Word Alignment Annotation in a Japanese-Chinese Parallel Corpus

Yujie Zhang, Zhulong Wang, Kiyotaka Uchimoto, Qing Ma, Hitoshi Isahara
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan, 619-0289
E-mail: {yujie.zhang, uchimoto, isahara} @nict.go.jp, wangzhulong@cn.fujitsu.com, qma@math.ryukoku.ac.jp

Abstract

Parallel corpora are critical resources for machine translation research and development since parallel corpora contain translation equivalences of various granularities. Manual annotation of word alignment is of significance to provide gold-standard for developing and evaluating both example-based machine translation model and statistical machine translation model. This paper presents the work of word alignment annotation in the NICT Japanese-Chinese parallel corpus, which is constructed at the National Institute of Information and Communications Technology (NICT). We describe the specification of word alignment annotation and the tools specially developed for the manual annotation. The manual annotation on 17,000 sentence pairs has been completed. We examined the manually annotated word alignment data and extracted translation knowledge from the word aligned corpus.

1. Introduction

Parallel corpora contain translation equivalences of various granularities and therefore are critical resources for machine translation research and development. Manual annotation of word alignment is of significance for providing gold-standard to development and evaluation of both example-based machine translation model and statistical machine translation model. Because parallel corpora of Asian languages are less developed, the National Institute of Information and Communications Technology (NICT) started a multilingual corpora construction project in 2002, which is focused on Asian language pairs (Uchimoto et al., 2004). The project makes effort on the annotation of detailed information, including syntactic structure and alignment at word & phrase levels. We call the corpora the NICT Multilingual Corpora.

This paper presents the word alignment annotation in the Japanese-Chinese parallel corpus, one parallel corpus of the NICT Multilingual Corpora. We will describe the specification for manual annotation and the tools specially developed for the manual annotation. The experience we obtained and points we paid special attentions are also introduced for share with other researches who are engaged in parallel corpora construction. To the best of our knowledge, the corpus is the first Japanese-Chinese parallel corpus annotated with the detailed information in the world. It will provide materials for investigation into the characteristic of translation from Japanese to Chinese.

2. NICT Japanese-Chinese Parallel Corpus

The NICT Japanese-Chinese parallel corpus consists of the original Japanese text and its Chinese translations (Zhang et al., 2005-a). 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 consistency of terminology in Chinese. The Chinese translations were then revised by other translators and lastly revised by Chinese natives. Each article was translated one sentence to one sentence, so the obtained parallel corpus is already sentence aligned. In Japanese side, morphological and syntactic structure information has been annotated following the specification of the Corpus of Spontaneous Japanese (Maekawa et al., 2000). In Chinese side, word segmentations and parts-of-speech have been annotated following the specification of Peking University (Yu, 1997). The detail of the corpus is listed in Table 1.

	Japanese	Chinese
Sentences	38,383	
Words	947,066	877,859
Vocabulary	36,657	33,425
Singletons	15,036	13,238
Aver. Sentence length	24.7	22.9

Table 1. Characteristics of NICT Japanese-Chinese Parallel Corpus.

3. Tool for Word Alignment Annotation

We specially designed and developed a tool for manual annotation of word alignment based on the investigation into a few work of word alignment annotation (Melamed, 2001; LDC, 2006). Our motivation is as follows. (a) Since automatic alignment technologies are applicable, the annotation here is manual revision on the automatically obtained alignments. A multi-aligner developed by (Zhang et al. 2005-b) is used in this work. The multi-aligner consists of a lexical knowledge based aligner, Chinese to Japanese direction application of GIZA++ and Japanese to Chinese direction application of GIZA++ (Och & Ney, 2000). The evaluation of the multi-aligner showed that 63% recall rate and 79% precision have been achieved. In order to use the results of the multi-aligner, the tool should be able to display the results of the multi-

aligner. (b) The tool should be able to provide a visualized interface for annotators to easily revise the automatically obtained results. (c) The tool should be able to display larger syntactic granularities in addition to words when syntactic structure information is available. In this way, annotators may select larger syntactic granularities and then align them effectively and conveniently. So the translation equivalences between larger units or between syntactic structures can be annotated through this tool.

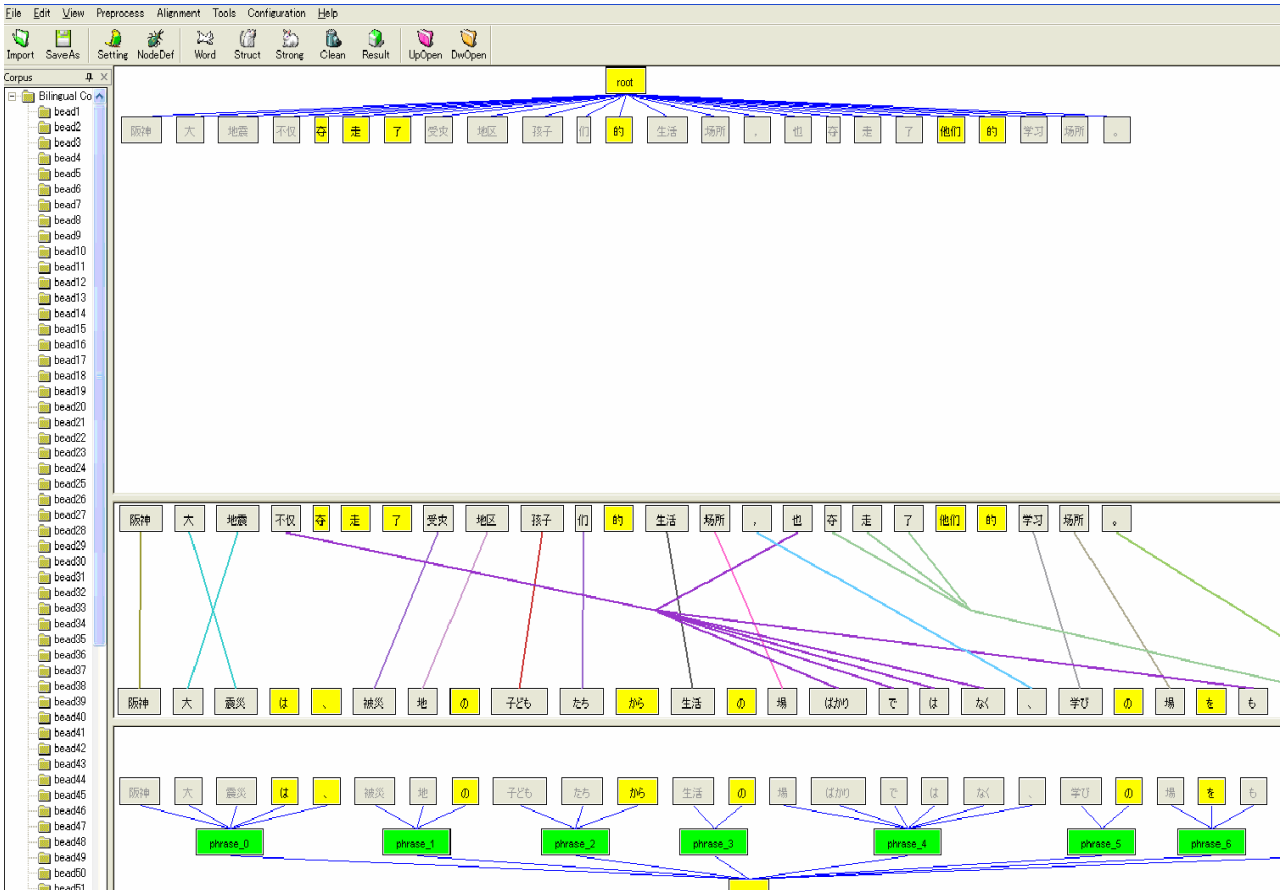

Figure 1. Visualized interface of the tool for manual alignment annotation (word level).

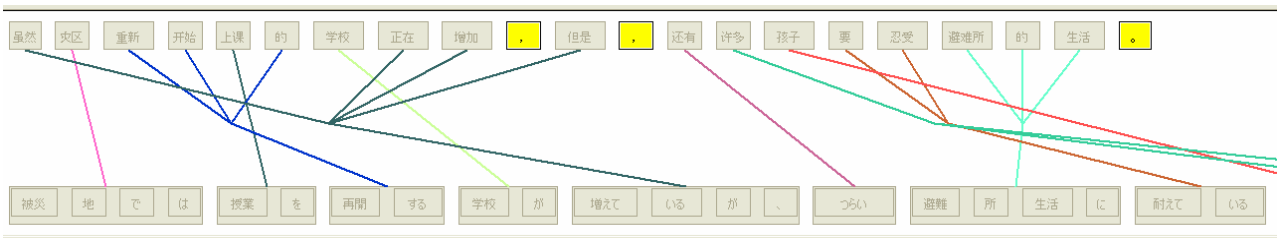

Figure 2. Visualized interface of the tool for manual alignment annotation (phrase level).

3.1 Visualized Interface

The visualized interface of our tool is showed in Figure 1 and Figure 2.

In Figure 1, the left area is used to display the ID list of the input sentence pairs. On the right, the upper area and the lower area are used to display the syntactic structures of Chinese sentence and Japanese sentence, respectively. The middle area is operation area where alignments are displayed and are to be revised. Each word of both Chinese sentence and the Japanese sentence is displayed in one quadrilateral button separately. Annotators click the buttons to select words for annotation. The lines connecting the Chinese words and the Japanese words mean alignments between them. The lines of different alignments are displayed in different colour. The quadrilateral buttons of the unaligned words are displayed in yellow, while ones of the aligned ones are displayed in grey. After one word is selected and aligned, the colour of the quadrilateral button will be changed from yellow to grey.

For labeling one alignment, annotators select Chinese (Japanese) words and then the corresponding words in the Japanese (Chinese) side by clicking the left button of the mouse. After the click operations, the selected Chinese words and the Japanese words are linked by lines. If the alignment is one-to-one, there is one line between them. If the alignment is many-to-many, i.e. multiple Chinese words being aligned to multiple Japanese words, the lines from each word of both the Chinese side and the Japanese side are got together at the same point, which is located at the middle area. See the 2-to-5 alignment in Figure 1, "不仅…也-to-ばかり で は なく…も" (not only … but). This presentation is different from the tool proposed by (Melamed, 2001), where each word of the group of the one side is linked to each word of the group on another side and therefore too many lines look redundancy. Strictly speaking, the correspondences between two groups of words do not mean each word of one group correspond to every word of another group. In our tool, the lines from each word of one group are got together at one point first and then from the point the lines are radiated to each word of another group. One-to-many and many-to-one alignments are the special cases of many-to-many alignments.

For adding one word to an alignment, just click the word and then click any word or line of the alignment. For deleting one word from one alignment, just click the word. Then the corresponding line will disappear.

If phrase alignment mode is selected, the quadrilateral buttons of phrase will be displayed in the middle area. At present, only Japanese sentences have syntactic structure information. As shown in Figure 2, words of each phrase are contained in a larger quadrilateral button. Annotators can click the quadrilateral button to select phrases for annotation. In this way, larger units can be considered and therefore be annotated more effectively.

3.2 Data Structure

We store the alignment data in a XML file and encode them in Unicode. The following tags are designed for different types of data.

<srctext> Chinese sentence
<srcword> Chinese word sequence with Part of Speech
<tgttext> Japanese sentence
<tgtword> Japanese word sequence with Part of Speech;
<wordalignment_1> automatically aligned result
<srctree> syntactic structure of the Chinese sentence
<tgttree> syntactic structure of the Japanese sentence
<wordalignment_2> manually annotated word alignment
<structalignment> manually annotated phrase alignment.

The alignment data of the example displayed in Figure 1 is shown in Figure 3.

Figure 3. The illustration of the data structure.

4. Specification for Manual Word Alignment Annotation

In alignment annotation, the semantic equivalences between Chinese words and Japanese words are detected and are aligned. The annotation is carried out according to the following criteria.

(1) Content words are considered first. After all content words are aligned, the left words are processed.

(2) The semantic equivalences should be approved based on a few pre-specified Japanese-Chinese translation dictionaries. The translation dictionaries are used to qualify the annotation of semantic equivalences in order to avoid annotation of free translations, whose semantic equivalences only appear in the certain sentence pairs and therefore are not applicable in other case. In word alignment annotation, we only consider semantic equivalences that will be applicable in general. We will deal with free translations in phrase level alignment.

(3) Alignment unit should be the minimum granularity, i.e. the smallest number of words. The words within one unit should correlate each other in semantics and therefore can not be separated further. This criterion aims at increasing the coverage of the translation knowledge that will be extracted from the aligned corpus.

(4) In the case of idioms and frozen expressions, the larger units should be preferred on both sides to ensure the two groups to be equivalent semantically and grammatically.

(5) For some Japanese postpositions, the counterpart in Chinese usually consists of one preposition and one suffix, appearing discontinuously. For instance, in the Japanese postposition phrase “机の上で(on the table)” and its Chinese translation “在 桌子 上(on the table)”, Japanese postposition “で(on)” corresponds to the Chinese preposition “在(on)” and the suffix “上(on)”. The Chinese preposition “在(on)” and the suffix “上(on)” should be glued together first and then aligned to the Japanese postposition “で(on)”.

(6) For some Japanese conjunctions, the counterpart in Chinese usually consists of two conjunctions, appearing discontinuously. For instance, in the Japanese sub sentence “遅いが” (it is late, but) and its Chinese translation “虽然晚了，但是 (it is late, but)”, the Japanese conjunction “が (but)” corresponds to the Chinese conjunction “虽然 (but)” and “但是 (but)”. The separated two Chinese conjunctions “虽然 (but)” and “但是 (but)” should be glued together first and then aligned to the Japanese conjunction “が (but)”.

(7) One big difference between Japanese and Chinese language is that the former has inflectional morphology but the latter has not. In Chinese, the words such as “了 (past tense particle)”, “过 (perfect aspect particle)” and “着 (progressive aspect particle)” are used to express tense and aspect, “被 (by)” are used to express passive voice, “使 (causative morpheme)” are used to express causative aspect. If the Japanese inflection suffix is segmented from its root, i.e. appearing as one independent morpheme, the Chinese particle should be aligned to the suffix. Otherwise, the Chinese particle should be glued to its main verb first and then aligned to the Japanese verb, which consists of the root and the inflection suffix. When the subject of the active sentence appears in the passive sentence, in Japanese it is expressed as “SUBJECT に (by SUBJECT)” and in Chinese it is expressed as “被 SUBJECT (by SUBJECT)”. In this case, the Chinese word “被 (by)” is aligned to the Japanese word “に (by)”.

5. What are Extracted from the Annotated Corpus

The manual annotation of word and phrase alignment on 17,000 sentence pairs has been completed. We examined the manually annotated data and extracted translation knowledge from them. From the word alignment data, a translation dictionary is obtained which may be used without restriction on context. From the phrase alignment data, the translation templates are obtained, in which context restrictions are contained. The former knowledge aims at increasing the coverage of applying the translation knowledge, while the later aims at increasing the accuracy of applying the translation knowledge. Some examples of the extracted knowledge are shown in Figure 4 and Figure 5. Figure 4 shows the examples of the obtained semantic equivalences at phrase level. Figure 5 shows the examples of the obtained translation templates which are obtained by replacing the specified aligned words, displayed in parentheses, with variables, like X1.

```
<sid id="950107002-001">
  <CP>完全</CP>
  <JP>すっぽり と</JP>
  <CP>积雪</CP>
  <JP>雪 に</JP>
  <CP>被 覆盖 的</CP>
  <JP>包ま れ た</JP>
  <CP>在 富士山 顶峰</CP>
  <JP>富士 山頂 に 、</JP>
  <CP>壮观 的</CP>
  <JP>壮大 な</JP>
  <CP>山帽云</CP>
  <JP>笠雲 が</JP>
```

Figure 4. Examples of the obtained semantic equivalences at phrase level.

```
<CS>一 x1(出) 便</CS>
<JS>x1(出る) と 、</JS>
<CS>在 x1(政策) 方面</CS>
<JS>x1(政策) 面 では</JS>
<CS>随地 乱 扔 的</CS>
<JS>道 に 捨てる</JS>
<CS>x1(商船) 所剩无几</CS>
<JS>残った x1(商船) は 皆無 に 等しく 、</JS>
<CS>将 可能 x1(引起)</CS>
<JS>x1(生み出し) そうだ 。</JS>
<CS>令 x1(震撼) 的</CS>
<JS>x1(震かん) さ せ た</JS>
```

Figure 5. Examples of the obtained translation templates.

6. Conclusion

This paper presents word and phrase alignment annotation in the NICT Japanese-Chinese parallel corpus. A visualized tool is developed to assistant the manual annotation. The data structure and the general guideline are described. The translation knowledge extracted from the aligned corpus is also reported. At present each sentence pair is annotated by only one annotator. We plan to select a small part of sentence pairs and ask different annotators to annotate alignment on them, in order to examine divergence among different annotators.

References

Linguistic Data Consortium. (2006). *Guidelines for Chinese Word Alignment Annotation.*

Maekawa, K., Koiso, H., Furui, F., Isahara, H. (2000). Spontaneous Speech Corpus of Japanese. In Proc. of LREC2000, pp. 947--952.

Melamed, I. Dan. (2001). *Empirical Methods for Exploiting Parallel Texts.* The MIT Press.

Och, Franz J., Ney, H. (2000). Giza++: Training of statistical translation models. Available at http://www-i6.informatik.rwthaachen.de/~och/software/GIZA++.html.

Uchimoto, K., Zhang, Y., Sudo, K., Murata, M., Sekine, S. and Isahara, H. (2004). Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information and Its Applications. In Proc. of the MLR2004: PostCOLING Workshop on Multilingual Linguistic Resources,pp.63--70.

Yu, S. (1997). *Grammatical Knowledge Base of Contemporary Chinese*. Tsinghua Publishing Company.

Zhang, Y., Liu, Q., Ma, Q., Isahara, H. (2005-a). A Multi-aligner for Japanese-Chinese Parallel Corpora. In The Tenth Machine Translation Summit Proceedings, pp.133-140.

Zhang, Y., Uchimoto, K., Ma, Q., Isahara H. (2005-b). Building an Annotated Japanese-Chinese Parallel Corpus – A Part of NICT Multi lingual Corpora. In the Tenth Machine Translation Summit Proceedings, pp.71-78.