Reliable Measures for Aligning Japanese-English News Articles and Sentences

Masao Utiyama and **Hitoshi Isahara**
Communications Research Laboratory
3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto 619-0289 Japan
mutiyama@crl.go.jp and isahara@crl.go.jp

Abstract

We have aligned Japanese and English news articles and sentences to make a large parallel corpus. We first used a method based on cross-language information retrieval (CLIR) 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 results included many incorrect alignments. To remove these, we propose two measures (scores) that evaluate the validity of alignments. The measure for article alignment uses similarities in sentences aligned by DP matching and that for sentence alignment uses similarities in articles aligned by CLIR. They enhance each other to improve the accuracy of alignment. Using these measures, we have successfully constructed a large-scale article and sentence alignment corpus available to the public.

1 Introduction

A large-scale Japanese-English parallel corpus is an invaluable resource in the study of natural language processing (NLP) such as machine translation and cross-language information retrieval (CLIR). It is also valuable for language education. However, no such corpus has been available to the public.

We recently have obtained a noisy parallel corpus of Japanese and English newspapers consisting of issues published over more than a decade and have tried to align their articles and sentences. We first aligned the articles using a method based on CLIR (Collier et al., 1998; Matsumoto and Tanaka, 2002) and then aligned the sentences in these articles by using a method based on dynamic programming (DP) matching (Gale and Church, 1993; Utsuro et al., 1994). However, the results included many incorrect alignments due to noise in the corpus.

To remove these, we propose two measures (scores) that evaluate the validity of article and sentence alignments. Using these, we can selectively extract valid alignments.

In this paper, we first discuss the basic statistics on the Japanese and English newspapers. We next explain methods and measures used for alignment. We then evaluate the effectiveness of the proposed measures. Finally, we show that our aligned corpus has attracted people both inside and outside the NLP community.

2 Newspapers Aligned

The Japanese and English newspapers used as source data were the Yomiuri Shimbun and the Daily Yomiuri. They cover the period from September 1989 to December 2001. The number of Japanese articles per year ranges from 100,000 to 350,000, while English articles ranges from 4,000 to 13,000. The total number of Japanese articles is about 2,000,000 and the total number of English articles is about 110,000. The number of English articles represents less than 6 percent that of Japanese articles. Therefore, we decided to search for the Japanese articles corresponding to each of the English articles.

The English articles as of mid-July 1996 have tags indicating whether they are translated from Japanese articles or not, though they don't have explicit links to the original Japanese articles. Consequently, we only used the translated English articles for the article alignment. The number of English articles used was 35,318, which is 68 percent of all of the articles. On the other hand, the English articles before mid-July 1996 do not have such tags. So we used all the articles for the period. The number of them was 59,086. We call the set of articles before mid-July 1996 "1989-1996" and call the set of articles after mid-July 1996 "1996-2001."

If an English article is a translation of a Japanese article, then the publication date of the Japanese article will be near that of the English article. So we searched for the original Japanese articles within 2 days before and after the publication of each English article, i.e., the corresponding article of an English article was searched for from the Japanese articles of 5 days' issues. The average number of English articles per day was 24 and that of Japanese articles per 5 days was 1,532 for 1989-1996. For 1996-2001, the average number of English articles was 18 and that of Japanese articles was 2,885. As there are many candidates for alignment with English articles, we need a reliable measure to estimate the validity of article alignments to search for appropriate Japanese articles from these ambiguous matches.

Correct article alignment does not guarantee the existence of one-to-one correspondence between English and Japanese sentences in article alignment because literal translations are exceptional. Original Japanese articles may be restructured to conform to the style of English newspapers, additional descriptions may be added to fill cultural gaps, and detailed descriptions may be omitted. A typical example of a restructured English and Japanese article pair is:

> **Part of an English article:** ⟨e1⟩ Two bullet holes were found at the home of Kengo Tanaka, 65, president of Bungei Shunju, in Akabane, Tokyo, by his wife Kimiko, 64, at around 9 a.m. Monday. ⟨/e1⟩ ⟨e2⟩ Police suspect right-wing activists, who have mounted criticism against articles about the Imperial family appearing in the Shukan Bunshun, the publisher's weekly magazine, were responsible for the shooting. ⟨/e2⟩ ⟨e3⟩ Police received an anonymous phone call shortly after 1 a.m. Monday by a caller who reported hearing gunfire near Tanaka's residence. ⟨/e3⟩ ⟨e4⟩ Police found nothing after investigating the report, but later found a bullet in the Tanakas' bedroom, where they were sleeping at the time of the shooting. ⟨/e4⟩
>
> **Part of a literal translation of a Japanese article:** ⟨j1⟩ At about 8:55 a.m. on the 29th, Kimiko Tanaka, 64, the wife of Bungei Shunju's president Kengo Tanaka, 65, found bullet holes on the eastern wall of their two-story house at 4 Akabane Nishi, Kita-ku, Tokyo.⟨/j1⟩ ⟨j2⟩ As a result of an investigation, the officers of the Akabane police station found two holes on the exterior wall of the bedroom and a bullet in the bedroom.⟨/j2⟩ ⟨j3⟩ After receiving an anonymous phone call shortly after 1 a.m. saying that two or three gunshots were heard near Tanaka's residence, police officers hurried to the scene for investigation, but no bullet holes were found.⟨/j3⟩ ⟨j4⟩ When gunshots were heard, Mr. and Mrs. Tanaka were sleeping in the bedroom.⟨/j4⟩ ⟨j5⟩ Since Shukan Bunshun, a weekly magazine published by Bungei Shunju, recently ran an article criticizing the Imperial family, Akabane police suspect right-wing activists who have mounted criticism against the recent article to be responsible for the shooting and have been investigating the incident.⟨/j5⟩

where there is a three-to-four correspondence between $\{e1, e3, e4\}$ and $\{j1, j2, j3, j4\}$, together with a one-to-one correspondence between e2 and j5.

Such sentence matches are of particular interest to researchers studying human translations and/or stylistic differences between English and Japanese newspapers. However, their usefulness as resources for NLP such as machine translation is limited for the time being. It is therefore important to extract sentence alignments that are as literal as possible. To achieve this, a reliable measure of the validity of sentence alignments is necessary.

3 Basic Alignment Methods

We adopt a standard strategy to align articles and sentences. First, we use a method based on CLIR to align Japanese and English articles (Collier et al., 1998; Matsumoto and Tanaka, 2002) and then a method based on DP matching to align Japanese and English sentences (Gale and Church, 1993; Utsuro et al., 1994) in these articles. As each of these methods uses existing NLP techniques, we describe them briefly focusing on basic similarity measures, which we will compare with our proposed measures in Section 5.

3.1 Article alignment

Translation of words

We first convert each of the Japanese articles into a set of English words. We use ChaSen[1] to segment each of the Japanese articles into words. We next extract content words, which are then translated into English words by looking them up in the EDR Japanese-English bilingual dictionary,[2] EDICT, and ENAMDICT,[3] which have about 230,000, 100,000,

[1] http://chasen.aist-nara.ac.jp/

[2] http://www.iijnet.or.jp/edr/

[3] http://www.csse.monash.edu.au/~jwb/edict.html

and 180,000 entries, respectively. We select two English words for each of the Japanese words using simple heuristic rules based on the frequencies of English words.

Article retrieval

We use each of the English articles as a query and search for the Japanese article that is most similar to the query article. The similarity between an English article and a (word-based English translation of) Japanese article is measured by BM25 (Robertson and Walker, 1994). BM25 and its variants have been proven to be quite efficient in information retrieval. Readers are referred to papers by the Text REtrieval Conference (TREC)[4], for example.

The definition of BM25 is:

$$\text{BM25}(J,E) = \sum_{T \in E} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf}$$

where

> J is the set of translated English words of a Japanese article and E is the set of words of an English article. The words are stemmed and stop words are removed.
>
> T is a word contained in E.
>
> $w^{(1)}$ is the weight of T, $w^{(1)} = \log \frac{(N-n+0.5)}{(n+0.5)}$.
>
> N is the number of Japanese articles to be searched.
>
> n is the number of articles containing T.
>
> K is $k_1((1-b)+b\frac{dl}{avdl})$. k_1, b and k_3 are parameters set to 1, 1, and 1000, respectively. dl is the document length of J and $avdl$ is the average document length in words.
>
> tf is the frequency of occurrence of T in J. qtf is the frequency of T in E.

To summarize, we first translate each of the Japanese articles into a set of English words. We then use each of the English articles as a query and search for the most similar Japanese article in terms of BM25 and assume that it corresponds to the English article.

3.2 Sentence alignment

The sentences[5] in the aligned Japanese and English articles are aligned by a method based on DP matching (Gale and Church, 1993; Utsuro et al., 1994). We allow 1-to-n or n-to-1 ($1 \leq n \leq 6$) alignments when aligning the sentences. Readers are referred to Utsuro et al. (1994) for a concise description of the algorithm. Here, we only discuss the similarities between Japanese and English sentences for alignment. Let J_i and E_i be the words of Japanese and English sentences for i-th alignment. The similarity[6] between J_i and E_i is:

$$\text{SIM}(J_i, E_i) = \frac{\text{co}(J_i \times E_i) + 1}{l(J_i) + l(E_i) - 2\text{co}(J_i \times E_i) + 2}$$

where

> $l(X) = \sum_{x \in X} f(x)$
>
> $f(x)$ is the frequency of x in the sentences.
>
> $\text{co}(J_i \times E_i) = \sum_{(j,e) \in J_i \times E_i} \min(f(j), f(e))$
>
> $J_i \times E_i = \{(j,e) | j \in J_i, e \in E_i\}$ and $J_i \times E_i$ is a one-to-one correspondence between Japanese and English words.

J_i and E_i are obtained as follows. We use ChaSen to morphologically analyze the Japanese sentences and extract content words, which consists of J_i. We use Brill's tagger (Brill, 1992) to POS-tag the English sentences, extract content words, and use WordNet's library[7] to obtain lemmas of the words, which consists of E_i. We use simple heuristics to obtain $J_i \times E_i$, i.e., a one-to-one correspondence between the words in J_i and E_i, by looking up Japanese-English and English-Japanese dictionaries made up by combining entries in the EDR Japanese-English bilingual dictionary and the EDR English-Japanese bilingual dictionary. Each of the constructed dictionaries has over 300,000 entries.

We evaluated the implemented program against a corpus consisting of manually aligned Japanese and English sentences. The source texts were Japanese white papers (JEIDA, 2000). The style of translation was generally literal reflecting the nature of government documents. We used 12 pairs of texts for evaluation. The average number of Japanese sentences per text was 413 and that of English sentences was 495.

The recall, R, and precision, P, of the program against this corpus were $R = 0.982$ and $P = 0.986$, respectively, where

[4]http://trec.nist.gov/

[5]We split the Japanese articles into sentences by using simple heuristics and split the English articles into sentences by using MXTERMINATOR (Reynar and Ratnaparkhi, 1997).

[6]$\text{SIM}(J_i, E_i)$ is different from the similarity function used in Utsuro et al. (1994). We use SIM because it performed well in a preliminary experiment.

[7]http://www.cogsci.princeton.edu/˜wn/

$$R = \frac{\text{number of correctly aligned sentence pairs}}{\text{total number of sentence pairs aligned in corpus}}$$

$$P = \frac{\text{number of correctly aligned sentence pairs}}{\text{total number of sentence pairs proposed by program}}$$

The number of pairs in a one-to-n alignment is n. For example, if sentences $\{J_1\}$ and $\{E_1, E_2, E_3\}$ are aligned, then three pairs $\langle J_1, E_1 \rangle$, $\langle J_1, E_2 \rangle$, and $\langle J_1, E_3 \rangle$ are obtained.

This recall and precision are quite good considering the relatively large differences in the language structures between Japanese and English.

4 Reliable Measures

We use BM25 and SIM to evaluate the similarity in articles and sentences, respectively. These measures, however, cannot be used to reliably discriminate between correct and incorrect alignments as will be discussed in Section 5. This motivated us to devise more reliable measures based on basic similarities.

BM25 measures the similarity between two bags of words. It is not sensitive to differences in the order of sentences between two articles. To remedy this, we define a measure that uses the similarities in sentence alignments in the article alignment. We define $\mathrm{AVSIM}(J, E)$ as the similarity between Japanese article, J, and English article, E:

$$\mathrm{AVSIM}(J, E) = \frac{\sum_{k=1}^{m} \mathrm{SIM}(J_k, E_k)}{m}$$

where $(J_1, E_1), (J_2, E_2), \dots (J_m, E_m)$ are the sentence alignments obtained by the method described in Section 3.2. The sentence alignments in a correctly aligned article alignment should have more similarity than the ones in an incorrectly aligned article alignment. Consequently, article alignments with high AVSIM are likely to be correct.

Our sentence alignment program aligns sentences accurately if the English sentences are literal translations of the Japanese as discussed in Section 3.2. However, the relation between English news sentences and Japanese news sentences are not literal translations. Thus, the results for sentence alignments include many incorrect alignments. To discriminate between correct and incorrect alignments, we take advantage of the similarity in article alignments containing sentence alignments so that the sentence alignments in a similar article alignment will have a high value. We define

$$\mathrm{SntScore}(J_i, E_i) = \mathrm{AVSIM}(J, E) \times \mathrm{SIM}(J_i, E_i)$$

$\mathrm{SntScore}(J_i, E_i)$ is the similarity in the i-th alignment, (J_i, E_i), in article alignment J and E. When we compare the validity of two sentence alignments in the same article alignment, the rank order of sentence alignments obtained by applying SntScore is the same as that of SIM because they share a common AVSIM. However, when we compare the validity of two sentence alignments in different article alignments, SntScore prefers the sentence alignment with the more similar (high AVSIM) article alignment even if their SIM has the same value, while SIM cannot discriminate between the validity of two sentence alignments if their SIM has the same value. Therefore, SntScore is more appropriate than SIM if we want to compare sentence alignments in different article alignments, because, in general, a sentence alignment in a reliable article alignment is more reliable than one in an unreliable article alignment.

The next section compares the effectiveness of AVSIM to that of BM25, and that of SntScore to that of SIM.

5 Evaluation of Alignment

Here, we discuss the results of evaluating article and sentence alignments.

5.1 Evaluation of article alignment

We first estimate the precision of article alignments by using randomly sampled alignments. Next, we sort them in descending order of BM25 and AVSIM to see whether these measures can be used to provide correct alignments with a high ranking. Finally, we show that the absolute values of AVSIM correspond well with human judgment.

Randomly sampled article alignments

Each English article was aligned with a Japanese article with the highest BM25. We sampled 100 article alignments from each of 1996-2001 and 1989-1996. We then classified the samples into four categories: “A”, “B”, “C”, and “D”. “A” means that there was more than 50% to 60% overlap in the content of articles. “B” means more than 20% to 30% and less than 50% to 60% overlap. “D” means that there was no overlap at all. “C” means that alignment was not included in “A”,“B” or “D”. We regard alignments that were judged to be A or B to be suitable for NLP because of their relatively large overlap.

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

Table 1: Ratio of article alignments

The results of evaluations are in Table 1.[8] Here, “ratio” means the ratio of the number of articles judged to correspond to the respective category against the total number of articles. For example, 0.59 in line “A” of 1996-2001 means that 59 out of 100 samples were evaluated as A. “Lower” and “upper” mean the lower and upper bounds of the 95% confidence interval for ratio.

The table shows that the precision (= sum of the ratios of A and B) for 1996-2001 was higher than that for 1989-1996. They were 0.71 for 1996-2001 and 0.44 for 1989-1996. This is because the English articles from 1996-2001 were translations of Japanese articles, while those from 1989-1996 were not necessarily translations as explained in Section 2. Although the precision for 1996-2001 was higher than that for 1989-1996, it is still too low to use them as NLP resources. In other words, the article alignments included many incorrect alignments.

We want to extract alignments which will be evaluated as A or B from these noisy alignments. To do this, we have to sort all alignments according to some measures that determine their validity and extract highly ranked ones. To achieve this, AVSIM is more reliable than BM25 as is explained below.

[8]The evaluations were done by the authors. We double checked the sample articles from 1996-2001. Our second checks are presented in Table 1. The ratio of categories in the first check were A=0.62, B=0.09, C=0.09, and D=0.20. Comparing these figures with those in Table 1, we concluded that first and second evaluations were consistent.

Sorted alignments: AVSIM vs. BM25

We sorted the same alignments in Table 1 in decreasing order of AVSIM and BM25. Alignments judged to be A or B were regarded as correct. The number, N, of correct alignments and precision, P, up to each rank are shown in Table 2.

rank	1996-2001				1989-1996			
	AVSIM		BM25		AVSIM		BM25	
	N	P	N	P	N	P	N	P
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

Table 2: Rank vs. precision

From the table, we can conclude that AVSIM ranks correct alignments higher than BM25. Its greater accuracy indicates that it is important to take similarities in sentence alignments into account when estimating the validity of article alignments.

AVSIM and human judgment

Table 2 shows that AVSIM is reliable in ranking correct and incorrect alignments. This section reveals that not only rank order but also absolute values of AVSIM are reliable for discriminating between correct and incorrect alignments. That is, they correspond well with human evaluations. This means that a threshold value is set for each of 1996-2001 and 1989-1996 so that valid alignments can be extracted by selecting alignments whose AVSIM is larger than the threshold.

We used the same data in Table 1 to calculate statistics on AVSIM. They are shown in Tables 3 and 4 for 1996-2001 and 1989-1996, respectively.

type	N	lower	av.	upper	th.	sig.
A	59	0.176	0.193	0.209	0.168	**
B	12	0.122	0.151	0.179	0.111	**
C	8	0.077	0.094	0.110	0.085	*
D	21	0.065	0.075	0.086		

Table 3: Statistics on AVSIM (1996-2001)

In these tables, “N” means the number of alignments against the corresponding human judgment.

type	N	lower	av.	upper	th.	sig.
A	29	0.153	0.175	0.197	0.157	*
B	15	0.113	0.141	0.169	0.131	
C	8	0.092	0.123	0.154	0.097	**
D	48	0.076	0.082	0.088		

Table 4: Statistics on AVSIM (1989-1996)

"Av." means the average value of AVSIM. "Lower" and "upper" mean the lower and upper bounds of the 95% confidence interval for the average. "Th." means the threshold for AVSIM that can be used to discriminate between the alignments estimated to be the corresponding evaluations. For example, in Table 3, evaluations A and B are separated by 0.168. These thresholds were identified through linear discriminant analysis. The asterisks "**" and "*" in the "sig." column mean that the difference in averages for AVSIM is statistically significant at 1% and 5% based on a one-sided Welch test.

In these tables, except for the differences in the averages for B and C in Table 4, all differences in averages are statistically significant. This indicates that AVSIM can discriminate between differences in judgment. In other words, the AVSIM values correspond well with human judgment. We then tried to determine why B and C in Table 4 were not separated by inspecting the article alignments and found that alignments evaluated as C in Table 4 had relatively large overlaps compared with alignments judged as C in Table 3. It was more difficult to distinguish B or C in Table 4 than in Table 3.

We next classified all article alignments in 1996-2001 and 1989-1996 based on the thresholds in Tables 3 and 4. The numbers of alignments are in Table 5. It shows that the number of alignments estimated to be A or B was 46738 $(= 31495 + 15243)$. We regard about 47,000 article alignments to be sufficiently large to be useful as a resource for NLP such as bilingual lexicon acquisition and for language education.

	1996-2001	1989-1996	total
A	15491	16004	31495
B	9244	5999	15243
C	4944	10258	15202
D	5639	26825	32464
total	35318	59086	94404

Table 5: Number of articles per evaluation

In summary, AVSIM is more reliable than BM25 and corresponds well with human judgment. By using thresholds, we can extract about 47,000 article alignments which are estimated to be A or B evaluations.

5.2 Evaluation of sentence alignment

Sentence alignments in article alignments have many errors even if they have been obtained from correct article alignments due to free translation as discussed in Section 2. To extract only correct alignments, we sorted whole sentence alignments in whole article alignments in decreasing order of SntScore and selected only the higher ranked sentence alignments so that the selected alignments would be sufficiently precise to be useful as NLP resources.

The number of whole sentence alignments was about 1,300,000. The most important category for sentence alignment is one-to-one. Thus, we want to discard as many errors in this category as possible. In the first step, we classified whole one-to-one alignments into two classes: the first consisted of alignments whose Japanese and English sentences ended with periods, question marks, exclamation marks, or other readily identifiable characteristics. We call this class "one-to-one". The second class consisted of the one-to-one alignments not belonging to the first class. The alignments in this class, together with the whole one-to-n alignments, are called "one-to-many". One-to-one had about 640,000 alignments and one-to-many had about 660,000 alignments.

We first evaluated the precision of one-to-one alignments by sorting them in decreasing order of SntScore. We randomly extracted 100 samples from each of 10 blocks ranked at the top-300,000 alignments. (A block had 30,000 alignments.) We classified these 1000 samples into two classes: The first was "match" (A), the second was "not match" (D). We judged a sample as "A" if the Japanese and English sentences of the sample shared a common event (approximately a clause). "D" consisted of the samples not belonging to "A". The results of evaluation are in Table 6.[9]

[9]Evaluations were done by the authors. We double checked all samples. In the 100 samples, there were a maximum of two or three where the first and second evaluations were different.

range	# of A's	# of D'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 6: One-to-one: Rank vs. judgment

This table shows that the number of A's decreases rapidly as the rank increases. This means that SntScore ranks appropriate one-to-one alignments highly. The table indicates that the top-150,000 one-to-one alignments are sufficiently reliable.[10] The ratio of A's in these alignments was 0.982.

We then evaluated precision for one-to-many alignments by sorting them in decreasing order of SntScore. We classified one-to-many into three categories: "1-90000", "90001-180000", and "180001-270000", each of which was covered by the range of SntScore of one-to-one that was presented in Table 6. We randomly sampled 100 one-to-many alignments from these categories and judged them to be A or D (see Table 7). Table 7 indicates that the 38,090 alignments in the range from "1-90000" are sufficiently reliable.

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

Table 7: One-to-many: Rank vs. judgment

Tables 6 and 7 show that we can extract valid alignments by sorting alignments according to SntScore and by selecting only higher ranked sentence alignments.

Overall, evaluations between the first and second check were consistent.

[10]The notion of "appropriate (correct) sentence alignment" depends on applications. Machine translation, for example, may require more precise (literal) alignment. To get literal alignments beyond a sharing of a common event, we will select a set of alignments from the top of the sorted alignments that satisfies the required literalness. This is because, in general, higher ranked alignments are more literal translations, because those alignments tend to have many one-to-one corresponding words and to be contained in highly similar article alignments.

Comparison with SIM

We compared SntScore with SIM and found that SntScore is more reliable than SIM in discriminating between correct and incorrect alignments.

We first sorted the one-to-one alignments in decreasing order of SIM and randomly sampled 100 alignments from the top-150,000 alignments. We classified the samples into A or D. The number of A's was 93, and that of D's was 7. The precision was 0.93. However, in Table 6, the number of A's was 491 and D's was 9, for the 500 samples extracted from the top-150,000 alignments. The precision was 0.982. Thus, the precision of SntScore was higher than that of SIM and this difference is statistically significant at 1% based on a one-sided proportional test.

We then sorted the one-to-many alignments by SIM and sampled 100 alignments from the top 38,090 and judged them. There were 89 A's and 11 D's. The precision was 0.89. However, in Table 7, there were 98 A's and 2 D's for samples from the top 38,090 alignments. The precision was 0.98. This difference is also significant at 1% based on a one-sided proportional test.

Thus, SntScore is more reliable than SIM. This high precision in SntScore indicates that it is important to take the similarities of article alignments into account when estimating the validity of sentence alignments.

6 Related Work

Much work has been done on article alignment. Collier et al. (1998) compared the use of machine translation (MT) with the use of bilingual dictionary term lookup (DTL) for news article alignment in Japanese and English. They revealed that DTL is superior to MT at high-recall levels. That is, if we want to obtain many article alignments, then DTL is more appropriate than MT. In a preliminary experiment, we also compared MT and DTL for the data in Table 1 and found that DTL was superior to MT.[11] These

[11]We translated the English articles into Japanese with an MT system. We then used the translated English articles as queries and searched the database consisting of Japanese articles. The direction of translation was opposite to the one described in Section 3.1. Therefore this comparison is not as objective as it could be. However, it gives us some idea into a comparison of MT and DTL.

experimental results indicate that DTL is more appropriate than MT in article alignment.

Matsumoto and Tanaka (2002) attempted to align Japanese and English news articles in the Nikkei Industrial Daily. Their method achieved a 97% precision in aligning articles, which is quite high. They also applied their method to NHK broadcast news. However, they obtained a lower precision of 69.8% for the NHK corpus. Thus, the precision of their method depends on the corpora. Therefore, it is not clear whether their method would have achieved a high accuracy in the Yomiuri corpus treated in this paper.

There are two significant differences between our work and previous works.

(1) We have proposed AVSIM, which uses similarities in sentences aligned by DP matching, as a reliable measure for article alignment. Previous works, on the other hand, have used measures based on bag-of-words.

(2) A more important difference is that we have actually obtained not only article alignments but also sentence alignments on a large scale. In addition to that, we are distributing the alignment data for research and educational purposes. This is the first attempt at a Japanese-English bilingual corpus.

7 Availability

As of late-October 2002, we have been distributing the alignment data discussed in this paper for research and educational purposes.[12] All the information on the article and sentence alignments are numerically encoded so that users who have the Yomiuri data can recover the results of alignments. The data also contains the top-150,000 one-to-one sentence alignments and the top-30,000 one-to-many sentence alignments as raw sentences. The Yomiuri Shimbun generously allowed us to distribute them for research and educational purposes.

We have sent over 30 data sets to organizations on their request. About half of these were NLP-related. The other half were linguistics-related. A few requests were from high-school and junior-high-school teachers of English. A psycho-linguist was also included. It is obvious that people from both inside and outside the NLP community are interested in this Japanese-English alignment data.

[12] http://www.crl.go.jp/jt/a132/members/mutiyama/jea/index.html

8 Conclusion

We have proposed two measures for extracting valid article and sentence alignments. The measure for article alignment uses similarities in sentences aligned by DP matching and that for sentence alignment uses similarities in articles aligned by CLIR. They enhance each other and allow valid article and sentence alignments to be reliably extracted from an extremely noisy Japanese-English parallel corpus.

We are distributing the alignment data discussed in this paper so that it can be used for research and educational purposes. It has attracted the attention of people both inside and outside the NLP community.

We have applied our measures to a Japanese and English bilingual corpus and these are language independent. It is therefore reasonable to expect that they can be applied to any language pair and still retain good performance, particularly since their effectiveness has been demonstrated in such a disparate language pair as Japanese and English.

References

Eric Brill. 1992. A simple rule-based part of speech tagger. In *ANLP-92*, pages 152–155.

Nigel Collier, Hideki Hirakawa, and Akira Kumano. 1998. Machine translation vs. dictionary term translation – a comparison for English-Japanese news article alignment. In *COLING-ACL'98*, pages 263–267.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Japan Electronic Industry Development Association JEIDA. 2000. *Sizen Gengo Syori-ni Kan-suru Tyousa Houkoku-syo (Report on natural language processing systems).*

Kenji Matsumoto and Hideki Tanaka. 2002. Automatic alignment of Japanese and English newspaper articles using an MT system and a bilingual company name dictionary. In *LREC-2002*, pages 480–484.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *ANLP-97*.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *SIGIR'94*, pages 232–241.

Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, Yuji Matsumoto, and Makoto Nagao. 1994. Bilingual text matching using bilingual dictionary and statistics. In *COLING'94*, pages 1076–1082.