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, uchimoto,qma, 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 the scheme of detailed information annotation, and the tools we developed in the corpus construction. 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. So, 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. 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 translators 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. We finish the task by 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 in revision. The annotation task for each language included morphological and syntactic structure annotation. The annotation task for each language pair included alignments at word and phrase level.

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

(1) Since the original data is from newspaper and journals, the domain of each corpus is therefore rich.

(2) Each corpus consists of original sentences and their translations, so they are already sentence aligned. In translation of each sentence, the context of the article is also considered. Thus, the context of each original article is also well maintained in its translation, which can be exploited in the future.

(3) 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) To obtain intelligible Chinese translations, information of the proceeding sentences in the same article should be added. Especially, a subject should be supplemented because a subject is usually required in Chinese, while in Japanese subjects are often omitted .

(4) To obtain natural Chinese translations, supplement, deletion, replacement, and paraphrase should be made when necessary. When a translation is very long, word order can be changed or commons can be inserted. These are the restrictions on (2), i.e., the naturalness of the Chinese translations is the priority.

One problem in translation is how to translate proper nouns in the newspaper articles. We pay special attentions to them in the following way.

(1) Proper nouns

When proper nouns did not exist in Japanese-Chinese dictionaries, new translations were created and then confirmed using the Chinese web. For kanji in proper nouns, if there was a Chinese character having the same orthography as the kanji, the Chinese character was used in the Chinese translation; if there was a traditional Chinese character having the same orthography as the kanji, the simplified character of the traditional Chinese character was used in the translation; otherwise, a Chinese character whose orthography is similar to that of the kanji was used in the translation.

(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 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.

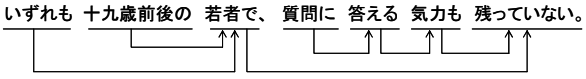

Figure 2 Example of syntactic structure

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.

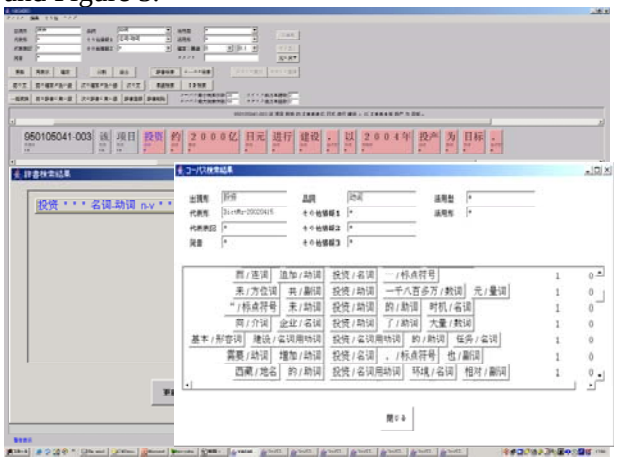

Figure 4 Interface of the manual revision tool (Retrieves a word in the grammatical knowledge base of contemporary Chinese)

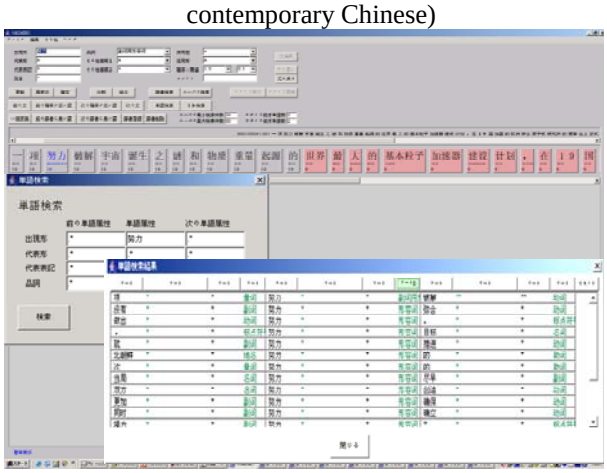

Figure 5 Interface of the manual revision tool (Retrieves a word in the current file)

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 automatic word alignment techniques cannot reach as high a level as the morphological analyses, we adopt a practical method of using multiple aligners. One aligner is a lexical knowledge-based approach, which was implemented by us based on the work of Ker (Ker and Chang, 1997). Another aligner is the well-known GIZA++ toolkit, which is a statistics-based approach. For GIZA++, two directions were adopted: the Chinese sentences were used as source sentences and the Japanese sentences as target sentences, and vice versa.

The results produced by the lexical knowledge-based aligner, 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 2 showed the evaluation results of the multi-aligner on 1,127 test sentence pairs, which were manually annotated with gold standards, totally 17,332 alignments.

	Precision (%)	Recall (%)	F-measure
Multi-aligner	79.3	62.7	70

Table 2 Evaluation results of the multi-aligner

The multi-aligner produced satisfactory results. This performance is evidence that the multi-aligner is feasible for use in assisting word alignment annotation.

For manual revision, we also developed an assisting 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, 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 of the manual annotation 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 and syntactic structure information, 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 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, which were greatly assisted by the revision tools we developed in the project. To the best of our knowledge, this will be the first annotated Japanese-Chinese parallel corpus in the world.

In the future, we will finish the annotation on the remaining data and add syntactic structures to the Chinese sentences.

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.

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.