A Multi-aligner for Japanese-Chinese Parallel Corpora

Yujie Zhang and *Qun Liu and Qing Ma and Hitoshi Isahara

National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan, 619-0289
(yujie, qma, isahara)@nict.go.jp
*Institute of Computing Technology, Chinese Academy of Sciences
No.6 South KeXueYuan Road, Haidian District, Beijing 100080, China
liuqun@ict.ac.cn

Abstract

Automatic word alignment is an important technology for extracting translation knowledge from parallel corpora. However, automatic techniques cannot resolve this problem completely because of variances in translations. We therefore need to investigate the performance potential of automatic word alignment and then decide how to suitably apply it. In this paper we first propose a lexical knowledge-based approach to word alignment on a Japanese-Chinese corpus. Then we evaluate the performance of the proposed approach on the corpus. At the same time we also apply a statistics-based approach, the well-known toolkit GIZA++, to the same test data. Through comparison of the performances of the two approaches, we propose a multi-aligner, exploiting the lexical knowledge-based aligner and the statistics-based aligner at the same time. Quantitative results confirmed the effectiveness of the multi-aligner.

1 Introduction

In a parallel corpus, automatic word alignment is to identify the translation relations between the words in a source sentence and those in a target sentence. A word-aligned parallel corpus has many applications, such as machine translation, machine-aided translation, bilingual lexicography, and word-sense disambiguation. For these applications, much research on automatic word alignment has been conducted and reported.

The statistics-based approach is widely studied (Och and Ney, 2003), and is mainly based on the research of statistical machine translation (Brown et al., 1993). However, this approach incorrectly aligns less frequently occurring words when statistically significant evidence is not available. Instead of word-based statistics, Ker proposed a class-based approach by using lexicon resources (Ker and Chang, 1997). Based on this idea, various types of linguistic knowledge are taken into account for the heuristic (Huang and Choi, 2000), (Deng, 2004). No automatic techniques, neither statistics-based nor linguistics-based approaches, can resolve the problem of word alignment completely because of variance in translations. To decide how to best use these techniques, we need to know what level of performance they can provide.

This paper presents an investigation of the performance of the two approaches, linguistics and statistics-based, on a Japanese-Chinese parallel corpus. We first propose a lexical knowledge-based approach to word alignment and then evaluated its performance using the corpus. We also applied a statistics-based approach to the same test data. Through comparison of the performances of the two approaches, we propose a multi-aligner by exploiting the lexical knowledge-based aligner and the statistics-based aligner.

These efforts are part of a larger project to construct a Japanese-Chinese parallel corpus, which was started in 2002 at NICT. In this project, we need to annotate the alignment at word level. Previously reported research involved many language pairs, such as English-French, English-German, English-Japanese, Chinese-English, and Chinese-Korean. To our knowledge, there is no report on Japanese-Chinese word alignment. Word alignment is often thought to be easier for Japanese-Chinese because some Japanese characters are the same as Chinese characters. However, no quantitative result has been reported. The experimental results obtained in this work gave us new insights on aligning words in a Japanese-Chinese parallel corpus.

2 Japanese-Chinese Parallel Corpus

The corpus we used in this study consists of 38,383 Japanese sentences from Mainichi newspaper and their Chinese translations. The corpus has been morphological annotated (word segmented and part-of-speech tagged) in the first phase of the project. For Japanese morphological

annotation, the definition of the Corpus of Spontaneous Japanese was adopted (Maekawa, 2000). For Chinese, the definition of Peking University was adopted (Zhou and Yu, 1994). The average lengths of the sentences on both sides are about 30 words.

The study, word alignment, aims to assist to word alignment annotation, which is a task in the second phase of the project.

3 Word Alignment Approach

In our Japanese-Chinese parallel corpus, a Japanese sentence *J* and its Chinese translation *C* are given as a pair. Both *J* and *C* are segmented into words as described in Section 2. Let W_J and W_C denote their word lists, respectively. This section will describe how to align a word *j* in W_J with its translation *c* in W_C. Here, we propose a lexical knowledge-based approach that consists of two algorithms. The first algorithm aims to establish reliable alignments by using lexical knowledge. The second algorithm aims to select the most likely alignments from the remaining alignment candidates by using dislocation information.

3.1 Algorithm for Establishing Reliable Alignment

In measuring the degree of similarity between two strings, x and y, the Dice coefficient (Dice, 1945) is often used. It is defined as follows.

$$Sim(x,y) = \frac{2\times|x \bigcap y|}{|x|+|y|} \quad , \qquad (1)$$

where $|x|$ ($|y|$) is the number of morphemes in x (y), and $|x \bigcap y|$ is the number of the morphemes in the intersection of x and y. Based on this measure, we can estimate the likelihood of j in W_J being aligned with c in W_C by measuring the similarity between the Chinese translation of j and c.

When considering the case of one-to-more alignment, we also consider the case of j being aligned with a sequence of words from c_i to c_{i+k} $(1 \le i \le |W_C|, 0 \le k \le l)$. l is the largest number of words in a Chinese sentence that can be aligned with a Japanese word. We set $l = 4$ in this paper. Hereafter, we use $\ddot{c}$ to express any word sequence in W_C within the length of 4 and use $\ddot{c}(i,k)$ to identify a certain sequence that starts at the position i with a length of k $(1 \le i \le |W_C|, 0 \le k \le 4)$. One-to-one alignment is a special case when $k = 0$. Actually, the case of more-to-one has also been considered in the study. For simplicity of describtion, however, only the case of one-to-more is described here.

Three kinds of lexical resources used for the estimation are described below .

Orthography

About half of Japanese words contain *kanji*, the Chinese characters used in Japanese writing. We call them kanji words. Japanese words may also contain *hiragana* or *katakana*, which are phonetic characters. Because some kanji words were adapted directly from China, their Chinese translations are the same as the words themselves. For example, the Chinese translations for the Japanese words 人民 (people) and 国家(country) are 人民 and 国家, respectively, with the same orthographies. Based on this observation, we assume that $\ddot{c}$ is probably the translation of j in W_J if their orthographies are similar, and j is therefore probably aligned to $\ddot{c}$. The following formula is defined to estimate the possibility of j being aligned with $\ddot{c}$.

$$Poss_{ort}(j,\ddot{c}) = Sim(j,\ddot{c}). \qquad (2)$$

$Poss_{ort}$ expresses the possibility that is estimated by using orthography. The morpheme of j may be kanji, katakana, or hiragana, and the morpheme of $\ddot{c}$ is a Chinese character.

Simple and Traditional Chinese Characters

In Chinese, the traditional Chinese characters were simplified in the Chinese reformation. In Mandarin, simplified Chinese characters are used. At the same time, many Japanese kanji words maintain the form of the traditional Chinese characters as they were when they were introduced from China. The Chinese translations of such kanji words are usually the simplified character of the traditional characters. For example, the Chinese translation of the Japanese word 故郷 (hometown) is 故乡, in which 乡 is the simplified character of 郷. Based on this phenomenon, we assume that $\ddot{c}$ in W_C is probably the translation of j in W_J if its traditional form is similar to j, and therefore j is probably aligned with $\ddot{c}$. Therefore, converting $\ddot{c}$ into traditional characters and then measuring its similarity to j allows us to estimate the possibility of j being aligned to $\ddot{c}$. Let $Trad(\ddot{c})$ denote the traditional form of $\ddot{c}$ by converting each simplified character of $\ddot{c}$ into a traditional character. We then define the following formula.

$$Poss_{tra}(j,\ddot{c}) = Sim(j, Trad(\ddot{c})). \quad (3)$$

$Poss_{tra}$ expresses the possibility estimated by using the correspondence between simplified Chinese characters and traditional Chinese characters.

Bilingual Dictionary

A translation dictionary can help to identify the translation relations between j in W_J and $\ddot{c}$ in W_C. We assume that $\ddot{c}$ is probably the translation of j if $\ddot{c}$ is similar to the Chinese translation of j, and therefore j is probably aligned with $\ddot{c}$. Let C_j denote the Chinese translation set of j and let c' denote one translation ($c' \in C_j$). We can estimate the possibility of j being aligned with $\ddot{c}$ using the following formula (Ker and Chang, 1997).

$$Poss_{dic}(j,\ddot{c}) = \max_{c' \in C_j} Sim(c', \ddot{c}). \quad (4)$$

$Poss_{dic}$ expresses the possibility estimated by using a translation dictionary. In Section 4, we will describe how to automatically build a Japanese-Chinese dictionary.

We have described how to estimate the possibility of alignment between j and $\ddot{c}$ by using three kinds of lexical resources: orthography, the correspondence between the simplified Chinese characters and the traditional Chinese characters, and a translation dictionary. We will combine them in the following formula, where $Poss_{lex}$ expresses the possibility estimated by using the three kinds of lexical resources.

$$Poss_{lex}(j,\ddot{c}) = \max(Poss_{ort}(j,\ddot{c}), Poss_{tra}(j,\ddot{c}), Poss_{dic}(j,\ddot{c})) \quad (5)$$

Algorithm 1, to establish reliable alignment, is described as follows.

Algorithm 1

Align j in W_J with $\ddot{c}$ in W_C using lexical resources.

Input: Japanese word list W_J and Chinese word list W_C.

Output: Reliable alignment A_{rel}

Step 1. For all $\ddot{c}$ in W_C, get $Trad(\ddot{c})$ by converting them into traditional Chinese characters.

Step 2. For all j in W_J, search the translation dictionary to obtain Chinese translation set C_j.

Step 3. For all j in W_J and all $\ddot{c}$ in W_C, calculate $Poss_{ort}(j,\ddot{c})$ using formula (2), $Poss_{tra}(j,\ddot{c})$ using formula (3), $Poss_{dic}(j,\ddot{c})$ using formula (4), and $Poss_{lex}(j,\ddot{c})$ using formula (5).

Step 4. For each j in W_J, if $\max_{\ddot{c} \text{ in } W_C, 1 \le i \le |W_C|, 0 \le k \le 4} Poss_{lex}(j, \ddot{c}(i,k)) \ge \theta_{lex}$, output $(j, \hat{\ddot{c}})$ $(\hat{\ddot{c}}(\hat{i},\hat{k}) = \arg\max_{\ddot{c} \text{ in } W_C, 1 \le i \le |W_C|, 0 \le k \le 4} Poss_{lex}(j, \ddot{c}(i,k)))$ to A_{rel}, where θ_{lex} is a preset threshold.

3.2 Algorithm for Broadening Coverage

This section describes an augmentation algorithm for finding the most likely alignment from the remaining candidates. In this algorithm we only consider one to one alignment. Let $\overline{W_J}$ denote the list of words j $(\in W_J)$ that are still not aligned, and let $\overline{W_C}$ denote the list of words c $(\in W_C)$ that are still not aligned.

We observed that words in one syntactic structure are to be translated into words that belong to the same syntactic structure in the target sentence. When j_1 and j_2 belong to the same syntactic structure and we know the translation position of j_1, we can use this knowledge to infer the translation position of j_2. When syntactic analysis techniques are not available for source and target languages, the left context and right contexts are referred to instead. For this purpose, we used the alignments established in Algorithm 1 as the left and right context. For an alignment candidate $(\tilde{j}, \tilde{c})$, we take four established alignments into account: the two alignments that are the nearest to $\tilde{j}$ on the left and right and the two alignments that are the nearest to $\tilde{c}$ on the left and right. For $\tilde{j} (\in \overline{W_J})$, we estimate the possibility of $\tilde{j}$ being aligned with $\tilde{c} (\in \overline{W_C})$ as follows.

First, add ($Null_0, Null_0$) and ($Null_{|W_J|+1}, Null_{|W_C|+1}$) to A_{rel} as the leftmost and the rightmost alignments. For a j in W_J, we use $m(j)$ to express its position in W_J. For a c in W_C, we use $n(c)$ to express its position in W_C. For a

sequence of Chinese words $\ddot{c}$ that has been aligned in Algorithm 1, we use $n_s(\ddot{c})$ to express the starting position and $n_e(\ddot{c})$ to express the ending position of $\ddot{c}$ in W_C.

Second, for $\tilde{j}$ and $\tilde{c}$, search for the following four alignments in A_{rel}.

(1) $a_{\tilde{j}_L} = (j_{\tilde{j}_L}, \ddot{c}_{\tilde{j}_L})$ in which $j_{\tilde{j}_L}$ is the nearest word to the left of $\tilde{j}$.

(2) $a_{\tilde{j}_R} = (j_{\tilde{j}_R}, \ddot{c}_{\tilde{j}_R})$ in which $j_{\tilde{j}_R}$ is the nearest word to the right of $\tilde{j}$.

(3) $a_{\tilde{c}_L} = (j_{\tilde{c}_L}, \ddot{c}_{\tilde{c}_L})$ in which the last word in $\ddot{c}_{\tilde{c}_L}$ is the nearest word to the left of $\tilde{c}$.

(4) $a_{\tilde{c}_R} = (j_{\tilde{c}_R}, \ddot{c}_{\tilde{c}_R})$ in which the first word in $\ddot{c}_{\tilde{c}_R}$ is the nearest word to the right of $\tilde{c}$.

The following quantitative variables measure the degree at which $\tilde{j}$ and $\tilde{c}$ dislocate from the four reliable alignments.

$$\Delta m_{\tilde{j}_L} = m(\tilde{j}) - m(j_{\tilde{j}_L}), \Delta n_{\tilde{j}_L} = n(\tilde{c}) - n_e(\ddot{c}_{\tilde{j}_L}), \quad (6)$$

$$\Delta m_{\tilde{j}_R} = m(\tilde{j}) - m(j_{\tilde{j}_R}), \Delta n_{\tilde{j}_R} = n(\tilde{c}) - n_s(\ddot{c}_{\tilde{j}_R}),$$

$$\Delta m_{\tilde{c}_L} = m(\tilde{j}) - m(j_{\tilde{c}_L}), \Delta n_{\tilde{c}_L} = n(\tilde{c}) - n_e(\ddot{c}_{\tilde{c}_L}), \text{and}$$

$$\Delta m_{\tilde{c}_R} = m(\tilde{j}) - m(j_{\tilde{c}_R}), \Delta n_{\tilde{c}_R} = n(\tilde{c}) - n_s(\ddot{c}_{\tilde{c}_R}).$$

Third, estimate the possibility of $\tilde{j}$ being aligned with $\tilde{c}$ by referring to the four alignments $a_{\tilde{j}_L}$, $a_{\tilde{j}_R}$, $a_{\tilde{c}_L}$ and $a_{\tilde{c}_R}$ as follows (Deng, 2004).

$$Poss_{\tilde{j}_L}(\tilde{j}, \tilde{c}) = \frac{2}{(|\Delta m_{\tilde{j}_L}| + |\Delta n_{\tilde{j}_L}|) e^{|\Delta m_{\tilde{j}_L} - \Delta n_{\tilde{j}_L}|}}, \quad (7)$$

$$Poss_{\tilde{j}_R}(\tilde{j}, \tilde{c}) = \frac{2}{(|\Delta m_{\tilde{j}_R}| + |\Delta n_{\tilde{j}_R}|) e^{|\Delta m_{\tilde{j}_R} - \Delta n_{\tilde{j}_R}|}},$$

$$Poss_{\tilde{c}_L}(\tilde{j}, \tilde{c}) = \frac{2}{(|\Delta m_{\tilde{c}_L}| + |\Delta n_{\tilde{c}_L}|) e^{|\Delta m_{\tilde{c}_L} - \Delta n_{\tilde{c}_L}|}}, \text{and}$$

$$Poss_{\tilde{c}_R}(\tilde{j}, \tilde{c}) = \frac{2}{(|\Delta m_{\tilde{c}_R}| + |\Delta n_{\tilde{c}_R}|) e^{|\Delta m_{\tilde{c}_R} - \Delta n_{\tilde{c}_R}|}}.$$

The first item in the denominator lays penalty using the degree at which $\tilde{j}$ and $\tilde{c}$ dislocate from one reliable alignment. The larger the sum of them is, the smaller the possibility of the alignment is. The second item in the denominator lays penalty using the degree at which $\tilde{j}$ and $\tilde{c}$ dislocate from one reliable alignment in an opposite direction. When $\tilde{j}$ and $\tilde{c}$ dislocate from one reliable alignment in an opposite direction, the possibility is smaller. When $\tilde{j}$ and $\tilde{c}$ dislocate from one reliable alignment in a parallel direction, the possibility is larger. The exponential function is used in the second item because the fact that $\tilde{j}$ and $\tilde{c}$ dislocate from one reliable alignment in the same direction or not is thought more important.

Finally, select the reliable alignment with the largest value as the final reference context and use it for estimation.

$$Poss_{dis}(\tilde{j}, \tilde{c}) = \max(Poss_{\tilde{j}_L}(\tilde{j}, \tilde{c}), Poss_{\tilde{j}_R}(\tilde{j}, \tilde{c}), Poss_{\tilde{c}_L}(\tilde{j}, \tilde{c}), Poss_{\tilde{c}_R}(\tilde{j}, \tilde{c})) \quad (8)$$

$Poss_{dis}$ expresses the possibility estimated by using dislocation information. Algorithm 2, to broaden coverage, is described as follows.

Algorithm 2

Align $\tilde{j} \in \overline{W_J}$ with $\tilde{c} \in \overline{W_C}$ by referring to the established alignments.

Input: Japanese word list $\overline{W_J}$, Chinese word lists $\overline{W_C}$ and A_{rel}.

Output: Augmented alignment A_{aug}

Step 1. For all $\tilde{j} \in \overline{W_J}$ and all $\tilde{c} \in \overline{W_C}$, search for $a_{\tilde{j}_L}$, $a_{\tilde{j}_R}$, $a_{\tilde{c}_L}$ and $a_{\tilde{c}_R}$ in A_{rel}.

Step 2. For all $\tilde{j} \in \overline{W_J}$ and all $\tilde{c} \in \overline{W_C}$, calculate $Poss_{\tilde{j}_L}$, $Poss_{\tilde{j}_R}$, $Poss_{\tilde{c}_L}$, $Poss_{\tilde{c}_R}$ using formula (6) and (7), and then $Poss_{dis}(\tilde{j}, \tilde{c})$ using formula (8).

Step 3. For each $\tilde{j} \in \overline{W_J}$, if $\max_{\tilde{c} \in \overline{W_C}} Poss_{dis}(\tilde{j}, \tilde{c}) > \theta_{dis}$ and $Poss_{lex}(\tilde{j}, \hat{c}) > \theta'_{lex}$ ($\hat{c} = \arg\max_{\tilde{c} \in \overline{W_C}} Poss_{dis}(\tilde{j}, \tilde{c}) > \theta_{dis}$), output $(\tilde{j}, \hat{c})$ to A_{aug}, where θ_{dis} and $\theta'_{lex} (< \theta_{lex})$ are preset thresholds.

In Step 3, $Poss_{le}(\tilde{j}, \hat{c}) > \theta_{le-l}$ means that the lexical-knowledge is also used to filter out canditated alignmetns.

Finally, we output A_{rel} and A_{aug} as alignment results.

4 Automatically Building a Japanese-Chinese

Dictionary

Although a bilingual dictionary is a very important resource in word alignment, we have no machine-readable Japanese-Chinese dictionary. We do, however, have a machine-readable Japanese-English dictionary and a machine-readable English-Chinese dictionary. We have automatically built a Japanese-Chinese dictionary by applying a method of using the third language as an intermediary (Zhang et al., 2005). One aim is to use the dictionary in this study, word alignment, as described in formula (4). Another aim is to confirm the efficiency the automatically built dictionary in word alignment.

4.1 Obtain Chinese Translation Candidates for Japanese Words

Two machine-readable dictionaries are as follows.

EDR Japanese-English Dictionary (NICT, 2002)
It contains 364,430 records, each of which consists of Japanese word, part-of-speech, English translations, etc.

LDC English-Chinese Dictionary (LDC, 2002)
It contains 110,834 records, each of which consists of English word and Chinese translations.

The first step is to obtain Chinese translation candidates. For each EDR record, the procedure is as follows. First, collect the English translations that are single words. Second, for each collected English translation, look up the word in the LDC English-Chinese Dictionary and obtain the Chinese translations. Then designate all the obtained Chinese translations as the set of Chinese translation candidates for the Japanese record. As a result, 144,002 records of EDR obtained their sets of Chinese translation candidates.

4.2 Selecting Correct Chinese Translations Using Heuristics

To select correct translations, we ranked candidates by referring to their possibilities of being correct translations. To estimate the possibilities, we utilized three sources of heuristic information: the number of English translations in common, the part of speech, and Japanese kanji information. The scores estimated from the three sources of information are denoted as S_e , S_{pos} , and S_{kanji} respectively. We then defined a scoring function as follows, where the three scores are integrated into one measurement scale.

$$Score(j,c) = k_e \times S_e(j,c) + k_{pos} \times S_{pos}(j,c) + k_{kanji} \times S_{kanji}(j,c) \quad (9)$$

$Score(j,c)$ thus gives the score of a Chinese word c being a correct translation of a Japanese word j . k_e , k_{pos} , and k_{kanji} are the weights of S_e , S_{pos} , and S_{kanji} respectively, with the restriction that $k_e + k_{pos} + k_{kanji} = 1.0$. Next, we introduce each source of information and explain how to use them for estimating scores.

Number of English Translations in Common

If a translation candidate and a source word share multiple English translations, the two words may be considered nearer to each other in meaning, and therefore, the candidate may be regarded as more likely to be correct. $S_e(j,c)$ is calculated according to formula (10), where $E(j)$ and $E(c)$ are the sets of English translations for j and c , respectively (Bond et al., 2001).

$$S_e(j,c) = sim(E(j), E(c)) \quad (10)$$

Part of Speech

The Japanese words and the corresponding Chinese translations have some similarities in syntactic function. Based on this observation, we selected candidates whose categories were similar or nearly similar to the category of the original Japanese word. We used the degree of part of speech similarity between the source word and the translation candidate to measure their similarity in meaning. Let $S_{pos}(j,c)$ denote the degree of similarity between the part of speech of j and the part of speech of c . We manually determined the degree of similarity between the Japanese and Chinese parts of speech. This degree has four levels: “similar”, “approximately similar”, “unknown”, and “not similar”. This manual definition work took a bilingual expert three days. Then, the four qualitative degrees of similarity were assigned to the four levels as 1.0, 0.8, 0.2, and 0, respectively. For example, if the parts of speech of j and c are common noun and noun, respectively, $S_{pos}(j,c) = 1.0$.

Use of Kanji Information

We acquired a correspondence between kanji and Chinese characters from EDR and LDC, using the same method by focusing on the Japanese record that is single kanji and the Chinese translations that are single Chinese characters. As a result, we

obtained Chinese translation candidates in single characters for 2,847 kanji, which were then ranked by using two kinds of heuristics, the number of English translations in common and orthography.

We then utilized the obtained correspondences to measure the similarity between Japanese words and Chinese translation candidates. First, each kanji of j is translated into Chinese characters according to the obtained correspondences. Let $Tran(j)$ denote this translation. Second, the distance between $Tran(j)$ and one Chinese translation candidate can be computed by simply using the edit distance algorithm (Levenshtein, 1965). Here, the edit unit is a Chinese character. Then the edit distance is normalized by the following formula (11) and then is used to measure the similarity between a Japanese word j and a Chinese translation candidate c.

$$S_{kanji}(j,c) = 1 - \frac{\text{EditDistance (Tran(j), c)}}{\max(|Tran(j)|, |c|)} \qquad (11)$$

4.3 Obtained Japanese-Chinese Dictionary

To evaluate the method, we carried out a few experiments. Test data was selected from Japanese words that have more than 20 Chinese translation candidates. We randomly selected 109 Japanese words. Through tests on various combinations of weights, the best performance was obtained with k_e =0.3, k_{pos} =0.3, and k_{kanji} =0.4. The evaluation results showed that for 90.8% of the tested Japanese words, the method found one correct translation in the top results, with an accuracy of 81.4%. Using the best combinations of the three weights, we ranked Chinese translation candidates for each Japanese word of the 144,002 records and took the results that were ranked within top 10 for building a Japanese-Chinese dictionary. We then used the obtained Japanese-Chinese dictionary in the proposed word alignment approach.

5 Evaluation of Word Alignment

In this section we evaluate the automatic word alignments using our Japanese-Chinese parallel corpus. The automatic word alignments include the proposed lexical knowledge-based approach and a statistics-based approach. The test data is 1,127 sentence pairs. For each Japanese sentence and its Chinese translation, word alignments were annotated manually as gold standards. In total, 17,332 alignments were obtained.

5.1 Evaluation of Proposed Approach

In the application of Algorithm 1, orthography, the correspondence between the simplified Chinese characters and the traditional Chinese characters and the automatically built Japanese-Chinese dictionary were used first individually, and then in combination. In addition, Algorithm 2 was applied for augmentation. Thresholds in Algorithms 1 and 2 were set as $\theta_{lex} = 0.85, \theta'_{lex} = 0.4,$ and $\theta_{dis} = 0.8$, by referring to the empirical knowledge in (Ker and Chang, 19973) and (Deng, 2004). An example of the automatically aligned results is shown in Figure 1.

The results were evaluated in terms of three measures, Precision, Recall and F-measure. The evaluation results are shown in Table 1.

Heuristics	Precision (%)	Recall (%)	F-measure
Orthography (Ort)	98	25	39.8
Traditional (Tra)	97	11	19.8
Dictionary (Dic)	87	19	31.2
Ort+Tra	98	27	42.3
Ort+Tra+Dic	92	36	51.8
Ort+Tra+Dic+Dislocation	69	58	63.02

Table 1. The evaluation results of the proposed approach by using different heuristics and their combinations.

It is showed that using orthography, traditional Chinese characters, and the automatically built dictionary individually obtained high precisions, between 87% and 97%, but low recall rates, between 11% and 25%. Using orthography obtained recall rates of 25%. This implies that the degree of similarity in orthography between Japanese sentences and their Chinese translations is about 25%.

Using orthography and traditional Chinese characters obtained recall rates of 27%. Kanji information can therefore help to determine 27% of alignments.

After the bilingual dictionary was added, the recall rate and F-measure were both improved by about 9% while maintaining a high precision at 92%. The automatically built dictionary thus broadened the coverage with only a small decrease in precision. This proved the efficacy of the method of automatically building a translation dictionary.

After applying Algorithm 2, using dislocation information for augmentation, the recall rate and F-measure were further increased to 58% and 63.2%, respectively, although precision decreases to 69%.

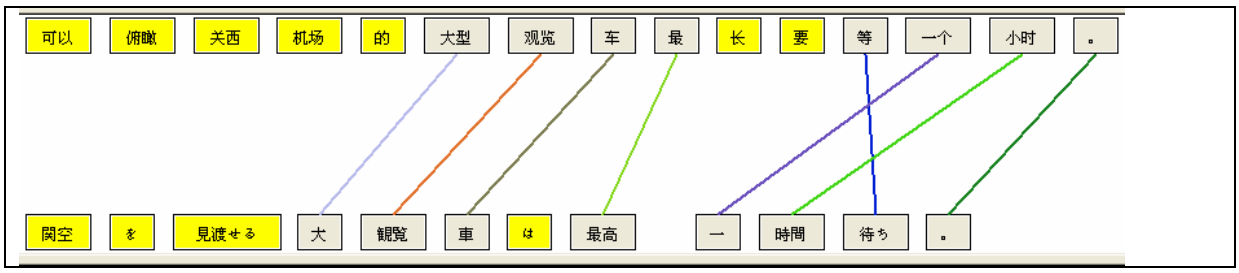

Figure 1. Example of the automatically aligned results.

5.2 Comparison with a Statistics-Based Approach

We compared the proposed approach with a statistics-based approach, the well-known toolkit, GIZA++. In the application of GIZA++, two directions were tested: the Chinese sentences were used as source sentences and the Japanese sentences as target sentences, and vice versa. The comparison results are shown in Table 2.

Method	Precision (%)	Recall (%)	F-measure
GIZA(C→J)	55	73	62.7
GIZA(J→C)	46	55	50
Proposed approach	69	58	63

Table 2. Comparison of performances of GIZA++ and the proposed lexical knowledge-based approach.

The results produced by C→J of GIZA++ were better than those produced by J→C of GIZA++. Compared with the results produced by J→C of GIZA++, our approach achieved better performances. Compared with the results produced by C→J of GIZA++, our approach achieved the same performance in F-measure, but with higher precision and a lower recall rate. By comparing the results produced by the three aligners, we found that each has its own advantage in certain aspects. The proposed approach obtained a higher precision but GIZA++ (C→J) obtained a higher recall rate. The proposed approach could correctly align the less frequently occurring words, while GIZA++ could not because statistically significant evidence was not available. On the other hand, GIZA++ could correctly align the often occurring words, for some of which the proposed approach could not because of the deficiency of the obtained translation dictionary. We further considered using the three aligners together.

5.3 Method of Multi-aligner

In this method, the results produced by the proposed knowledge-based approach, C → J of GIZA++, and J→C of GIZA++ were selected in a majority decision. If an alignment result was produced by two or three aligners at the same time, the result was accepted. Otherwise, was abandoned. In this way, we aimed to utilize the results of each aligner and maintain high precision at the same time. Table 3 shows the evaluation results of the multi-aligner.

	Precision (%)	Recall (%)	F-measure
Multi-aligner	79	63	70.1

Table 3. Evaluation results of the multi-aligner consisting of our proposed approach, J→C of GIZA++, and C→J of GIZA++.

The multi-aligner produced satisfactory results. This performance is evidence that the multi-aligner is feasible for use in word alignment annotation in the construction of a Japanese-Chinese parallel corpus. Comparing Table 3 with Table 2 reveals that the multi-aligner was superior to the proposed approach and J→C of GIZA++ in precision, recall rate, and F-measure. Compared with C → J of GIZA++, the multi-aligner achieved higher precision and as a result achieved a higher F-measure. We therefore conclude that the performance of the multi-aligner consisting of the proposed lexical knowledge-based approach, J→C of GIZA++, and C→J of GIZA++ is superior to each of them individually.

6 Conclusion

This paper presented a lexical knowledge-based approach for word alignment. The approach consists of two algorithms. The first algorithm is used to obtain reliable alignments by using three types of heuristics: orthography, the correspondence between the simplified Chinese characters and the traditional Chinese characters, and an automatically built Japanese-Chinese dictionary. The second algorithm is used to broaden coverage by estimating the dislocation of a candidate from the established reliable alignments. The two algorithms and three heuristics were evaluated by application to test data. After a comparison with the results produced by Japanese to Chinese and Chinese to Japanese alignment of GIZA++, we proposed a multi-aligner method. The

experimental results on the same test data confirmed the superior performance of the multi-aligner. In the future research, we will improve the lexical knowledge-based approach to increase coverage further while maintaining high precision.

References

Brown, P.F., Pietra, S.A.D., Pietra, V. J. D., Mercer, R. L. 1993. *The Mathematics of Statistical Machine Translation: Parameters Estimation.* Computational Linguistics, Vol. 19, Num. 2, pages 263-311.

Bond, F., Yamazaki, T., Sulong, R. B., Okura, K. 2001. *Design and Construction of a Machine-tractable Japanese-Malay Lexicon.* Proceedings of 7th Annual Meeting of the Association for Natural Language Processing, pages 62–65.

Dice, L.R. 1945. *Measures of the amount of ecologic association between species.* Journal of Ecology (26), pages 297-302.

Huang, J.X., Choi, K.S. 2000. *Chinese-Korean word alignment based on linguistic comparison.* Annual Meeting of the Association for Computational Linguistics, pages 392–399.

Ker, S.J., Chang, J.S. 1997. *A Class-based Approach to Word Alignment.* Computational Linguistics, Vol. 23, Num. 2, pages 313–343.

LDC. 2002. *English-to-Chinese Wordlist (version 2.)* http://www.ldc.upenn.edu/Projects/Chinese/.

Levenshtein, V.I. 1965. *Binary codes capable of correcting deletions, insertions and reversals.* Doklady Akademii Nauk SSSR,Vol. 163, Num. 4, pages 845–848.

Deng D. 2004. *Research on Chinese-English Word Alignment.* Master Thesis, Institute of Computing Technology, Chinese Academy of Sciences.

Maekawa, K., Koiso, H., Furui, F., Isahara, H. 2000. *Spontaneous Speech Corpus of Japanese.* Proceedings of LREC2000, pages 947–952.

NICT (National Institute of Information and Communications Technology). 2002. *EDR Electronic Dictionary Version 2.0 Technical Guide.*

Och, F.J., Ney, H. 2003. *A systematic comparison of various statistical alignment models.* Computational Linguistics, Vol. 29, Num. 1, pages 19–51.

Zhang, Y., Ma, Q., Isahara, H. 2005. *Automatic Construction of Japanese-Chinese Translation Dictionary Using English as Intermediary.* Journal of Natural Language Processing, Vol. 12, No. 2, pages 63-85.

Zhou, Q., Yu, S. 1994. *Blending Segmentation with Tagging in Chinese Language Corpus Processing.* In Proc. of COLING-94. pages 1274–1278.