Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information and Its Applications

Kiyotaka Uchimoto[†] **Yujie Zhang**[†] **Kiyoshi Sudo**[‡]

Masaki Murata[†] **Satoshi Sekine**[‡] **Hitoshi Isahara**[†]

[†]National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun,
Kyoto 619-0289, Japan
{uchimoto,yujie,murata,isahara}@nict.go.jp

[‡]New York University
715 Broadway, 7th floor
New York, NY 10003, USA
{sudo,sekine}@cs.nyu.edu

Abstract

This paper describes Japanese-English-Chinese aligned parallel treebank corpora of newspaper articles. They have been constructed by translating each sentence in the Penn Treebank and the Kyoto University text corpus into a corresponding natural sentence in a target language. Each sentence is translated so as to reflect its contextual information and is annotated with morphological and syntactic structures and phrasal alignment. This paper also describes the possible applications of the parallel corpus and proposes a new framework to aid in translation. In this framework, parallel translations whose source language sentence is similar to a given sentence can be semi-automatically generated. In this paper we show that the framework can be achieved by using our aligned parallel treebank corpus.

1 Introduction

Recently, accurate machine translation systems can be constructed by using parallel corpora (Och and Ney, 2000; Germann et al., 2001). However, almost all existing machine translation systems do not consider the problem of translating a given sentence into a natural sentence reflecting its contextual information in the target language. One of the main reasons for this is that we had many problems that had to be solved by one-sentence to one-sentence machine translation before we could solve the contextual problem. Another reason is that it was difficult to simply investigate the influence of the context on the translation because sentence correspondences of the existing bilingual documents are rarely one-to-one, and are usually one-to-many or many-to-many.

On the other hand, high-quality treebanks such as the Penn Treebank (Marcus et al., 1993) and the Kyoto University text corpus (Kurohashi and Nagao, 1997) have contributed to improving the accuracies of fundamental techniques for natural language processing such as morphological analysis and syntactic structure analysis. However, almost all of these high-quality treebanks are based on monolingual corpora and do not have bilingual or multilingual information. There are few high-quality bilingual or multilingual treebank corpora because parallel corpora have mainly been actively used for machine translation between related languages such as English and French, therefore their syntactic structures are not required so much for aligning words or phrases. However, syntactic structures are necessary for machine translation between languages whose syntactic structures are different from each other, such as in Japanese-English, Japanese-Chinese, and Chinese-English machine translations, because it is more difficult to automatically align words or phrases between two unrelated languages than between two related languages. Actually, it has been reported that syntactic structures contribute to improving the accuracy of word alignment between Japanese and English (Yamada and Knight, 2001). Therefore, if we had a high-quality parallel treebank corpus, the accuracies of machine translation between languages whose syntactic structures are different from each other would improve. Furthermore, if the parallel treebank corpus had word or phrase alignment, the accuracy of automatic word or phrase alignment would increase by using the parallel treebank corpus as training data. However, so far, there is no aligned parallel treebank corpus whose domain is not restricted. For example, the Japanese Electronics Industry Development Association's (JEIDA's) bilingual corpus (Isahara and Haruno, 2000) has sentence, phrase, and proper noun alignment. However, it does not have morphological and syntactic information, the alignment is partial, and the target is restricted to a white paper. The Advance Telecommunications Research dialogue database (ATR, 1992) is a parallel treebank corpus between Japanese and English. However, it does not have word or phrase alignment, and the target domain is restricted to travel conversation.

Therefore, we have been constructing aligned parallel treebank corpora of newspaper articles between languages whose syntactic structures are different from each other since 2001; they

meet the following conditions.

1. It is easy to investigate the influence of the context on the translation, which means the sentences that come before and after a particular sentence, and that help us to understand the meaning of a particular word such as a pronoun.
2. The annotated information in the existing monolingual high-quality treebanks can be utilized.
3. They are open to the public.

To construct parallel corpora that satisfy these conditions, each sentence in the Penn Treebank (Release 2) and the Kyoto University text corpus (Version 3.0) has been translated into a corresponding natural sentence reflecting its contextual information in a target language by skilled translators, revised by native speakers, and each parallel translation has been annotated with morphological and syntactic structures, and phrasal alignment. Henceforth, we call the parallel corpus that is constructed by pursuing the above policy an *aligned parallel treebank corpus reflecting contextual information*. In this paper, we describe an aligned parallel treebank corpus of newspaper articles between Japanese, English, and Chinese, and its applications.

2 Construction of Aligned Parallel Treebank Corpus Reflecting Contextual Information

2.1 Human Translation of Existing Monolingual Treebank

The Penn Treebank is a tagged corpus of Wall Street Journal material, and it is divided into 24 sections. The Kyoto University text corpus is a tagged corpus of the *Mainichi* newspaper, which is divided into 16 sections according to the categories of articles such as the sports section and the economy section. To maintain the consistency of expressions in translation, a few particular translators were assigned to translate articles in a particular section, and the same translator was assigned to the same section. The instructions to translators for Japanese-English translation is basically as follows.

1. One-sentence to one-sentence translation as a rule

   Translate a source sentence into a target sentence. In case the translated sentence becomes unnatural by pursuing this policy, leave a comment.
2. Natural translation reflecting contextual information

   Except in the case that the translated sentence becomes unnatural by pursuing policy 1, translate a source sentence into a target sentence naturally.

   By deletion, replacement, or supplementation, let the translated sentence be natural in the context.

   In an entire article, the translated sentences must maintain the same meaning and information as those of the original sentences.
3. Translations of proper nouns

   Find out the translations of proper nouns by looking up the nouns in a dictionary or by using a web search. In case a translation cannot be found, use a temporary name and report it.

We started the construction of Japanese-Chinese parallel corpus in 2002. The Japanese sentences of the Kyoto University text corpus were also translated into Chinese by human translators. Then each translated Chinese sentence was revised by a second Chinese native. The instruction to the translators is the same as that given in the Japanese-English human translations.

The breakdown of the parallel corpora is shown in Table 1. We are planning to translate the remaining 18,714 sentences of the Kyoto University text corpus and the remaining 30,890 sentences of the Penn Treebank. As for the naturalness of the translated sentences, there are 207 (1%) unnatural English sentences of the Kyoto University text corpus, and 462 (2.5%) unnatural Japanese sentences of the Penn Treebank generated by pursuing policy 1.

2.2 Morphological and Syntactic Annotation

In the following sections, we describe the annotated information of the parallel treebank corpus based on the Kyoto University text corpus.

2.2.1 Morphological and Syntactic Information of Japanese-English corpus

Translated English sentences were analyzed by using the Charniak Parser (Charniak, 1999). Then, the parsed sentences were manually revised. The definitions of part-of-speech (POS) categories and syntactic labels follow those of the Treebank I style (Marcus et al., 1993). We have finished revising the 10,328 parsed sentences that appeared from January 1st to 11th. An example of morphological and syntactic structures is shown in Figure 1. In this figure, "S-ID" means the sentence ID in the Kyoto University text corpus. EOJ means the boundary between a Japanese parsed sentence and an English parsed sentence. The definition of Japanese morphological and syntactic information follows that of the Kyoto University text corpus (Version 3.0). The syntactic structure is represented by dependencies between Japanese phrasal units called *bunsetsu*s. The phrasal

Table 1: Breakdown of the parallel corpora

Original corpus	Languages	# of parallel sentences
Kyoto University text corpus	Japanese-English	19,669 (from Jan. 1st to 17th in 1995)
	Japanese-Chinese	38,383 (all)
Penn Treebank	Japanese-English	18,318 (from section 0 to 9)
Total	Japanese-English	37,987 (Approximately 900,000 English words)
	Japanese-Chinese	38,383 (Approximately 900,000 Chinese words)

```
# 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 * 子音動詞ラ行 タ系連用テ形
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 形容詞性述語接尾辞 イ形容詞アウオ段 基本形
。 。 * 特殊 句点 * *
EOJ
(S1 (S (NP (PRP They))
       (VP (VP (VBD were)
               (NP (DT all))
               (ADJP (NP (QP (RB about)
                             (CD nineteen))
                         (NNS years))
                     (JJ old)))
           (CC and)
           (VP (VBD had)
               (S (NP (DT no)
                      (NN strength))
                  (VP (VBN left)
                      (SBAR (S (VP (ADVP (RB even))
                                   (TO to)
                                   (VP (VB answer)
                                       (NP (NNS questions))))))))))
       (. .)))
EOE
```

Figure 1: Example of morphological and syntactic information.

units or *bunsetsu*s are minimal linguistic units obtained by segmenting a sentence naturally in terms of semantics and phonetics, and each of them consists of one or more morphemes.

2.2.2 Chinese Morphological Information of Japanese-Chinese corpus

Chinese sentences are composed of strings of Hanzi and there are no spaces between words. The morphological annotation, therefore, includes providing tags of word boundaries and POSs of words. We analyzed the Chinese sentences by using the morphological analyzer developed by Peking University (Zhou and Duan, 1994). There are 39 categories in this POS set. Then the automatically tagged sentences were revised by the third native Chinese. In this pass the Chinese translations were revised again while the results of word segmentation and POS tagging were revised. Therefore the Chinese translations are obtained with a high quality. We have finished revising the 12,000 tagged sentences. The revision of the remaining sentences is ongoing. An example of tagged Chinese sentences is shown in Figure 2. The letters shown

```
S-ID:950104141-008
这些(ZheXie)/r
俄军(EJun)/j
士兵(ShiBing)/n
均(Jun)/d
为(Wei)/v
十九(ShiJiu)/m
岁(Sui)/q
左右(ZuoYou)/m
的(De)/u
年青人(NianQingRen)/n
,/w
他们(TaMen)/r
甚至(ShenZhi)/d
连(Lian)/p
回答(HuiDa)/v
问题(WenTi)/n
的(De)/u
气力(QiLi)/n
也(Ye)/d
没有(MeiYou)/v
。/w
```

Figure 2: Example of morphological information of Chinese corpus.

after '/' indicate POSs. The Chinese sentence is the translation of the Japanese sentence in Figure 1. The Chinese sentences are GB encoded. The 38,383 translated Chinese sentences have 1,410,892 Hanzi and 926,838 words.

2.3 Phrasal Alignment

This section describes the annotated information of 19,669 sentences of the Kyoto University text corpus.

The minimum alignment unit should be as small as possible, because bigger units can be constructed from units of the minimum size. However, we decided to define a *bunsetsu* as the minimum alignment unit. One of the main reasons for this is that the smaller the unit is, the higher the human annotation cost is. Another reason is that if we define a word or a morpheme as a minimum alignment unit, expressions such as post-positional particles in Japanese and articles in English often do not have alignments. To

effectively absorb those expressions and to align as many parts as possible, we found that a bigger unit than a word or a morpheme is suitable as the minimum alignment unit. We call the minimum alignment based on *bunsetsu* alignment units the *bunsetsu unit translation pair*. Bigger pairs than the *bunsetsu* unit translation pairs can be automatically extracted based on the *bunsetsu* unit translation pairs. We call all of the pairs, including *bunsetsu* unit translation pairs, *translation pairs*. The *bunsetsu* unit translation pairs for idiomatic expressions often become unnatural. In this case, two or more *bunsetsu* units are combined and handled as a minimum alignment unit. The breakdown of the *bunsetsu* unit translation pairs is shown in Table 2.

Table 2: Breakdown of the *bunsetsu* unit translation pairs.

(1) total # of translation pairs	172,255
(2) # of different translation pairs	146,397
(3) # of Japanese expressions	110,284
(4) # of English expressions	111,111
(5) average # of English expressions corresponding to a Japanese expression	1.33 ((2)/(3))
(6) average # of Japanese expressions corresponding to a English expression	1.32 ((2)/(4))
(7) # of ambiguous Japanese expressions	15,699
(8) # of ambiguous English expressions	12,442
(9) # of *bunsetsu* unit translation pairs consisting of two or more *bunsetsu*s	17,719

An example of phrasal alignment is shown in Figure 3. A Japanese sentence is shown from the line after the S-ID to the EOJ. Each line indicates a *bunsetsu*. Each rectangular line indicates a dependency between *bunsetsu*s. The leftmost number in each line indicates the *bunsetsu* ID. The corresponding English sentence is shown in the next line after that of the EOJ (End of Japanese) until the EOE (End of English). The English expressions corresponding to each *bunsetsu* are tagged with the corresponding *bunsetsu* ID such as <P id="*bunsetsu* ID"></P>. When there are two or more figures in the tag id such as id="1,2", it means two or more *bunsetsu*s are combined and handled as a minimum alignment unit.

For example, we can extract the following translation pairs from Figure 3.

- (J) 輸入が (*yunyuu-ga*) / 解禁された (*kaikin-sa-reta*); (E)that had been under the ban
- (J) 米国産リンゴの (*beikoku-san-ringo-no*); (E)of apples imported from the U.S.
- (J) 第１便が (*dai-ichi-bin-ga*); (E)The first cargo
- (J) 売り出された。(*uridasa-reta*); (E)was brought to the market.
- (J) 米国産リンゴの (*beikoku-san-ringo-no*) / 第１便が (*dai-ichi-bin-ga*); (E)The first cargo / of apples imported from the U.S.

```
# S-ID:950110003-001
1       輸入が┐
2     解禁された┐
3   米国産リンゴの┐
4           第１便が──┐
5               ９日、─┤
6           検疫手続きを┤
7                 終え、────P
8           首都圏の┐    │
9           大手スーパーなどで─┤
10                    初めて┤
11                売り出された。
EOJ
<P id="4">The first cargo</P> <P id="3">of apples
imported from the U.S.</P> <P id="1,2">that had been
under the ban</P> <P id="7">completed</P> <P id="6">
quarantine</P> <P id="7">and</P> <P id="11">was brought
to the market</P> <P id="10">for the first time</P>
<P id="5">on the 9th</P> <P id="9">at major supermarket
chain stores</P> <P id="8">in the Tokyo metropolitan
area</P> <P id="11">.</P>
EOE
```

Figure 3: Example of phrasal alignment.

- (J) 米国産リンゴの (*beikoku-san-ringo-no*) / 第１便が (*dai-ichi-bin-ga*) / 売り出された。(*uridasa-reta*); (E)The first cargo / of apples imported from the U.S. / was brought to the market.

Here, Japanese and English expressions are divided by the symbol ";", and "/" means a *bunsetsu boundary*.

An overview of the criteria of the alignment is as follows. Align as many parts as possible, except if a certain part is redundant. More detailed criteria will be attached with our corpus when it is open to the public.

1. Alignment of English grammatical elements that are not expressed in Japanese

   English articles, possessive pronouns, infinitive *to*, and auxiliary verbs are joined with nouns and verbs.
2. Alignment between a noun and its substitute expression

   A noun can be aligned with its substitute expression such as a pronoun.
3. Alignment of Japanese ellipses

   An English expression is joined with its related elements. For example, the English subject is joined with its related verb.
4. Alignment of supplementary or explanatory expression in English

   Supplementary or explanatory expressions in English are joined with their related words.

■ Ex.：

```
# S-ID:950104142-003
1   「佳」には──────┐
2     「美しい」P     │
3     「立派な」と┐   │
4               いう┐ │
5               意味が┤
6                 ある。
EOJ
<P id="1">The Chinese character used for "ka"</P>
```

```
has such meanings as "beautiful" and "splendid."
EOE
・"「佳 (ka)」には (niwa)" corresponds to
  "The Chinese character used for "ka""
```

5. Alignment of date and time

   When a Japanese noun representing date and time is adverbial, the English preposition is joined with the date and time.

6. Alignment of coordinate structures

   When English expressions represented by "X (A + B)" correspond to Japanese expressions represented by "XA + XB", the alignment of X overlaps.

```
■ Ex.：
# S-ID:950106149-005
1   近畿圏では──────────┐
2       尼崎沖でＩ          │
3   八九年度から、ーＰ      │
4         泉大津沖でＩ      │
5       九一年度から、──┤
6           廃棄物などの┐  │
7                 投棄が┤
8               始まった。
EOJ
In the Kinki Region, disposal of wastes started
<P id="2"><P id="4"> at offshore sites of</P>
Amagasaki</P> and <P id="4">Izumiotsu</P> from
1989 and 1991 respectively.
EOE
・"尼崎沖 (Amagasaki-oki) で (de)" corresponds to
"at offshore sites of Amagasaki"
・"泉大津沖 (Izumiotsu-oki) で (de)" corresponds to
"at offshore sites of … Izumiotsu"
```

3 Applications of Aligned Parallel Treebank Corpus

3.1 Use for Evaluation of Conventional Methods

The corpus as described in Section 2 can be used for the evaluation of English-Japanese and Japanese-English machine translation. We can directly compare various methods of machine translation by using this corpus. It can be summarized as follows in terms of the characteristics of the corpus.

One-sentence to one-sentence translation can be simply used for the evaluation of various methods of machine translation.

Morphological and syntactic information can be used for the evaluation of methods that actively use morphological and syntactic information, such as methods for example-based machine translation (Nagao, 1981; Watanabe et al., 2003), or transfer-based machine translation (Imamura, 2002).

Phrasal alignment is used for the evaluation of automatically acquired translation knowledge (Yamamoto and Matsumoto, 2003).

An actual comparison and evaluation is our future work.

3.2 Analysis of Translation

One-sentence to one-sentence translation reflects contextual information. Therefore, it is suitable to investigate the influence of the context on the translation. For example, we can investigate the difference in the use of demonstratives and pronouns between English and Japanese. We can also investigate the difference in the use of anaphora.

Morphological and syntactic information and phrasal alignment can be used to investigate the appropriate unit and size of translation rules and the relationship between syntactic structures and phrasal alignment.

3.3 Use in Conventional Systems

One-sentence to one-sentence translation can be used for training a statistical translation model such as GIZA++ (Och and Ney, 2000), which could be a strong baseline system for machine translation.

Morphological and syntactic information and phrasal alignment can be used to acquire translation knowledge for example-based machine translation and transfer-based machine translation.

In order to show what kind of units are helpful for example-based machine translation, we investigated whether the Japanese sentences of newspaper articles appearing on January 17, 1995, which we call test-set sentences, could be translated into English sentences by using translation pairs appearing from January 1st to 16th as a database. First, we found that only one out of 1,234 test-set sentences agreed with one out of 18,435 sentences in the database. Therefore, a simple sentence search will not work well. On the other hand, 6,659 *bunsetsu*s out of 12,632 *bunsetsu*s in the test-set sentences agreed with those in the database. If words in *bunsetsu*s are expanded into their synonyms, the combination of the expanded *bunsetsu*s sets in the database may cover the test-set sentences. Next, therefore, we investigated whether the Japanese test-set sentences could be translated into English sentences by simply combining translation pairs appearing in the database. Given a Japanese sentence, words were extracted from it and translation pairs that include those words or their synonyms, which were manually evaluated, were extracted from the database. Then, the English sentence was manually generated by just combining English expressions in the extracted translation pairs. One hundred two relatively short sentences (the average number of *bunsetsu*s is about 9.8) were selected as inputs. The number of equivalent translations, which mean that the translated sentence is grammatical and has the same meaning as the source

sentence, was 9. The number of similar translations, which mean that the translated sentence is ungrammatical, or different or wrong meanings of words, tenses, and prepositions are used in the translated sentence, was 83. The number of other translations, which mean that some words are missing, or the meaning of the translated sentence is completely different from that of the original sentence, was 10. For example, the original parallel translation is as follows:

Japanese: さきがけ側は通常国会に向け、政策や国会運営をテーマとする協議機関を両党に設置することを提案した。
English: New Party Sakigake proposed that towards the ordinary session, both parties found a council to discuss policy and Diet management.

Given the Japanese sentence, the translated sentence was:

Translation: Sakigake Party suggested to set up an organization between the two parties towards the regular session of the Diet to discuss under the theme of policies and the management of the Diet.

This result shows that only 9% of input sentences can be translated into sentences equivalent to the original ones. However, we found that approximately 90% of input sentences can be translated into English sentences that are equivalent or similar to the original ones.

3.4 Similar Parallel Translation Generation

The original aim of constructing an aligned parallel treebank corpus as described in Section 2 is to achieve a new framework for translation aid as described below.

It would be very convenient if multilingual sentences could be generated by just writing sentences in our mother language. Today, it can be formally achieved by using commercial machine translation systems. However, the automatically translated sentences are often incomprehensible. Therefore, we have to revise the original and translated sentences by finding and referring to parallel translation whose source language sentence is similar to the original one. In many cases, however, we cannot find such similar parallel translations to the input sentence. Therefore, it is difficult for users who do not have enough knowledge of the target languages to generate comprehensible sentences in several languages by just searching similar parallel translations in this way. Therefore, we propose to generate similar parallel translations whose source language sentence is similar to the input sentence. We call this framework for translation aid *similar parallel translation generation*.

We investigated whether the framework can be achieved by using our aligned parallel treebank corpus. As the first step of this study, we investigated whether an appropriate parallel translation can be generated by simply combining translation pairs extracted from our aligned parallel treebank corpus in the following steps.

1. Extract each content word with its adjacent function word in each *bunsetsu* in a given sentence
2. The extracted content words and their adjacent function words are expanded into their synonyms and class words whose major and minor POS categories are the same
3. Find translation pairs including the expanded content words with their expanded adjacent function words in the given sentence
4. For each *bunsetsu*, select a translation pair that has similar dependency relationship to those in the given sentence
5. Generate a parallel translation by combining the selected translation pairs

The input sentences were randomly selected from 102 sentences described in Section 3.3. The above steps, except the third step, were basically conducted manually. The Examples of the input sentences and generated parallel translations are shown in Figure 4.

The basic unit of translation pairs in our aligned parallel treebank corpus is a *bunsetsu*, and the basic unit in the selection of translation pairs is also a *bunsetsu*. One of the advantages of using a *bunsetsu* as a basic unit is that a Japanese expression represented as one of various expressions in English, or omitted in English, such as Japanese post-positional particles, is paired with a content word. Therefore, the translation of such an expression is appropriately selected together with the translation of a content word when a certain translation pair is selected. If the translation of such an expression was selected independently of the translation of a content word, the combination of each translation would be ungrammatical or unnatural. Another advantage of the basic unit, *bunsetsu*, is that we can easily refer to dependency information between *bunsetsu*s when we select an appropriate translation pair because the original treebank has the dependency information between *bunsetsu*s. These advantages are utilized in the above generation steps. For example, in the first step, a content word "国会 (*kokkai*, Diet session)" in the second example in Figure 4 was extracted from the *bunsetsu* "通常国会 (*tsuujo-kokkai*, the ordinary Diet session) に (*ni*, case marker)", and it was expanded into its class word "会 (*kai*, meeting)" in the second step. Then, a translation pair "(J) 国連子どもの権利委員会 (*kokuren-kodomo-no-kenri-iinkai*) に (*ni*, case marker); (E)the UN Committee on the Rights of the Child /(J) 対し (*taishi*); (E)towards" was extracted as a translation pair in the third step. Since the dependency between "国連子どもの権利委員会

(*kokuren-kodomo-no-kenri-iinkai*, the UN Committee on the Rights of the Child)" and "対し (*taishi*, towards)" is similar to that between "通常国会 (*tsuujo-kokkai*, the ordinary Diet session) に (*ni*, case marker)" and "向け (*muke*, towards)" in the input sentence, this translation pair was selected in the fourth step. Finally, the *bunsetsu* "国連子どもの権利委員会 (*kokuren-kodomo-no-kenri-iinkai*, the UN Committee on the Rights of the Child) に (*ni*, case marker)" and its translation "the UN Committee on the Rights of the Child" was used for generation of a parallel translation in the fifth step.

When we use the generated parallel translation for the exact translation of the input sentence, we should replace "国連子どもの権利委員会 (*kokuren-kodomo-no-kenri-iinkai*)" and its translation "the UN Committee on the Rights of the Child" with "通常国会 (*tsuujo-kokkai*, the ordinary Diet session)" and its translation "the ordinary Diet session" by consulting a bilingual dictionary. In this example, "その (*sono*)" and "them" should also be replaced with "両党 (*ryoto*)" and "both parties". It is easy to identify words in the generated translation that should be replaced with words in the input sentence because each *bunsetsu* in translation pairs is already aligned. In such cases, templates such as "[会議 (*kaigi*)] に (*ni*) 向け (*muke*)" and "towards [council]" can be automatically generated by generalizing content words expanded in the second step and their translation in the generated translation. The average number of English expressions corresponding to a Japanese expression is 1.3 as shown in Table 2. Even when there are two or more possible English expressions, an appropriate English expression can be chosen by selecting a Japanese expression by referring to dependencies in extracted translation pairs. Therefore, in many cases, English sentences can be generated just by reordering the selected expressions. The English word order was estimated manually in this experiment. However, we can automatically estimate English word order by using a language model or an English surface sentence generator such as FERGUS (Bangalore and Rambow, 2000). Unnatural or ungrammatical parallel translations are sometimes generated in the above steps. However, comprehensible translations can be generated as shown in Figure 4. The biggest advantage of this framework is that comprehensible target sentences can be generated basically by referring only to source sentences. Although it is costly to search and select appropriate translation pairs, we believe that human labor can be reduced by developing a human interface. For example, when we use a Japanese text generation system from keywords (Uchimoto et al., 2002), users should only select appropriate keywords.

We are investigating whether or not we can generate similar parallel translations to all of the Japanese sentences appearing on January 17, 1995. So far, we found that we can generate similar parallel translations to 691 out of 840 sentences (the average number of *bunsetsu*s is about 10.3) including the 102 sentences described in Section 3.3. We found that we could not generate similar parallel translations to 149 out of 840 sentences.

In the proposed framework of similar parallel translation generation, the language appearing in a corpus corresponds to a controlled language, and users are allowed to use only the controlled language to write sentences in the source language. We believe that high-quality bilingual or multilingual documents can be generated by letting us adapt ourselves to the controlled environment in this way.

4 Conclusion

This paper described aligned parallel treebank corpora of newspaper articles between languages whose syntactic structures are different from each other; they meet the following conditions.

1. It is easy to investigate the influence of the context on the translation.
2. The annotated information in the existing monolingual high-quality treebanks can be utilized.
3. It is open to the public.

To construct parallel corpora that satisfy these conditions, each sentence in the existing monolingual high-quality treebanks has been translated into a corresponding natural sentence reflecting its contextual information in a target language by skilled translators, and each parallel translation has been annotated with morphological and syntactic structures and phrasal alignment.

This paper also described the possible applications of the parallel corpus and proposed a similar parallel translation generation framework. In this framework, a parallel translation whose source language sentence is similar to a given sentence can be semi-automatically generated. In this paper we demonstrated that the framework could be achieved by using our aligned parallel treebank corpus.

In the near future, the aligned parallel treebank corpora will be open to the public, and expanded. We are planning to use the corpora actively for machine translation, as a translation aid, and for second language learning. We are also planning to develop automatic or semi-automatic alignment system and an efficient interface for machine translation aid.

Input sentence (Japanese only)	村山首相と武村蔵相との会談は首相公邸で行われ、新民主連合の新会派結成を中心に意見交換。(Prime Minister Murayama and Finance Minister Takemura met in the presidential office and they exchanged their opinions, mainly on the issue of the new faction being formed by the New Democratic Union.)
Generated parallel translation	(J) 村山富市首相と武村蔵相は会談を首相公邸で開き、新民主連合の新党結成を中心に意見交換した。 (E) Finance Minister Takemura held the meeting at the official residence to exchange views about the formation of the new party of the New Democratic Union.
Input sentence (Japanese only)	さきがけ側は通常国会に向け、政策や国会運営をテーマとする協議機関を両党に設置することを提案した。(New Party Sakigake proposed that towards the ordinary session, both parties found a council to discuss policy and Diet management.)
Generated parallel translation	(J) さきがけは国連子どもの権利委員会に向け、政策や国会運営をテーマに、話し合う機関をその間に設置することを提案した (E) Sakigake proposed to set up an organization between them towards the UN Committee on the Rights of the Child to discuss under the theme of policies and the management of the Diet.
Input sentence (Japanese only)	会談は新民連による新党が新進党との連携を深める形で進んでいることをけん制する意味もあった。(The meeting was also intended to slow the movement towards the new party by the New Democratic Union, which is trying to deepen the relationship with the New Frontier Party.)
Generated parallel translation	(J) 会談は新民連による新党が新進党との連携を深める形に進んでいることをけん制した意味があった。 (E) The meeting had meanings to restrict the movement that the new party of New Democratic Union is progressing to strengthen the coalition with The New Frontier Party.
Input sentence (Japanese only)	新進党の川端達夫衆院議員は十六日、山花貞夫氏らとの新会派結成のため、十七日に同党に離党届を提出することを決めた。(Lower House Diet Member Tatsuo Kawabata of the New Frontier Party decided on the 16th that he would hand in notification of his secession to the party on the 17th, in order to form a new faction with Sadao Yamahana's group.)
Generated parallel translation	(J) 新進党の川端達夫衆院議員は、十六日、天野祐吉氏らと新会派結成のため、十七日に新生党に離党届は提出することを決めた。 (E) On 16th Tatsuo Kawabata, a member of the House of Representatives of the New Frontier Party decided to submit The notice to leave the party to the Shinsei Party on the 17th in order to establish a new faction with Yuukichi Amano and others.
Input sentence (Japanese only)	参院会派名は民主改革連合との関係を詰めてから決定する。(As for the faction name in the Upper House, they will decide after they consider how to form a relationship with Democratic Reform Union.)
Generated parallel translation	(J) 会派名は連合との関係を話し合って決定する。 (E) The name of the faction will be decided after discussing the relationship with the JTUC.

Figure 4: Example of generated similar parallel translations.

Acknowledgments

We thank the Mainichi Newspapers for permission to use their data.

References

ATR. 1992. Dialogue Database. http://www.red.atr.co.jp/database_page/taiwa.html.

S. Bangalore and O. Rambow. 2000. Exploiting a Probabilistic Hierarchical Model for Generation. In *Proceedings of the COLING*, pages 42–48.

E. Charniak. 1999. A Maximum-Entropy-Inspired Parser. Technical Report CS-99-12.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada 2001. Fast Decoding and Optimal Decoding for Machine Translation. In *Proceedings of the ACL-EACL*, pages 228–235.

K. Imamura. 2002. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT. In *Proceedings of the TMI*, pages 74–84.

H. Isahara and M. Haruno. 2000. Japanese-English aligned bilingual corpora. In Jean Veronis, editor, *Parallel Text Processing - Alignment and Use of Translation Corpora*, pages 313–334. Kluwer Academic Publishers.

S. Kurohashi and M. Nagao. 1997. Building a Japanese Parsed Corpus while Improving the Parsing System. In *Proceedings of the NLPRS*, pages 451–456.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

M. Nagao. 1981. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*.

F. J. Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the ACL*, pages 440–447.

K. Uchimoto, S. Sekine, and H. Isahara. 2002. Text Generation from Keywords. In *Proceedings of the COLING*, pages 1037–1043.

H. Watanabe, S. Kurohashi, and E. Aramaki. 2003. Finding Translation Patterns from Paired Source and Target Dependency Structures. In Michael Carl and Andy Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 397–420. Kluwer Academic Publishers.

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

K. Yamamoto and Y. Matsumoto. 2003. Extracting Translation Knowledge from Parallel Corpora. In Michael Carl and Andy Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 365–395. Kluwer Academic Publishers.

Q. Zhou and H. Duan. 1994. Segmentation and POS Tagging in the Construction of Contemporary Chinese Corpus. *Journal of Computer Science of China*, Vol.85. (in Chinese)