Building an Annotated Japanese-Chinese Parallel Corpus – A Part of NICT Multilingual Corpora

Yujie Zhang and Kiyotaka Uchimoto 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, uchimoto,isahara)@nict.go.jp

Abstract

We are constricting a Japanese-Chinese parallel corpus, which is a part of the NICT Multilingual Corpora. The corpus is general domain, of large scale of about 40,000 sentence pairs, long sentences, annotated with detailed information and high quality. To the best of our knowledge, this will be the first annotated Japanese-Chinese parallel corpus in the world. We created the corpus by selecting Japanese sentences from Mainichi Newspaper and then manually translating them into Chinese. We then annotated the corpus with morphological and syntactic structures and alignments at word and phrase levels. This paper describes the specification in human translation and detailed information annotation, and the tools we developed in the project. The experience we obtained and points we paid special attentions are also introduced for share with other researches in corpora construction.

1 Introduction

A parallel corpus is a collection of articles, paragraphs, or sentences in two different languages. Since a parallel corpus contains translation correspondences between the source text and its translations at different level of constituents, it is a critical resource for extracting translation knowledge in machine translation (MT). Although recently some versions of machine translation software have become available in the market, translation quality is still a significant problem. Therefore, a detailed examination into human translation is still required. This will provide a basis for radically improving machine translation in the near future. In addition, in MT system development, the example-based method and the statistics-based method are widely researched and applied. Parallel corpora are required by the translation studies and practical system development.

The raw text of a parallel corpus contains implicit knowledge. If we annotate some information, we can get explicit knowledge from the corpus. The more information that is annotated on a parallel corpus, the more knowledge we can get from the corpus. The parallel corpora of European languages are usually raw texts without annotation on syntactic structure since their syntactic structures are similar and MT does not require such annotation information. However, when language pairs are different in syntactic structures, such as the pair of English and Japanese and the pair of Japanese and Chinese, transformation between syntactic structures is difficult. A parallel corpus annotated with syntactic structures would thus be helpful to MT. Yamada has reported that alignment of syntactic structures in a parallel corpus could effectively improve translation quality. Therefore annotating a parallel corpus with syntactic structures and alignment is very meaningful. Besides MT, an annotated parallel corpus can be applied to cross-lingual information retrieval, language teaching, machine-aided translation, bilingual lexicography, and word-sense disambiguation.

Parallel corpora between European languages are well developed and are available through the Linguistic Data Consortium (LDC). However, parallel corpora between European languages and Asian languages are less developed, and parallel corpora between two Asian languages are even less developed.

The National Institute of Information and Communications Technology therefore started a project to build multilingual parallel corpora in 2002 (Uchimoto et al., 2004). The project focuses on Asian language pairs and annotation of detailed information, including syntactic structure and alignment at word and phrase levels. We call the corpus the NICT Multilingual Corpora. The corpus will be open to the public in the near future.

2 Overview of the NICT Multilingual Corpora

At present, a Japanese-English parallel corpus and a Japanese-Chinese parallel corpus are under construction following systematic specifications.

The parallel texts in each corpus consist of the original text in the source language and its translations in the target language. The original data is from newspaper articles or journals, such as Mainichi Newspaper in Japanese. The original articles were translated by skilled translators. In human translation, the articles of one domain were all assigned to the same translator to maintain consistent terminology in the target language. Different translators then revised the translated articles. Each article was translated one sentence to one sentence, so the obtained parallel corpora are already sentence aligned.

The details of the current version of the NICT Multilingual Corpora are listed in Table 1.

Corpora	Total	Original	Translation
Japanese-English Parallel Corpus	37,987 sentence pairs; (English 900,000 words)	Japanese (19,669 sentences, Mainichi Newspaper)	English Translation
		English (18,318 Sentences, Wall Street Journal)	Japanese Translation
Japanese-Chinese Parallel Corpus	38,383 sentence pairs; (Chinese 1,410,892 Characters, 926,838 words)	Japanese (38,383 sentences, Mainichi Newspaper)	Chinese Translation

Table 1 Details of current version of NICT Multilingual Corpora

The following is an example of English and Chinese translations of a Japanese sentence from Mainichi Newspaper.

[Ex. 1]

J: いずれも十九歳前後の若者で、質問に答える気力も残っていない。

E: They were all about nineteen years old and had no strength left even to answer questions.

C: 这些俄军士兵均为十九岁左右的年青人，他们甚至连回答问题的气力也没有。

In addition to the human translation, another big task is annotating the information. The annotation process has two steps: automatic annotation and human revision. In automatic annotation, we applied existing analysis techniques and tag sets. In human revision, we developed assisting tools that have powerful functions to help annotators. The annotation for each language included morphological and syntactic structures. After morphological annotation, alignments at word and phrase level were annotated.

The NICT Multilingual Corpora constructed in this way have the following characteristics.

(1) The original data is from newspaper and journals, and is therefore not domain limited.

(2) Each corpus consists of original sentences and their translations, so they are already sentence aligned.

(3) In translation of each sentence, the context of the article in which the sentence is contained is also considered. Thus, the context of each article is also well maintained in its translation, which can be exploited in the future.

(4) The corpora are annotated at high quality with morphological and syntactic structures and word/phrase alignment.

In the following section, we will describe the details in the construction of the Japanese-Chinese parallel corpus.

3 Human Translation from Japanese to Chinese

About 40,000 Japanese sentences from issues of Mainichi Newspaper were translated by skilled translators. The translation guidelines were as follows.

(1) One Japanese sentence is translated into one Chinese sentence.

(2) Among several translation candidates, the one that is close to the original sentence in syntactic structure is preferred. The aim is to avoid translating a sentence too freely, i.e., paraphrasing.

(3) For intelligible translation, information from the proceeding sentences should be added. Especially, a subject should be supplemented in Chinese while subjects are often omitted in Japanese.

(4) For a natural Chinese translation, text should be supplemented, deleted, replaced, and paraphrased when necessary. When a translation is very long, word order can be changed or commons can be inserted. These are the restrictions on (2). The naturalness of the Chinese translations is the priority.

Proper nouns in the newspaper articles, such as names of people and places and special things in Japan, created a problem for translation. We pay special attentions to them in the following way.

(1) Proper nouns

When words did not exist in available Japanese-Chinese dictionaries, new translations were created and then confirmed using the Chinese web. For kanji in Japanese words, if the same orthography existed in Chinese characters, the kanji was used directly in the Chinese translation; if the kanji was a traditional Chinese character, its simplified Chinese character was used in the translation; otherwise, the kanji was introduced into the Chinese characters and used in the translation directly.

(2) Special things in Japan

Explanations were added if necessary. For example, “大相扑”, translated from “大相撲” (grand sumo tournament), is well known in China, while “春斗”, translated from “春闘” (spring labor offensive), is not known in China. In this case, an explanation “春季劳资纠纷” was added behind the unfamiliar term. We think any word has its origin in a language, and we attempt to introduce new words about Japanese culture into Chinese through the construction of the corpus.

Producing high-quality Chinese translations is crucial to this parallel corpus. We controlled the quality by the following treatments.

(1) The first revision of a translated article was conducted by a different translator after the first translation. The reviewers checked whether the meanings of the Chinese translations corresponded accurately to the meanings of the original sentences and modified the Chinese translations if necessary.

(2) The second revision was conducted by Chinese natives without referring to the original sentences. The reviewers checked whether the Chinese translations were natural and passed the unnatural translations back to translators for modification.

(3) The third revision was conducted by a Chinese native in the annotation process of Chinese morphological information. The words that did not exist in the dictionary of contemporary Chinese were checked to determine whether they were new words. If not, the words were designated as informal or not written language and were replaced with suitable words. The word sequences that missed the Chinese language model’s part-of-speech chain were also adjusted.

Until now, 38,383 Japanese sentences have been translated to Chinese, and of those, 22,000 Chinese translations have been revised three times, and we are still working on the remaining 18,000 Chinese translations.

4 Morphological Information Annotation

Annotation consists of automatic analyses and manual revision.

4.1 Annotation on Japanese Sentences

Japanese morphological and syntactic analyses follow the definitions of part-of-speech categories and syntactic labels of the Corpus of Spontaneous Japanese (Maekawa, 2000).

A morphological analyzer developed in that project was applied for automatic annotation on the Japanese sentences and then the automatically tagged sentences were revised manually. An annotated senetence is illustrated in Figure 1, which is the Japanese sentence in Ex. 1 in Section 2.

```
# S-ID:950104141-008
＊ 0 2D
いずれも いずれも ＊ 副詞 ＊ ＊ ＊
＊ 1 2D
十九 じゅうきゅう ＊ 名詞 数詞 ＊ ＊
歳 さい ＊ 接尾辞 名詞性名詞助数辞 ＊ ＊
前後 ぜんご ＊ 接尾辞 名詞性名詞接尾辞 ＊ ＊
の の ＊ 助詞 接続助詞 ＊ ＊
＊ 2 6D
若者 わかもの ＊ 名詞 普通名詞 ＊ ＊
で で だ 判定詞 ＊ 判定詞 ダ列タ系連用テ形
、 、 ＊ 特殊 読点 ＊ ＊
＊ 3 4D
質問 しつもん ＊ 名詞 サ変名詞 ＊ ＊
に に ＊ 助詞 格助詞 ＊ ＊
＊ 4 5D
答える こたえる 答える 動詞 ＊ 母音動詞 基本形
＊ 5 6D
気力 きりょく ＊ 名詞 普通名詞 ＊ ＊
も も ＊ 助詞 副助詞 ＊ ＊
＊ 6 -1D
残って のこって 残る 動詞 ＊ 子音ラ行 タ系テ形
い い いる 接尾辞 動詞性接尾辞 母音動詞 未然形
ない ない ない 接尾辞 形容詞辞 イ形容段 基本形
。 。 ＊ 特殊 句点 ＊ ＊
EOJ
```

Figure 1. An annotated Japanese sentence

The data of one sentence begins from the line “# S-ID... ” and ends with the mark “EOJ”. The line headed by “*” indicates the beginning of a phrase and the following lines are morphemes in that phrase. For example, the line “＊ 0 2D” indicates the phrase whose number is 0. The following line “いずれも いずれも ＊ 副詞 ＊ ＊ ＊” indicates the morpheme in the phrase. There are seven fields in each morpheme line, token form, phonetic alphabet, dictionary form, part-of-speech, sub-part-of-speech, verbal category and conjugation form. In the line “＊ 0 2D”, the numeral 2 in “2D” indicates that the phrase 0 “いずれも” modifies the phrase 2 “若者で、”. The syntactic structure analysis adopts dependency-structure analysis in which modifier-modified relations between phrases are determined. The dependency-structure of the example in Figure 1 is demonstrated in Figure 2.

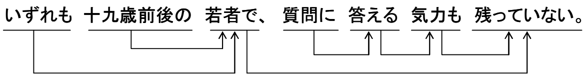

Figure 2 Example of syntactic structure

4.2 Annotation on Chinese Sentences

For Chinese morphological analysis, we used the analyser developed by Peking University, where the research on definition of Chinese words and the criteria of word segmentation has been conducted for over ten years. The achievements include a grammatical knowledge base of contemporary Chinese, an automatic morphological analyser, and an annotated People's Daily Corpus. Since the definition and tagset are widely used in Chinese language processing, we also took the criteria as the basis of our guidelines.

A morphological analyzer developed by Peking University (Zhou and Yu, 1994) was applied for automatic annotation of the Chinese sentences and then the automatically tagged sentences were revised by humans. An annotated sentence is illustrated in Figure 3, which is the Chinese sentence in Ex. 1 in Section 2.

S-ID: 950104141-008
这些/r 俄军/j 士兵/n 均/d 为/v 十九/m 岁/q
左右/m 的/u 年青人/n ，/w 他们/r 甚至/d
连/p 回答/v 问题/n 的/u 气力/n 也/d
没有/v 。/w

Figure 3 An annotated Chinese sentence

4.3 Tool for Manual Revision

We developed a tool to assist annotators in revision. The tool has both Japanese and Chinese versions. Here, we introduce the Chinese version. The input of the tool is the automatically segmented and part-of-speech tagged sentences and the output is revised data. The basic functions include separating a sequence of characters into two words, combining two segmented words into one word, and selecting a part-of-speech for a segmented word from a list of parts-of-speech. In addition, the tool has the following functions.

(1) Retrieves a word in the grammatical knowledge base of contemporary Chinese of Peking University (Yu et al., 1997).

This is convenient when annotators want to confirm whether a segmented word is authorized by the grammatical knowledge base, and when they want to know the parts-of-speech of a word defined by the grammatical knowledge base.

(2) Retrieves a word in other annotated corpora or the sentences that have been revised.

This is convenient when annotators want to see how the same word has been annotated before.

(3) Retrieves a word in the current file.

It collects all the sentences in the current file that contain the same word and then sorts their context on the left and right of the word. By referring to the sorted contexts, annotators can select words with the same syntactic roles and change all of the parts-of-speech to a certain one all in one operation. This is convenient when annotators want to process the same word in different sentences, aiming for consistency in annotation.

(4) Adds new words to the grammatical knowledge base dynamically.

The updated grammatical knowledge base can be used by the morphological analyser in the next analysis.

(5) Indexes to sentences by an index file.

The automatically discovered erroneous annotations can be stored in one index file, pointing to the sentences that are to be revised.

The interface of the tool is shown in Figure 4 and Figure 5.

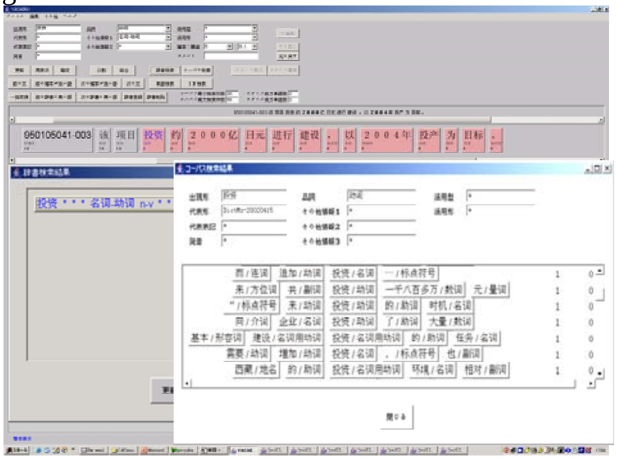
```
而/连词  追加/动词  投资/名词  —/标点符号                    1  0
未/方位词  共/副词  投资/动词  一千八百多万/数词  元/量词    1  0
"/标点符号  未/动词  投资/动词  的/助词  时机/名词            1  0
向/介词  企业/名词  投资/动词  了/助词  大量/数词              1  0
基本/形容词  建设/名词用动词  投资/名词用动词  的/助词  任务/名词  1  0
需要/动词  增加/动词  投资/名词  ，/标点符号  也/副词          1  0
西藏/地名  的/助词  投资/名词用动词  环境/名词  相对/副词      1  0
```

Figure 4 Interface of the manual revision tool (Retrieves a word in the grammatical knowledge base of contemporary Chinese)

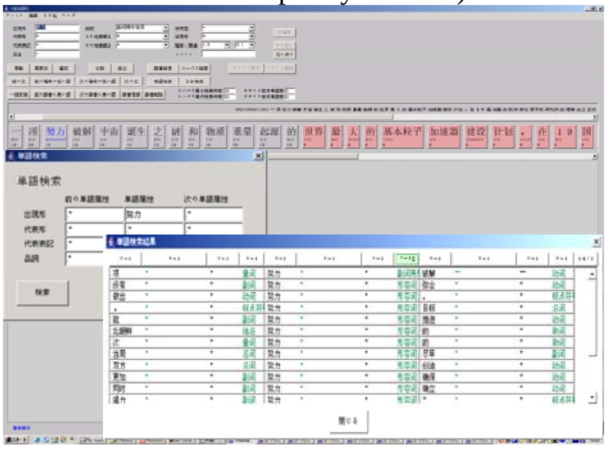

Figure 5 Interface of the manual revision tool (Retrieves a word in the current file)

In Figure 4, the small window in the lower left displays the retrieved result of the word "投资" in the grammatical knowledge base; the lower right window displays the retrieved result of the same word in the annotated People's Daily Corpus.

In Figure 5, the small window in the lower left is used to define retrieval conditions in the current file. In this example, the orthography of "努力" is defined. The lower right window displays the sentences containing the word "努力" retrieved from the current file. The left and right contexts of one word are shown with the retrieved word. The contents of any column can be sorted by clicking the top line of the column.

5 Annotation of word alignment

Since completely manual alignment is time consuming and expensive, automatic word alignment is required for assistance, although the techniques cannot reach as high a level as the morphological analyses. Therefore, we should determine the potential performance of automatic techniques and then combine automatic word alignment with manual alignment. In this section, we first evaluate two approaches to word alignment: a lexical knowledge-based approach proposed by us and a statistics-based approach, the well-known GIZA++ toolkit. We then present a practical method of using multiple aligners. After that, we describe an assisting tool developed for manual alignment.

5.1 Lexical-knowledge based approach

We propose a lexical knowledge-based approach for the Japanese-Chinese parallel corpus. In the 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 the above section. Let W_J and W_C denote their word lists, respectively. The approach consists of two algorithms: Algorithm 1 for establishing reliable alignments and Algorithm 2 for broading coverage.

5.1.1 Algorithm for establishing reliable alignment

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

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

where $|x|$ ($|y|$) is the number of morphemes in x (y), and $|x \cap y|$ is the number of 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, j may be 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 experiment. Hereafter, we use $\ddot{c}(i,k)$ to express a 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. Based on this phenomenon, we assume that j in W_J and $\ddot{c}$ in W_C probably have a translation relation if their orthographies are similar, and j is therefore probably aligned with $\ddot{c}$. Formula (2) 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 Mandarin, simplified Chinese characters are used. At the same time, many Japanese kanji words maintain the traditional Chinese characters as they were when they were introduced from China. Based on this phenomenon, we assume that j and $\ddot{c}$ probably have a translation relation if j and the traditional form of $\ddot{c}$ are similar, and therefore j is probably aligned with $\ddot{c}$. Let $Trad(\ddot{c})$ denote the traditional form of $\ddot{c}$ by converting each simplified character of $\ddot{c}$ into a traditional character. Formula (3) is defined to estimate the possibility of j being aligned with $\ddot{c}$.

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

$Poss_{tra}$ expresses the possibility estimated by using the traditional Chinese characters.

Bilingual Dictionary

A translation dictionary can help to identify the translation relations. Let C_j denote the Chinese translation set of j. We can estimate the possibility of j being aligned with $\ddddot{c}$ using the following formula (Ker and Chang, 1997).

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

$Poss_{dic}$ expresses the possibility estimated by using a translation dictionary. An automatically built Japanese-Chinese dictionary is used here, which was built from EDR Japanese-English Dictionary (NICT, 2002) and LDC English-Chinese Dictionary (LDC, 2002) by using English as an intermediary (Zhang et al., 2005).

From the three possibilities estimated as above, the largest one is selected as follows. It is denoted as $Poss_{lex}$.

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

Algorithm 1 is described below.

Algorithm 1 (Establishing Reliable Alignment)

Step 1. For all $\ddddot{c}$ in W_C, get $Trad(\ddddot{c})$ 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 $\ddddot{c}$ in W_C, calculate $Poss_{ort}(j,\ddddot{c})$ using formula (2), $Poss_{tra}(j,\ddddot{c})$ using formula (3), $Poss_{dic}(j,\ddddot{c})$ using formula (4), and $Poss_{lex}(j,\ddddot{c})$ using formula (5).

Step 4. For each j in W_J, if

$$\max_{\ddddot{c}\ in\ W_C,\ 1 \le i \le |W_C|,\ 0 \le k \le 4} Poss_{lex}(j, \ddddot{c}(i,k)) \ge \theta_{lex},$$

output

$$(j, \hat{\ddddot{c}})\ (\hat{\ddddot{c}}(\hat{i}, \hat{k}) = \arg\max_{\ddddot{c}\ in\ W_C,\ 1 \le i \le |W_C|,\ 0 \le k \le 4} Poss_{lex}(j, \ddddot{c}(i,k)))$$

to A_{rel}, where θ_{lex} is a preset threshold.

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.

5.1.2 Algorithm for broadening coverage

Let $\overline{W_J}$ ($\overline{W_C}$) denote the list of words $j \in W_J (c \in W_C)$ that are still not aligned. In this phase, we only consider one to one alignment. 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.

For an alignment candidate $(\tilde{j}, \tilde{c})$, we estimate its likelihood by taking the established alignments into account. Here we consider four established alignments: 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.

First, add ($Null_0, Null_0$) and ($Null_{|W_J|+1}, Null_{|W_C|+1}$) to A_{rel} as the leftmost and rightmost alignments.

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

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

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

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

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

Then calculate the degrees at which $\tilde{j}$ and $\tilde{c}$ dislocate from the four established reliable alignments as follows.

$$\Delta m_{\tilde{j}_L} = m(\tilde{j}) - m(j_{\tilde{j}_L}), \Delta n_{\tilde{j}_L} = n(\tilde{c}) - n_e(c_{\tilde{j}_L}), (6)$$
$$\Delta m_{\tilde{j}_R} = m(\tilde{j}) - m(j_{\tilde{j}_R}), \Delta n_{\tilde{j}_R} = n(\tilde{c}) - n_s(c_{\tilde{j}_R}),$$
$$\Delta m_{\tilde{c}_L} = m(\tilde{j}) - m(j_{\tilde{c}_L}), \Delta n_{\tilde{c}_L} = n(\tilde{c}) - n_e(c_{\tilde{c}_L}), \text{and}$$
$$\Delta m_{\tilde{c}_R} = m(\tilde{j}) - m(j_{\tilde{c}_R}), \Delta n_{\tilde{c}_R} = n(\tilde{c}) - n_s(c_{\tilde{c}_R}),$$

where $m(\tilde{j}) (n(\tilde{c}))$ denotes the position of $\tilde{j}$ ($\tilde{c}$) in W_J (W_C) and $n_s(\hat{\ddddot{c}})$ ($n_e(\hat{\ddddot{c}})$) denotes the starting (ending) position of $\hat{\ddddot{c}}$ in W_C.

Third, estimate the possibility of $\tilde{j}$ being aligned with $\tilde{c}$ by referring to the alignments

$a_{\tilde{j}_L}$ as follows (Deng, 2004; Liu, 2004).

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

The first item in the denominator lays penalty using the degree at which $\tilde{j}$ and $\tilde{c}$ dislocate from the 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 the reliable alignment in an opposite direction. When $\tilde{j}$ and $\tilde{c}$ dislocate from the reliable alignment in an opposite direction, the possibility is smaller. When $\tilde{j}$ and $\tilde{c}$ dislocate from the 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 the reliable alignment in the same direction or not is thought more important.

Similarly, we can estimate the possibility by referring to the alignments $a_{\tilde{j}_R}$, $a_{\tilde{c}_L}$, and $a_{\tilde{c}_R}$ in the same way.

Finally, from the four possibilities estimated, select the largest one as follows. It is denoted as $Poss_{dis}$.

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

Algorithm 2 is described below.

Algorithm 2 (Broadening Coverage)

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.

5.2 Performance Evaluation

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. Experimental results obtained in this work gave us new insights on aligning words in a Japanese-Chinese parallel corpus.

The test data was 1,127 sentence pairs with gold standards, totalling 17,332 alignments. Thresholds in Algorithms 1 and 2 were empirically 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).

We also applied a statistics-based approach, the well-known toolkit, GIZA++. Two directions were tested: the Chinese sentences were used as source sentences and the Japanese sentences as target sentences, and vice versa.

The results were evaluated in terms of three measures, Precision, Recall and F-measure. The evaluation results are shown in Table 2.

Method	Precision (%)	Recall (%)	F-measure
Proposed approach	69	58	63.02
GIZA(C→J)	54.8	72.8	63
GIZA(J→C)	45.5	54.6	50

Table 2 Evaluation results of the proposed approach and GIZA++

The performance of C→J was better than that of J →C in GIZA++. Compared with C→J of GIZA++, the proposed approach achieved the same performance in F-measure, but with higher precision and a lower recall rate. No one approach could achieve a superior performance, so we 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	79.3	62.7	70

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 assisting 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.

5.3 Manual Alignment Tool

We developed a manual alignment tool, which consist of a graphical interface and internal data management. Annotators can correct the output of the automatic aligner and add alignments that it has not identified. In addition to assisting with word alignment, the tool also supports annotation on phrase alignment. Since Japanese sentences have been annotated with phrase structures, the tool can display a Japanese sentence in phrase units. Annotators can select each phrase on the Japanese side and then align them with words on the Chinese side. For idioms in Japanese sentences, two or more phrases can be selected.

The input and output file is in XML format. The data of one sentence pair consists of the Chinese sentence annotated with morphological information, the Japanese sentence annotated with morphological information, and the Japanese sentence annotated with syntactic structure, word alignment, and phrase alignment.

The alignment annotation at word and phrase is ongoing, the former focusing on lexical translations and the latter focusing on pattern translations. After a certain amount of data is annotated, we plan to exploit the annotated data to further improve the performance of automatic word alignment. We will also investigate a method to automatically identify phrase alignments from the annotated word alignment and a method to automatically discover the syntactic structures on the Chinese side from the annotated phrase alignments.

6 Conclusion

We have described the construction of a Japanese-Chinese parallel corpus, a part of the NICT Multilingual Corpus. The corpus consists of about 40,000 pairs of Japanese sentences and their Chinese translations. The Japanese sentences are annotated with morphological and syntactic structures and the Chinese sentences are annotated with morphological information. In addition, word and phrase alignments are annotated. A high quality of annotation was obtained through manual revisions. To the best of our knowledge, this will be the first annotated Japanese-Chinese parallel corpus in the world.

References

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

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

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

Liu Q. 2004. *Research into some aspects of Chinese-English machine translation.* Doctoral Dissertation.

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

LDC. 1992. *Linguistic data Consortium.* http://www.ldc.upenn.edu/.

Uchimoto, K. and Zhang,Y., Sudo, K., Murata, M., and Sekine, S., Isahara, H. *Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information* and Its Applications. Proceedings of the MLR2004: PostCOLING Workshop on Multilingual Linguistic Resources, pages 63-70.

Yamada, K., Knight, K. 2001.A syntax-based Statistical Translation Model. In Proceedings of the ACL , pages 523-530.

Yu, Shiwen. 1997. *Grammatical Knowledge Base of Contemporary Chinese*. Tsinghua Publishing Company.

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.