2-3 Automatic Construction Technology for Parallel Corpora

UTIYAMA Masao and TANIMURA Midori

We have aligned Japanese and English news articles and sentences, extracted from the Yomiuri and the Daily Yomiuri newspapers, to make a large parallel corpus. We first used a method based on cross-lingual information retrieval to align the Japanese and English articles and then used a method based on dynamic programming (DP) matching to align the Japanese and English sentences in these articles. However, the articles and sentences included many incorrect alignments. To remove these, we propose two measures that evaluate the validity of the alignments. Using these measures, we successfully extracted valid article and sentence alignments.

Keywords
Japanese-English parallel corpus, Article alignment, Sentence alignment

1 Introduction

A Japanese-English parallel corpus is a necessary element in the study of natural language processing — including studies relating to machine translation, for example — and represents an invaluable linguistic resource in areas such as English-language studies, comparative linguistics, and English and Japanese language education. However, to date no large-scale Japanese-English parallel corpus has been available for public use.

Given this background, we undertook to construct a large-scale Japanese-English parallel corpus based on a relatively large-scale collection of Japanese-language newspaper articles, in addition to English-language newspaper articles partially corresponding to the content of the Japanese-language articles.

Our approach consisted of aligning the Japanese and English newspaper articles by content, and then aligning sentences within the corresponding articles.

When the contents of a subject English newspaper article corresponded to those of a subject Japanese newspaper article, in many cases the English newspaper article had been written based on the Japanese newspaper article. However, even in such cases the Japanese newspaper article was not necessarily translated directly. The English newspaper article often included non-literal translations, and in some cases omitted some of the content of the corresponding Japanese newspaper article or included content that was not in the Japanese article. In addition, the collection of English newspaper articles used for this alignment process was relatively small: less than 6% the number of corresponding Japanese articles.

In aligning articles and sentences it is critical to identify the appropriate alignments from collections of articles that contain a great deal of "noise"; as a result the measures used to judge the quality of these alignments must be highly reliable.

In this paper we propose a number of such measures in the alignment of both articles and sentences, and evaluate the reliability of these measures.

Below is a summary of the Japanese and

English newspaper articles selected for alignment, followed by a discussion of the methods used in the alignment of articles and sentences; and finally, an evaluation of the precision of each type of alignment.

2 The Japanese and English newspaper articles used in alignment

The source data used in alignment consists of articles published in the Japanese-language newspaper *The Yomiuri Shimbun* and the English-language newspaper *The Daily Yomiuri* over the period from September 1989–December 2001. The number of articles published over this period totaled approximately 2 million Japanese and 110,000 English articles. Since the number of English articles was smaller, in the alignment process we looked specifically for a Japanese article corresponding to each English article.

Since mid-July 1996, *The Daily Yomiuri* has annotated each article with metadata indicating whether or not (Y/N) the article was written as a translation of an article in the *The Yomiuri Shimbun*. For this reason, for English articles published since mid-July 1996 we attempted alignment with Japanese articles only for those English articles annotated with the tag "Y". A total of 35,318 such articles were identified. At the same time, since articles published prior to mid-July 1996 had no such metadata, we attempted alignment with Japanese articles for all English articles published during this period. A total of 59,086 such articles were identified. Hereinafter, the collection of articles published in the period prior to mid-July 1996 is referred to as the "1989–1996 Group", while the collection of articles published since mid-July 1996 is referred to as the "1996–2001 Group".

Since we used all English articles in the 1989–1996 Group, unlike with the 1996–2001 Group, in some cases no Japanese article corresponded to the English article. For this reason, in order to estimate the rough percentage of English articles likely to have corresponding Japanese articles we investigated the ratio of articles for 1997 through 2001 tagged "Y" with regard to whether they were translated from Japanese articles. The resulting percentage was 67.9%.

In the alignment process, we considered it likely that if an English article was a translation of a Japanese article then the publication date of the Japanese article should be near that of the English article. For this reason, we searched for corresponding Japanese articles from the range of articles published within two days before and after the publication of each English article. In doing so, we searched for Japanese articles corresponding to a single day's English articles from five day's worth of Japanese articles. For the 1989–1996 Group, the average number of English articles per day was 24 and the average number of Japanese articles per five days was 1,532. For the 1996–2001 Group, the average number of English articles per day was 18 and the average number of Japanese articles per five days was 2,885.

Due to the need to be able to find corresponding articles amid the high level of noise resulting from the facts described above — i.e., the high ambiguity in alignment and lack in some cases of a corresponding Japanese article — a reliable measure is required to assess the validity of article alignment. In addition, in sentence alignment even between corresponding articles, the English articles appeared to be written solely based on the Japanese articles rather than as direct translations. Thus a reliable measure is also required to assess the validity of sentence alignment, in order to identify the alignment between sentences to the degree otherwise available with a direct translation.

3 Baseline article- and sentence-alignment method

For article alignment, we adopted a cross-lingual information-retrieval framework. In other words, we found Japanese articles corresponding to the English articles provided by

using parts of the English articles as search terms in the database of Japanese articles.

The above approach in general requires either conversion into Japanese of the English articles used as search terms or conversion of the database of Japanese articles into English. In this study, we converted the database of Japanese articles into a collection of English words. We used ChaSen to conduct morphological analysis of the Japanese articles and then used EDR dictionaries and other tools to convert the resulting words into English.

After first converting the Japanese articles into a collection of English words, this collection can then be used for ordinary information retrieval by searching for the Japanese articles (i.e., searching for the results of the conversion of the articles to a collection of English words) most similar to the English articles provided as search terms. Article alignment is based on the resulting Japanese articles. In doing so, we used BM 25 [1], well known as an effective tool in information-retrieval, to assess the degree of similarity between English and Japanese articles.

We used dynamic-programming (DP) matching to align sentences between the English and Japanese articles that had been aligned using BM 25 [2][3]. The reader is asked to refer to the relevant literature [3] for a brief description of the algorithms used for sentence alignment in DP matching. We used only SIM(J,E) to derive the degree of similarity between the collection of content words J extracted from the collection of Japanese articles and the collection of content words E extracted from the collection of English articles. Degree of similarity SIM is defined as follows:

$$SIM(J,E) = (co(J \cap E) + 1) / (|J| + |E| - 2\,co(J \cap E) + 2)$$

Above, $|J|$ and $|E|$ represent the numbers of words included in the collection of Japanese articles J and the collection of English articles E. In addition, $co(J \cap E)$ represents the number of words aligned on a one-to-one basis between the words in J and those in E. In one-to-one alignment between Japanese and English words, we used the EDR Japanese-English Dictionary and the EDR English-Japanese Dictionary.

We conducted sentence alignment using degree of similarity SIM as defined above. In doing so, the program we used allowed for sentence alignment only on a 1 : n or n : 1 basis, where $1 \leq n \leq 6$. When we derived the precision of the sentence-alignment program by applying it to white-paper data with sentence alignment conducted by hand, the result was greater than 98%. In other words, the precision of the sentence-alignment program can be considered sufficiently high when used with data such as white papers, in which the Japanese has been translated faithfully into English.

4 Proposal of highly reliable measures of article and sentence alignment

As described in Section **3** above, we adopted BM 25 as a measure of similarity in article alignment and SIM as a measure of similarity in sentence alignment. However, as shown in the test below, using only these measures of similarity in article and sentence alignment cannot provide adequate precision. For this reason, in this section we define a new, more reliable measure for use in both article and sentence alignment.

We will address article alignment first. We adopted BM 25(J,E) to assess the degree of similarity between Japanese articles J and English articles E. Since this approach provides the degree of similarity between collections of words, it cannot reflect sentence order or similar factors. Accordingly, we defined another article-alignment measure, AVSIM(J,E), as a method that takes sentence order into consideration. AVSIM(J,E) is represented by the following equation, where sentence alignment between J and E is represented by {(J1,E1), ..., (Jm,Em)}.

$$AVSIM(J,E) = (SIM(J1,E1) + \ldots + SIM(Jm,Em))/m$$

Since a high value for AVSIM means that individual degrees of similarity SIM in sentence alignment are also high, we consider articles with high AVSIM values to be highly aligned.

Next, we will address a measure of "fit" in sentence alignment. As mentioned in Section **3** above, our sentence-alignment program is highly precise when used in the alignment of documents such as white papers, in which the relation between the Japanese and English texts is that of an original document and its translation. However, as mentioned in Section **2**, in general the Japanese and English newspaper articles we used were not strictly related in this manner. As a result, when conducting sentence alignment using the method described in Section **3**, numerous cases arose of both appropriate and inappropriate alignment. Under such high-noise conditions, in order to identify only the appropriate alignments we considered it helpful to employ as a measure of sentence alignment not just a measure of the degree of similarity between sentences but also the measure of article alignment. We therefore defined the following measure of sentence alignment between sentences Jk and Ek, within the article alignment between Japanese article J and English article E:

$$\mathrm{SntScore}(Jk, Ek) = \mathrm{AVSIM}(J, E) \times \mathrm{SIM}(Jk, Ek)$$

Comparing sentence alignment within a single article alignment structure, this measure results in the same ranking as the sentence-similarity measure SIM. However, in comparing sentence alignment between different articles, this approach gives priority to sentence alignment in which not just the degree of similarity between sentences but also the value of the measured article alignment is high.

5 Precision of article alignment

5.1 Evaluation of precision using random sampling

Article alignment is conducted by retrieving Japanese articles with high degrees of BM 25 similarity with respect to English articles. The precision of article alignment for Japanese articles with top-ranking degrees of similarity is shown in Table 1 for the 1996–2001 Group and the 1989–1996 Group.

In Table 1, "assessment" represents the value derived in assessing the fit of article alignment by hand. The standards for such assessment are as follows: "A" represents semantic alignment achieved for 50–60% of the text in the entirety of each article, "B" represents semantic alignment achieved for 20–30% through 50–60% of the text in the entirety of each article, "D" represents no similarity, and "C" represents any assessment other than "A", "B", or "D". The percentages referred to here represent the percentages of article alignment with the relevant assessment values over 100 article alignments from each set of articles — in the 1996–2001 Group and in the 1989–1996 Group — selected through random sampling. "Max" and "min" represent upper and lower limits at a 95% confidence level.

As described under Section **2** above, for the 1996–2001 Group we conducted alignment only for English articles tagged "Y" with regard to whether they were translated from Japanese articles, while for 1989–1996 we conducted alignment for all English articles. For this reason, the precision of the 1989–1996 Group is lower than that of the 1996–2001 Group. In addition, even though the precision of the 1996–2001 Group was higher than that of the 1989–1996 Group, since its percentage of "A" assessments was approximately 60% and its percentage of "A" or "B" assessments was approximately 70%, the number of article alignments amounting to

Table 1 *Precision of article alignment with degrees of similarity in the top rank*

type	1996–2001			1989–1996		
	lower	ratio	upper	lower	ratio	upper
A	0.49	0.59	0.69	0.20	0.29	0.38
B	0.06	0.12	0.18	0.08	0.15	0.22
C	0.03	0.08	0.13	0.03	0.08	0.13
D	0.13	0.21	0.29	0.38	0.48	0.58

noise would be too high if the results of article alignment using BM 25 were used unchanged.

Our observations show that article alignments with "A" or "B" assessments were useful for identifying alignments between Japanese and English linguistic expressions. In order to identify only such article alignments — instead of using all results of article alignment by using BM 25 unchanged — it would be better to sort alignments by fit and to identify only the higher-ranking alignments.

5.2 Precision of article alignment after sorting

We compared AVSIM and BM 25 to see which was better suited for use as an indicator of article-alignment fit. For the same data used in Table 1, we sorted article alignments in descending order by assessment value and calculated the numbers of correct answers through each rank and the corresponding percentages, defining assessment values of "A" or "B" to represent correct answers. The results are shown in Table 2. From Table 2, we determined that AVSIM was better suited than BM 25 for use as a measure of article-alignment fit.

The precision of AVSIM may be higher than that of BM 25 because (as discussed in Section **4** above) unlike BM 25, AVSIM also takes into consideration the fit of alignment between individual sentences. Using AVSIM makes it possible to identify solely high-quality article alignments from article alignments featuring high noise.

Table 2 *Ranks and precision*

rank	1996–2001				1989–1996			
	AVSIM		BM25		AVSIM		BM25	
	No.	Prec.	No.	Prec.	No.	Prec.	No.	Prec.
5	5	1.00	5	1.00	5	1.00	2	0.40
10	10	1.00	8	0.80	10	1.00	4	0.40
20	20	1.00	16	0.80	19	0.95	9	0.45
30	30	1.00	25	0.83	28	0.93	16	0.53
40	40	1.00	34	0.85	34	0.85	24	0.60
50	50	1.00	39	0.78	37	0.74	28	0.56
60	60	1.00	47	0.78	42	0.70	30	0.50
70	66	0.94	55	0.79	42	0.60	35	0.50
80	70	0.88	62	0.78	43	0.54	38	0.47
90	71	0.79	68	0.76	43	0.48	40	0.44
100	71	0.71	71	0.71	44	0.44	44	0.44

6 Precision of sentence alignment

As described in Section **2** above, since even Japanese and English articles aligned in terms of content will not necessarily align between sentences, sentence alignment between aligned articles includes a high degree of noise. For this reason, we sorted all sentence alignments derived from article alignments with the top-ranking similarity in terms of BM 25 in descending order, using SntScore, identifying alignments with good fit by using only those ranked highly as a result.

Only about 1.3 million sentence alignments featured such good fits, from both the 1989–1996 and the 1996–2001 Groups combined. In sentence alignment, one-to-one alignments are the most important. Moreover, even in the process of sentence alignment it must be noted that newspaper articles include elements such as crossheadings that are not necessarily sentences. For this reason, from one-to-one alignments we extracted only those ending with periods, calling these 1 : 1 (one-to-one) alignments and others 1 : n (one-to-many) alignments. The total number of 1 : 1 alignments was approximately 640,000, while the total number of 1 : n alignments was approximately 660,000.

To achieve the precision of 1 : 1 alignments, we performed uniform random sampling on 100 per 30,000 of the 300,000 highest-ranking alignments when sorted in descending order using SntScore. We assessed each alignment using one of two values: x/o. Here, "x" indicates that the meanings differ completely, while "o" indicates that the meanings do not differ completely. Resulting numbers of x/o assessments are shown in Table 3.

As shown in this table, the number of "x" assessments rises exponentially as ranking descends. This indicates that SntScore efficiently ranks appropriate 1 : 1 alignments more highly. From Table 3, we can conclude

that we have achieved sufficiently reliable alignment through 150,000 alignments. The cumulative proportion of "o" assessments through 150,000 alignments was 0.982.

Next, we derived the precision of 1 : n alignments for the higher-ranking alignments when sorted in descending order using SntScore, keeping these within the bounds of SntScore scores for 1 : 1 alignments in each of the following ranges set forth in Table 3 : 1–90,000, 90,001–180,000, and 180,001–270,000. In deriving precision values, as with 1 : 1 alignments we performed uniform random sampling of 100 alignments and assessed each using one of two values: x/o. The results are shown in Table 4. From this table, we can conclude that the 38,090 1:n alignments in the 1–90,000 range featured good precision.

As described above, by sorting sentence alignments using SntScore we were able to achieve sentence alignment with sufficiently high precision using the higher-ranking SntScore results, for both 1 : 1 and 1 : n alignments. We have also confirmed that precision was higher with SntScore than with SIM. As discussed in Section **4**, precision may be higher with SntScore because this measure, unlike SIM, also takes fit of article alignment into consideration.

Table 3 *Ranking and 1 : 1 alignment precision*

range	# of o's	# of x's
1 –	100	0
30001 –	99	1
60001 –	99	1
90001 –	97	3
120001 –	96	4
150001 –	92	8
180001 –	82	18
210001 –	74	26
240001 –	47	53
270001 –	30	70

Table 4 *Ranking and 1 : n alignment precision*

range	# of one-to-many	# of o's	# of x's
1 –	38090	98	2
90001 –	59228	87	13
180001 –	71711	61	39

7 Data availability

With the generous permission of *The Yomiuri Shimbun*, since 2002 we have been distributing the sentence-alignment data discussed in Section **6** of this paper for research and educational purposes. The data contains the top-150,000 1 : 1 sentence alignments and the top-30,000 1 : n sentence alignments. To date, more than 100 institutions and individuals have applied to obtain this data, which is now being used for purposes from machine translation to English-language education [4]. We also have established the "Kotonoba" website, where users may search this data.

(http://www.kotonoba.net/~snj/cgi-bin/text-search/text-search.cgi).

8 Conclusions

We have proposed two highly reliable measures for identifying valid article and sentence alignments from a Japanese-English parallel corpus featuring high noise. We have used these measures to extract article and sentence alignments from articles published in *The Yomiuri Shimbun* and *The Daily Yomiuri* from 1989 through 2001. From these, we have extracted approximately 150,000 1 : 1 sentence alignments — considered to be of relatively high quality — and approximately 38,000 other types of sentence alignments. These results have been made available to the public for use in research and education.

References

1 S. E. Robertson and S. Walker, “Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval”, SIGIR, pp.232-241, 1994.

2 William A. Gale and Kenneth W. Church, “A Program for Aligning Sentences in Bilingual Corpora”, Computational Linguistics, 19:1, pp.75-102, 1993.

3 Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, Yuji Matsumoto, and Makoto Nagao, “Bilingual Text Matching using Bilingual Dictionary and Statistics”, COLING, pp.1076-1082, 1994.

4 Kiyomi Chujo, Masao Utiyama, and Shinji Miura, “Using a Japanese-English Parallel Corpus for Teaching English Vocabulary to Beginning-Level Students”, English Corpus Studies, 13, 153-172, 2006.

UTIYAMA Masao, *Ph.D.*

Senior Researcher, Computational Linguistics Group, Knowledge Creating Communication Research Center (former: Senior Researcher, Computational Linguistics Group, Keihanna Human Info-Communication Research Center, Information and Comunication Department)

Natural Language Processing

TANIMURA Midori, *Ph.D.*

Lecturer, Kyoto University of Foreign Studies (former: Expert researcher, Computational Linguistics Group, Keihanna Human Info-Communication Research Center, Information and Comunication Department)

English Education