Automatic Alignment of Japanese and English Newspaper Articles using an MT System and a Bilingual Company Name Dictionary

Kenji Matsumoto and Hideki Tanaka

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
{kenji.matsumoto, hideki.tanaka}@atr.co.jp

Abstract

One of the crucial parts of any corpus-based machine translation system is a large-scale bilingual corpus that is aligned at various levels such, as the sentence and phrase levels. This kind of corpus, however, is not easy to obtain, and accordingly, there is a great need for an efficient construction method. We approach this problem by integrating two large monolingual corpora in two different languages sharing the same source of information. We often see such a situation in journalistic texts where the same events are reported in many languages. Unfortunately, they often lack article-level alignment information and the recovery of this is the first problem to solve. In this paper, we report a method of automatically aligning Japanese and English newspaper articles in the financial and economic news domain. Although conventional methods require some manual work, the proposed method works fully automatically. We show that our method can align such newspaper articles with an accuracy of 97%.

1. Introduction

There is no need to explain the necessity of using parallel corpora in natural language processing. A large parallel corpus is an indispensable part of such natural language applications as a statistical machine-translation system (Brown et al., 1990), an example-based machine-translation system (Nagao, 1984), and many other applications including cross-lingual information retrieval and bilingual lexicon construction. The most famous parallel corpus is definitely the Canadian parliamentary records, i.e., the Hansard Corpus. The availability of such a voluminous and fairly clean corpus, however, is rather limited. Such a corpus between the Japanese and Chinese languages, for example, is not easily obtained. There is a strong need to develop a voluminous parallel corpus efficiently.

There have been attempts made to combine bilingual text data into the making of a parallel corpus. Hasan et al. (2001) constructed a parallel corpus of Chinese and Japanese in this way. World Wide Web is now receiving keen attention as a bilingual text data mining site (Resnik, 1999) since it produces an increasing number of bilingual documents day after day.

We obtained a Japanese and English newspaper corpus of non-trivial size. This corpus was provided by the Nihon Keizai Shinbun (Nikkei), a newspaper that specializes in reporting financial and economic news. English articles in the corpus are manual translations of Japanese articles, but no explicit cross-reference information is available. This situation encouraged us investigate restoring the article-level correspondence between the two languages toward making a gigantic parallel corpus.

Significant news items are often reported in several languages and therefore news texts in different languages can be highly related. This makes a news text collection a valuable resource in the construction of a parallel corpus. There have been many works that deals with it.

Xu (1999) attempted to align Chinese and English news articles and Collier et al. (1998) attempted the same with Japanese and English news articles provided by Reuters. Takahashi et al. (1997) dealt with the same newspaper article alignment problem by using anchor expressions, like numeral expressions and proper nouns. This work required some manual work in the alignment process. The work by Collier et al. formalized the article alignment problem as a multi lingual information retrieval task. They empirically compared the direct bilingual dictionary look-up method and machine-translation method as a means of a query construction.

		'95	*'96*	*'97*	*'98*	*'99*	*'00*	*'01*
The Nihon Keizai Shimbun	(Japanese)	135,928	146,347	148,282	150,431	148,529	147,529	128,135
	(English)	18,803	16,617	16,427	14,138	14,929	15,550	15,707
The Nikkei Industrial Daily	**(Japanese)**	**68,088**	**68,654**	**67,416**	**66,836**	**64,531**	**62,295**	**59,617**
	(English)	**9,658**	**8,440**	**9,795**	**10,402**	**8,823**	**8,227**	**7,731**
The Nikkei Financial Daily	(Japanese)	28,949	28,624	28,241	27,906	27,854	31,331	29,215
	(English)	5,232	5,937	4,972	4,792	4,700	5,349	4,972
The Nikkei Marketing Journal	(Japanese)	19,686	19,393	18,772	18,650	17,750	17,339	17,728
	(English)	148	114	204	226	102	78	20

Table 1: Yearly article counts of the four Nikkei newspapers

Our approach is basically the same as the latter approach and we used machine-translation system to construct query words. Our method, however, is different in that we selected possible candidate articles before the article matching. Our experiment with the Nikkei data revealed an accuracy of 97% and we thus confirmed the effectiveness of our method. We also tested our method on another journalistic bilingual document collection, broadcast news, in an effort to further demonstrate the basic effectiveness of our method. We, however, obtained lower accuracy in this undertaking and were thus compelled to investigate the reason.

In the following, we will explain our target corpus, Nikkei newspaper, in section 2. Then we will explain our alignment method in section 3 and report the results of the alignment experiment with the Nikkei newspaper in section 4 and NHK broadcast news in section 5. We will discuss the performance variance of the two experiments in section 6 and offer our conclusions and a discussion of further work in section 7.

2. Nikkei Database

Nihon Keizai Shimbun, Inc, or NIKKEI publishes four daily newspapers in Japanese, i.e., The Nihon Keizai Shimbun, The Nikkei Industrial Daily, The Nikkei Financial Daily and The Nikkei Marketing Journal. Some of their articles are translated into English for distribution via various Internet services. Currently about 30,000 English articles are accumulated every year. At NIKKEI, these Japanese and English articles are stored in separate databases and have no explicit correspondence information. However, we can expect to make a voluminous bilingual corpus by aligning the English and Japanese articles with each other. Table 1 shows the actual number of Japanese and English articles between 1995 and 2001.

We chose English articles derive from The Nikkei Industrial Daily as the alignment source, since this paper covers the smallest domain, i.e., popular industrial and corporate related news, and its data size is the second largest among the four. We searched for the Japanese articles corresponding to each source English article because the Japanese volume is about ten times as large as the English volume.

3. Alignment Methods

We adopted a two-stage process to achieve fast alignment as follows.

1) Japanese candidate article selection
2) Alignment decision.

Figure 1 shows the overview of the alignment process.

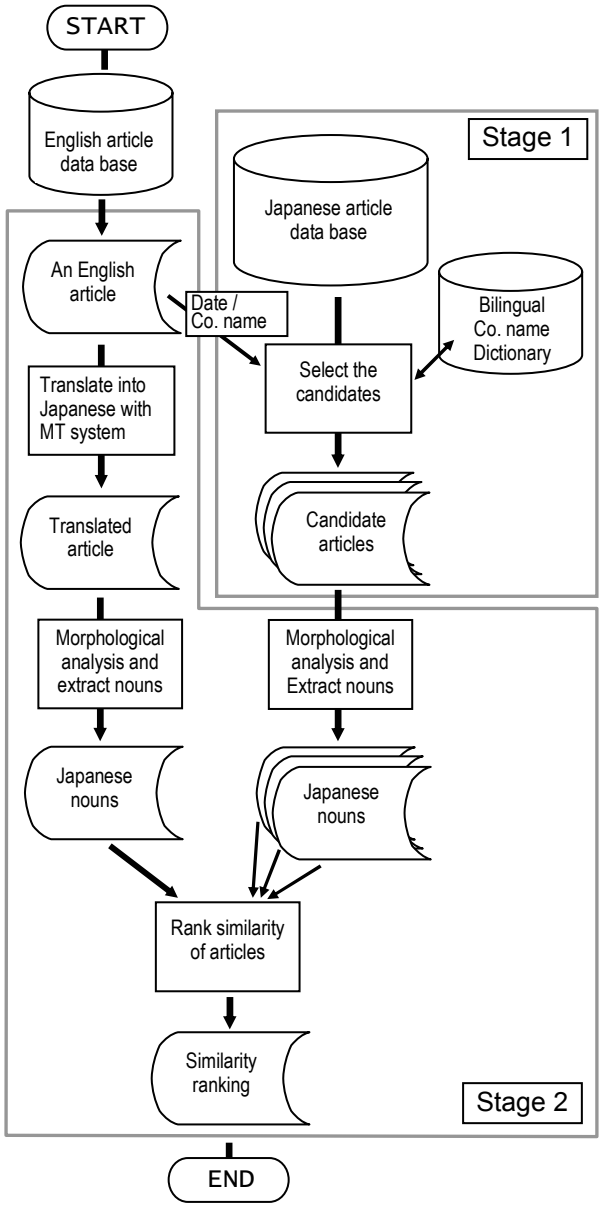


Figure 1: Alignment process

3.1. Japanese candidate article selection

We first extracted Japanese articles as the source candidates for each English article (Stage 1). We used chronological information (such as publication dates) and company names as the keys to find the Japanese candidate articles.

a) Chronological information

The publication dates of those Japanese articles can be fixed almost exactly with the information contained in the English source article. Figure 2 shows that the English article is translation of an article in The Nikkei Industrial Daily published Tuesday.

… choose between Microsoft Corp.'s Windows 2000 Server, Linux and other systems. (The Nikkei Industrial Daily ***Tuesday*** edition)

Figure 2: Publication name and date

b) Company names

In addition, those Japanese articles sharing the same company names appearing in the English article were selected, since these names conveyed important information in the Nikkei Industrial Daily and were always preserved even when the Japanese articles were translated into English.

Stock Code	Japanese Name	English Name
1911	住友林業	Sumitomo Forestry Co., Ltd.
9434	日本テレコム	Japan Telecom Co., Ltd
—	セイコーエプソン	Seiko Epson Corp.

Table 2: Bilingual company name dictionary

We used a bilingual company name dictionary to identify the English company names and obtain their Japanese translations. This dictionary, as shown in Table 2, lists 30,000 company names both in English and Japanese and their stock exchange codes, when applicable (similar to a ticker symbol).

Sumitomo Forestry Co. (***1911***) is broadening activities designed to help lumber processors a ..

Figure 3: Stock exchange code in Nikkei articles

The company names were identified in the following way. If an article contained a stock exchange code, such as "1911" shown in Figure 3, we used it; otherwise, we extracted sequences of capitalized words and matched them with the English entries in the dictionary. The matching was flexible in that it would evaluate the similarity between word sequences using Dice's coefficient derived by letter-level longest common subsequence (LCS) calculation (Cormen, 2001) as in formula (1). Any matching exceeding a pre-determined threshold was considered a match.

$$\frac{2\times\left|lcs(x,y)\right|}{|x|+|y|} \tag{1}$$

$|x|$ and $|y|$ are the lengths of the word sequences being matched, and $|lcs(x,y)|$ stands for the length of the LCS of x and y. The threshold for matching was a value of 0.75 in the experiments.

For example, "Victor Company of Japan Ltd" in the dictionary and "Victor Co. of Japan" in an article have longest common subsequence "Victor Co of Japan" considered to be a match since the score is: $2\times 18/(27+19)\approx 0.78$.

3.2. Alignment decision

After selecting Japanese candidate articles, the alignment decision stage consists of several steps (Stage 2).

In the first step, each source English article was translated into Japanese with a commercially available machine-translation system. A machine-translation system usually translates each noun and numerical word correctly except some ambiguous words, although it dose not always translate an English sentence into a Japanese one correctly.

In the second step, nouns were extracted from the results of the morphological analysis. The nouns of the candidate Japanese articles were also extracted.

In the last step, the similarity between the source article and the candidates was ranked in terms of Dice's coefficient (Manning et al., 1999) based on the word match count as in formula (2).

$$\frac{2\times|x\cap y|}{|x|+|y|} \tag{2}$$

$|x|$ and $|y|$ are the number of nouns extracted from each article. $|x\cap y|$ is the number of shared noun between both articls.

4. Experiment 1

We used 98 English translations of Nikkei articles published between 7 and 9 Mar., 2001 for an evaluation. Table 3 shows the result of the experiment. These were first associated with the candidate Japanese articles and the candidates were ranked using the criteria mentioned above. A total of 84 out of 98 articles in the evaluation set were associated with their candidates in terms of both company names and dates, and 83 were correctly aligned with the Japanese articles with the top rank. The remaining 14 articles were associated with their candidates in terms of only the date information and 12 of them were correctly aligned with the top-ranking Japanese candidate. The alignment correctness was evaluated by human judgment in this evaluation. The overall accuracy reached 97% (95/98). Consequently, we consider the proposed method to be quite useful when building a large bilingual corpus with aligned articles.

Date	Correct		Total	Success Rate
	With company names and dates	With dates		
3/7	33	4	38	97.4%
3/8	27	5	33	97.0%
3/9	23	3	27	96.3%
Total	83	12	98	96.9%

Table 3: Result of Experiment 1

5. Experiment 2

We were interested in whether or not our method can be effectively applied to a different type of journalistic corpora and conducted the same experiments with a

bilingual broadcast-news corpus compiled by NHK (Japan Broadcasting Corporation).

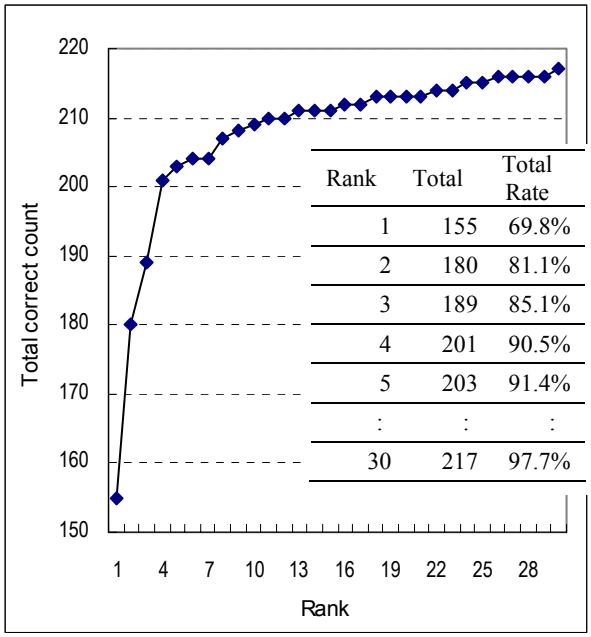
We used 222 English articles (nine days worth). They were translations of Japanese articles and had common identification numbers, which enabled an automatic objective result evaluation.[1]

The corpus covered all of news topics and the categories fell under six genres: politics, economy, international, society, sports, and others. These articles rarely contained company names and the candidates were extracted almost entirely with dates.

We were able to identify 155 correct Japanese articles out of the 222 articles (69.8%). The rate was increased to 91.4% when the top five candidates were included in the correct alignment.

When we restricted the category to politics and international, the accuracy with the top candidate reached about 80% and was further increased to 90% with the top three candidates. Although these figures were lower than those obtained with The Nikkei Industrial Daily, we consider the proposed method to be useful as a semi-automatic human assistance tool for article alignment.

Figure 4 and Table 4 show the result of the experiment.



Rank	Total	Total Rate
1	155	69.8%
2	180	81.1%
3	189	85.1%
4	201	90.5%
5	203	91.4%
:	:	:
30	217	97.7%

Figure 4: Graph of total correct articles in Experiment 2

	eco	*int*	*pol*	*soc*	*spo*	*genre*	*all*
Correct	24	59	36	24	1	161	155
Total	32	76	44	46	2	222	222
Rate	75	77.6	81.8	52.2	50	72.5	69.8

eco : Economy
int : International
pol : Politics
soc : Society
spo : Sports
genre : Total using genre information
all : Total not using genre information

Table 4: Result of Experiment 2

[1] This experiment was purely motivated by scientific interest and there was no practical necessity, since English and Japanese articles are linked with the id numbers.

6. Discussion

In this section, we discuss the difference in accuracy observed between the two experiments. The corpora we used were both from mass-communication media. We therefore expected similar alignment accuracy, but the result failed to meet our expectations: the Nikkei had an accuracy of 97% and NHK economic news 75%. How can we explain the difference and improve the alignment accuracy of the NHK corpus?

We hypothesized that the difference can be attributed to the two reasons below.

(1) Difference in translation style

Our alignment method used Dice's coefficient derived from the common noun counts between English and Japanese articles. The common nouns were obtained using an E-J machine-translation system that produces rather a direct and literal translation. If the translation style of the corpus is literal and the nouns are well preserved across the translation, alignment accuracy should be high, otherwise it should be low. We then questioned whether NHK (broadcast news) and Nikkei (newspaper) use different translation styles: is the Nikkei more literal than NHK?

(2) Difference in similarity distribution in Japanese corpus

We searched for a Japanese article from which an English article had been translated. If there are many similar Japanese articles in the corpus, selection will become hard since there will be many competing articles. Then, we questioned whether there is a difference in similarity distribution between the NHK and the Nikkei Japanese articles.

We measured the E-J similarity of aligned articles obtained in the aforementioned experiments using the same metrics described in Formula (2). The average similarity for the Nikkei was 0.28 and for NHK it was 0.30. NHK's similarity was found to be slightly higher and this indicates that NHK's translation style is slightly more literal than that of the Nikkei. The hypothesis (1), then, does not explain the lower accuracy of NHK.

We then investigated the possibility of the second hypothesis and measured inter-article similarities for each article both in the NHK and the Nikkei Japanese corpora. Table 5 shows the overall similarity and Table 6 shows the highest similarity for each article.

	NHK	*Nikkei*
Max.	1	0.629
Min.	0	0
Average	0.059	0.050

Table 5: Overall similarity

	NHK	*Nikkei*
Max.	1	0.629
Min.	0	0.073
Average	0.527	0.212

Table 6: Highest similarity for each article

The average of the overall similarity in Table 5 was 0.059 for NHK and 0.050 for the Nikkei and we can see no apparent difference. However, the maximum similarity for NHK was 1 and 0.63 for the Nikkei. The average of the highest similarity in Table 6 was 0.53 for NHK and 0.21 for Nikkei.

These figures indicate that the Nikkei and NHK corpora are similar in overall nature but the NHK corpus contains tight clusters.

In fact, we can observe quite similar articles in the NHK corpus: TV news repeats the same news every one hour with slight changes of the contents. Newspapers, on the other hand, are published everyday and a piece of news is reported once a day. Here, we can conclude that the second hypothesis was correct.

This suggests that we need to use verbs together with nouns and temporal information in similarity calculation to further increase the alignment accuracy.

7. Conclusions and future work

We proposed a method of aligning English and Japanese newspaper articles. This method is fully automated and consists of two stages: a candidate selection stage and an alignment stage. The candidate selection stage extracts articles that share common company names and close publication dates. Then alignment stage calculates the similarity rank based on common noun counts between the two texts. We used an E-J machine translation system to obtain noun translations for the English texts.

The experiments with the Nikkei new texts demonstrated the effectiveness of our method.

We are now conducting an automatic alignment of all our Japanese and English data obtained between 1995 and 2001 that was presented in Table 1. Furthermore, we would like to perform the alignments at the sentence level by using the obtained parallel corpus.

8. Acknowledgments

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus".

9. References

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer and P. S. Roosin. (1990). A Statistical Approach to Machine Translation. In *Computational Linguistics 16* (pp. 79-85).

Collier, N. H., H. Hirakawa and A. Kumano. (1998). Machine Translation vs. Dictionary Term Translation - a Comparison for English-Japanese News Article Alignment. In *COLING-ACL'98* (pp.263-267). University of Montreal, Canada.

Cormen, T. H. (2001). Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, Introduction to Algorithms, 2nd Ed., MIT Press, Cambridge.

Hasan M. M., and Y. Matsumoto. (2001). Multilingual Document Alignment – A study with Chinese and Japanese. In *NLPRS2001* (pp. 617-623).

Manning, C., and H. Schütze. (1999). Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA.

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by Analogy Principle. In Alick Elithorn and Ranan Banerji (eds.) *Artificial and Human Intelligence* (pp. 173-180). Amsterdam: North-Holland.

Resnik P. (1999). Mining the Web for Bilingual Text. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)* (pp. 527-534). College Park, Maryland.

Takahashi Y., S. Shirai and F. Bond. (1977). A Method of Automatically Aligning Japanese & English Newspaper Articles. In *NLPRS1997* (pp.657-660).

Xu, D. and C.L. Tan. (1999). Alignment and Matching of Bilingual English-Chinese News Texts. In *Machine Translation, Kluwer AcademicPublisher.* (vol. 14 pp. 1-33)