \mid i, l, n) = d_{\mathrm{CE}}(k \mid i, l, n)$$

For the length probability, the English sentence length n is independent of the word positions i. And we assume that it is uniformly distributed. Thus, we take it as a constant, and rewrite it as follows.

$$\Pr(n \mid i, l, m) = \Pr(n \mid l, m) = \text{constant}$$

According to the above three assumptions, we ignore the length probability $\Pr(n \mid l, m)$. Equation (9) is rewritten in (10).

$$\begin{aligned} & d_{\mathrm{CJ}}(j \mid i, l.m) \\ &= \sum_{k,n} d_{\mathrm{EJ}}(j \mid k, n, m) \cdot d_{\mathrm{CE}}(k \mid i, l, n) \end{aligned} \tag{10}$$

4.4 Distortion Probability in Model 4

In model 4, there are two parameters for the distortion probability: one for head words and the other for non-head words.

Distortion Probability for Head Words

The distortion probability $d_1(j - \odot_{i-1})$ for head words represents the relative position of the head word of the i^{th} cept and the center of the $(i\text{-}1)^{\text{th}}$ cept. Let $\Delta j = j - \odot_{i-1}$, then Δj is independent of the absolute position. Thus, we estimate the distortion probability by introducing another relative position $\Delta j'$ of English words, which is shown in (11).

$$\begin{aligned} & d_{1,\mathrm{CJ}}(\Delta j = j - \odot_{i-1}) \\ &= \sum_{\Delta j'} d_{1,\mathrm{CE}}(\Delta j') \cdot \Pr_{\mathrm{EJ}}(\Delta j \mid \Delta j') \end{aligned} \tag{11}$$

Where, $d_{1,\mathrm{CJ}}(\Delta j = j - \odot_{i-1})$ is the estimated distortion probability for head words in Chinese-Japanese alignment. $d_{1,\mathrm{CE}}(\Delta j')$ is the distortion probability for head word in Chinese-English alignment. $\Pr_{\mathrm{EJ}}(\Delta j \mid \Delta j')$ is the translation probability of relative Japanese position given relative English position.

In order to simplify $\Pr_{\mathrm{EJ}}(\Delta j \mid \Delta j')$, we introduce j' and $\odot_{i'-1}$ and let $\Delta j' = j' - \odot_{i'-1}$, where j' and $\odot_{i'-1}$ are positions of English words. We rewrite $\Pr_{\mathrm{EJ}}(\Delta j \mid \Delta j')$ in (12).

$$\begin{aligned} & \Pr_{\mathrm{EJ}}(\Delta j \mid \Delta j') \\ &= \Pr_{\mathrm{EJ}}(j - \odot_{i-1} \mid j' - \odot_{i'-1}) \\ &= \sum_{\substack{j, \odot_{i-1}: j - \odot_{i-1} = \Delta j \\ j', \odot_{i'-1}: j' - \odot_{i'-1} = \Delta j'}} \Pr_{\mathrm{EJ}}(j, \odot_{i-1} \mid j', \odot_{i'-1}) \end{aligned} \tag{12}$$

The English word in position j' is aligned to the Japanese word in position j, and the English word in position $\odot_{i'-1}$ is aligned to the Japanese word in position $\odot_{i-1}$.

We assume that j and $\odot_{i-1}$ are independent, j only depends on j', and $\odot_{i-1}$ only depends on $\odot_{i'-1}$. Then $\Pr_{\mathrm{EJ}}(j, \odot_{i-1} \mid j', \odot_{i'-1})$ can be estimated as shown in (13).

$$\Pr_{EJ}(j, \odot_{i-1} \mid j', \odot_{i'-1}) = \Pr_{EJ}(j \mid j') \cdot \Pr_{EJ}(\odot_{i-1} \mid \odot_{i'-1}) \quad (13)$$

Both of the two parameters in (13) represent the position translation probabilities. Thus, we can estimate them from the distortion probability in model 3. $\Pr_{EJ}(j \mid j')$ is estimated as shown in (14). And $\Pr_{EJ}(\odot_{i-1} \mid \odot_{i'-1})$ can be estimated in the same way. In (14), we also assume that the sentence length distribution $\Pr(l, m \mid j')$ is independent of the word position and that it is uniformly distributed.

$$\begin{aligned} \Pr_{EJ}(j \mid j') &= \sum_{l,m} \Pr_{EJ}(j, l, m \mid j') \\ &= \sum_{l,m} d_{EJ}(j \mid j', l, m) \cdot \Pr(l, m \mid j') \\ &= \sum_{l,m} d_{EJ}(j \mid j', l, m) \end{aligned} \quad (14)$$

Distortion Probability for Non-Head Words

The distortion probability $d_{>1}(j - p(j))$ describes the distribution of the relative position of non-head words. In the same way, we introduce relative position $\Delta j'$ of English words, and model the probability in (15).

$$d_{>1,CJ}(\Delta j = j - p(j)) = \sum_{\Delta j'} d_{>1,CE}(\Delta j') \cdot \Pr_{EJ}(\Delta j \mid \Delta j') \quad (15)$$

$d_{>1,CJ}(\Delta j = j - p(j))$ is the estimated distortion probability for the non-head words in Chinese-Japanese alignment. $d_{>1,CE}(\Delta j')$ is the distortion probability for non-head words in Chinese-English alignment. $\Pr_{EJ}(\Delta j \mid \Delta j')$ is the translation probability of the relative Japanese position given the relative English position.

In fact, $\Pr_{EJ}(\Delta j \mid \Delta j')$ has the same interpretation as in (12). Thus, we introduce two parameters j' and $p(j')$ and let $\Delta j' = j' - p(j')$, where j' and $p(j')$ are positions of English words. The final distortion probability for non-head words can be estimated as shown in (16).

$$d_{>1,CJ}(\Delta j = j - p(j)) = \sum_{\Delta j'} (d_{>1,CE}(\Delta j') \cdot \sum_{\substack{j, p(j): j - p(j) = \Delta j \\ j', p(j'): j' - p(j') = \Delta j'}} \Pr_{EJ}(j \mid j') \cdot \Pr_{EJ}(p(j) \mid p(j'))) \quad (16)$$

5 Interpolation Model

With the Chinese-English and English-Japanese corpora, we can build the induced model for Chinese-Japanese word alignment as described in section 4. If we have small amounts of Chinese-Japanese corpora, we can build another word alignment model using the method described in section 3, which is called the *original model* here. In order to further improve the performance of Chinese-Japanese word alignment, we build an interpolated model by interpolating the induced model and the original model.

Generally, we can interpolate the induced model and the original model as shown in equation (17).

$$\Pr(\mathbf{a}, \mathbf{f} \mid \mathbf{c}) = \lambda \cdot \Pr_{O}(\mathbf{a}, \mathbf{f} \mid \mathbf{c}) + (1 - \lambda) \cdot \Pr_{I}(\mathbf{a}, \mathbf{f} \mid \mathbf{c}) \quad (17)$$

Where $\Pr_{O}(\mathbf{a}, \mathbf{f} \mid \mathbf{c})$ is the original model trained from the Chinese-Japanese corpus, and $\Pr_{I}(\mathbf{a}, \mathbf{f} \mid \mathbf{c})$ is the induced model trained from the Chinese-English and English-Japanese corpora. λ is an interpolation weight. It can be a constant or a function of $\mathbf{f}$ and $\mathbf{c}$.

In both model 3 and model 4, there are mainly three kinds of parameters: translation probability, fertility probability and distortion probability. These three kinds of parameters have their own interpretation in these two models. In order to obtain fine-grained interpolation models, we interpolate the three kinds of parameters using different weights, which are obtained in the same way as described in Wu et al. (2005). λ_t represents the weights for translation probability. λ_n represents the weights for fertility probability. λ_{d3} and λ_{d4} represent the weights for distortion probability in model 3 and in model 4, respectively. λ_{d4} is set as the interpolation weight for both the head words and the non-head words. The above four weights are obtained using a manually annotated held-out set.

6 Experiments

In this section, we compare different word alignment methods for Chinese-Japanese alignment. The "Original" method uses the original model trained with the small Chinese-Japanese corpus. The "Basic Induced" method uses the induced model that employs the basic translation probability without introducing cross-language word similarity. The "Advanced Induced" method uses the induced model that introduces the cross-language word similarity into the calculation of the translation probability. The "Interpolated" method uses the interpolation of the word alignment models in the "Advanced Induced" and "Original" methods.

6.1 Data

There are three training corpora used in this paper: Chinese-Japanese (CJ) corpus, Chinese-English (CE) Corpus, and English-Japanese (EJ) Corpus. All of these tree corpora are from general domain. The Chinese sentences and Japanese sentences in the data are automatically segmented into words. The statistics of these three corpora are shown in table 1. "# Source Words" and "# Target Words" mean the word number of the source and target sentences, respectively.

Language Pairs	#Sentence Pairs	# Source Words	# Target Words
CJ	21,977	197,072	237,834
CE	329,350	4,682,103	4,480,034
EJ	160,535	1,460,043	1,685,204

Table 1. Statistics for Training Data

Besides the training data, we also have held-out data and testing data. The held-out data includes 500 Chinese-Japanese sentence pairs, which is used to set the interpolated weights described in section 5. We use another 1,000 Chinese-Japanese sentence pairs as testing data, which is not included in the training data and the held-out data. The alignment links in the held-out data and the testing data are manually annotated. Testing data includes 4,926 alignment links[2].

6.2 Evaluation Metrics

We use the same metrics as described in Wu et al. (2005), which is similar to those in (Och and Ney, 2000). The difference lies in that Wu et al. (2005) took all alignment links as sure links.

If we use S_G to represent the set of alignment links identified by the proposed methods and S_C to denote the reference alignment set, the methods to calculate the precision, recall, f-measure, and alignment error rate (AER) are shown in equations (18), (19), (20), and (21), respectively. It can be seen that the higher the f-measure is, the lower the alignment error rate is. Thus, we will only show precision, recall and AER scores in the evaluation results.

$$precision = \frac{|S_G \cap S_C|}{|S_G|} \quad (18)$$

$$recall = \frac{|S_G \cap S_C|}{|S_C|} \quad (19)$$

$$fmeasure = \frac{2|S_G \cap S_C|}{|S_G| + |S_C|} \quad (20)$$

$$AER = 1 - \frac{2|S_G \cap S_C|}{|S_G| + |S_C|} = 1 - fmeasure \quad (21)$$

6.3 Experimental Results

We use the held-out data described in section 6.1 to set the interpolation weights in section 5. λ_t is set to 0.3, λ_n is set to 0.1, λ_{d3} for model 3 is set to 0.5, and λ_{d4} for model 4 is set to 0.1. With these parameters, we get the lowest alignment error rate on the held-out data.

For each method described above, we perform bi-directional (source to target and target to source) word alignment and obtain two alignment results. Based on the two results, we get a result using "refined" combination as described in (Och and Ney, 2000). Thus, all of the results reported here describe the results of the "refined" combination. For model training, we use the GIZA++ toolkit[3].

Method	Precision	Recall	AER
Interpolated	0.6955	0.5802	0.3673
Advanced Induced	0.7382	0.4803	0.4181
Basic Induced	0.6787	0.4602	0.4515
Original	0.6026	0.4783	0.4667

Table 2. Word Alignment Results

The evaluation results on the testing data are shown in table 2. From the results, it can be seen that both of the two induced models perform better than the "Original" method that only uses the limited Chinese-Japanese sentence pairs. The "Advanced Induced" method achieves a relative error rate reduction of 10.41% as compared with the "Original" method. Thus, with the Chinese-English corpus and the English-Japanese corpus, we can achieve a good word alignment results even if no Chinese-Japanese parallel corpus is available. After introducing the cross-language word similarity into the translation probability, the "Advanced Induced" method achieves a relative error rate reduction of 7.40% as compared with the "Basic Induced" method. It indicates that cross-language word similarity is effective in the calculation of the translation probability. Moreover, the "interpolated" method further improves the result, which achieves relative error

[2] For a non one-to-one link, if *m* source words are aligned to *n* target words, we take it as one alignment link instead of *m*∗*n* alignment links.

[3] It is located at http://www.fjoch.com/ GIZA++.html.

rate reductions of 12.51% and 21.30% as compared with the "Advanced Induced" method and the "Original" method.

7 Conclusion and Future Work

This paper presented a word alignment approach for languages with scarce resources using bilingual corpora of other language pairs. To perform word alignment between languages L1 and L2, we introduce a pivot language L3 and bilingual corpora in L1-L3 and L2-L3. Based on these two corpora and with the L3 as a pivot language, we proposed an approach to estimate the parameters of the statistical word alignment model. This approach can build a word alignment model for the desired language pair even if no bilingual corpus is available in this language pair. Experimental results indicate a relative error reduction of 10.41% as compared with the method using the small bilingual corpus.

In addition, we interpolated the above model with the model trained on the small L1-L2 bilingual corpus to further improve word alignment between L1 and L2. This interpolated model further improved the word alignment results by achieving a relative error rate reduction of 12.51% as compared with the method using the two corpora in L1-L3 and L3-L2, and a relative error rate reduction of 21.30% as compared with the method using the small bilingual corpus in L1 and L2.

In future work, we will perform more evaluations. First, we will further investigate the effect of the size of corpora on the alignment results. Second, we will investigate different parameter combination of the induced model and the original model. Third, we will also investigate how simpler IBM models 1 and 2 perform, in comparison with IBM models 3 and 4. Last, we will evaluate the word alignment results in a real machine translation system, to examine whether lower word alignment error rate will result in higher translation accuracy.

References

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, and David Yarowsky. 1999. Statistical Machine Translation Final Report. *Johns Hopkins University Workshop*.

Niraj Aswani and Robert Gaizauskas. 2005. Aligning Words in English-Hindi Parallel Corpora. In *Proc. of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, pages 115-118.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): 263-311.

Colin Cherry and Dekang Lin. 2003. A Probability Model to Improve Word Alignment. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pages 88-95.

Sue J. Ker and Jason S. Chang. 1997. A Class-based Approach to Word Alignment. *Computational Linguistics*, 23(2): 313-343.

Adam Lopez and Philip Resnik. 2005. Improved HMM Alignment Models for Languages with Scarce Resources. In *Proc. of the ACL-2005 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, pages 83-86.

Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word Alignment for Languages with Scarce Resources. In *Proc. of the ACL-2005 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, pages 65-74.

Charles Schafer and David Yarowsky. 2002. Inducing Translation Lexicons via Diverse Similarity Measures and Bridge Languages. In *Proc. of the 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, pages 1-7.

Dan Tufis, Radu Ion, Alexandru Ceausu, and Dan Stefanescu. 2005. Combined Word Alignments. In *Proc. of the ACL-2005 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, pages 107-110.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 440-447.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51.

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

Hua Wu, Haifeng Wang, and Zhanyi Liu. 2005. Alignment Model Adaptation for Domain-Specific Word Alignment. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 467-474.

Hao Zhang and Daniel Gildea. 2005. Stochastic Lexicalized Inversion Transduction Grammar for Alignment. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 475-482.