Chinese-Japanese Clause Alignment

Xiaojie Wang [1] and Fuji Ren [2]

[1]School of Information Engineering,
Beijing University of Posts and Telecommunications,
Beijing, China, 100876
[2]Department of Information Science & Intelligent Systems,
Tokushima University,
Tokushima, Japan
xjwang@bupt.edu.cn, ren@is.tokushima-u.ac.jp

Abstract. Bi-text alignment is useful to many Natural Language Processing tasks such as machine translation, bilingual lexicography and word sense disambiguation. This paper presents a Chinese-Japanese alignment at the level of clause. After describing some characteristics in Chinese-Japanese bilingual texts, we first investigate some statistical properties of Chinese-Japanese bilingual corpus, including the correlation test of text lengths between two languages and the distribution test of length ratio data. We then pay more attention to n-m(n>1 or m>1) alignment modes which are prone to mismatch. We propose a similarity measure based on Hanzi characters information for these kinds of alignment modes. By using dynamic programming, we combine statistical information and Hanzi character information to find the overall least cost in aligning. Experiments show our algorithm can achieve good alignment accuracy.

1 Introduction

Text alignment is an important task in Natural Language Processing (NLP). It can be used to support many other NLP tasks. For example, it can be utilized to construct statistical translation models (Brown et al. 1991), and to acquire translation examples for example-based machine translation (Kaji et al. 1992). It can be helpful in bilingual lexicography (Tiedemann 2003). It is also used to improve monolingual word sense disambiguation (Diab and Resnik 2002).

The approaches to text alignment can be classified into two types: statistical-based and lexical-based. The statistical-based approaches rely on non-lexical information (such as sentence length, co-occurrence frequency, etc.) to achieve an alignment task. As illustrated in the research of Gale and Church (1991) for the sentence-level alignment, they start from the fact that the length of a source text sentence is highly correlated with the length of its target text translation.

The method proposed in Kay and Roscheisen (1993) is based on the assumption that in order for the sentences in a translation to correspond, the words in them must also correspond. Their method makes use of lexical anchor points to lead an alignment at the sentence level.

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 400 – 412, 2005.
© Springer-Verlag Berlin Heidelberg 2005

It has been shown that different language pairs are in favor of different information in alignment. For example, Wu (1994) found that the sentence-length correlation between English and Chinese is not as good as between English and French. Also, there is less cognate information between Chinese-English pair than that in English-French pair, while the alignment of Chinese-Japanese pair can make use of information of Hanzi commonly appearing in both languages (Tan and Nagao 1995). Currently, most methods rely on either or both of above two ideas (Veronis 2000). The approaches combining both length and lexical information, such as Melamed (2000), seem to represent the state of the art.

Standing on text alignment at sentence-level, structure-based alignment has been paid more and more attentions (Matsumoto et al. 1993, Wu 1997, Ding and Palmer 2004). Since bi-trees can bring more information and thus more helpful to machine translation. However, due to the limitation of parsers, especially for those non-English languages, current structure-based alignment cannot deal with complex or compound sentences well. Average length of Chinese sentences in test is about 12 Chinese words (Ding and Palmer 2004). Comparing with our bi-text corpora where there are nearly 25 words per Chinese sentence and 10 words per Chinese clause, we can reasonably believe that an alignment at the level of clause is more manageable for current parsers rather than sentence.

Some works have been done at the level of clause (Kit et al.2004) and phrase (Venugopal et al. 2003). As noted in Kit et al. (2004), a full Chinese sentence is often a compound, including several propositions. While in clause level, units are usually well-formed clauses (including only one proposition) and phrases or even words. It is thus more manageable for current parsers to do further structure-based alignment.

In this paper, we present our current work on Chinese-Japanese bilingual alignment at the level of clause. For Chinese-Japanese pairs, as noted in Tan and Nagao (1995), because of the different sentential structure, alignment at the level of sentence could sometimes result in pairing of quite a number of sentences en masse in both texts. This exacerbates burden for parsers. We will show that this problem is less severe in the level of clause.

Combination of both length-based information and lexicon-based information is proved to be the state of art approach to bi-text alignment (Veronis 2000). So we will also combine these two kinds of information in our alignment algorithm. In Chinese and Japanese pairs, up to now, there is no systematical investigation on length correlation between two languages. On the lexical information, Hanzi characters, which commonly occur in both languages, have been used for improving sentence alignment. But there are very different conclusions on the affects brought from Hanzi information. Tan and Nakao (1995) achieve a great improvement by combining Hanzi information with length information, and the accuracy is raised from 78% to 96% for their less than 500 Chinese and Japanese test sentences which are mostly from texts. But in Zaiman et al. (2001), authors draw a very different conclusion for the affect of Hanzi information. We will also dress these questions in this paper.

The remainder of paper is organized as follows: section 2 describes some characteristics in Chinese-Japanese bilingual corpora, and show what we can get in a clause level alignment. Section 3 describes several information sources we use in our approach, and how to combine these information sources for improving alignment. Where, we investigate the length correlation between two languages and the distribution of length ratio. We especially concern about the alignment modes of *n*-*m* (*n*>1 or *m*>1), which are well known to be prone to mismatch. We construct a deliberate measure to deal with these kinds of alignment. We then implement several experiments and evaluate different effects brought from different information sources in section 4. Finally we draw some conclusions.

2 Some Characteristics Chinese-Japanese Bi-texts

As noted in Tan and Nagao (1995), Chinese and Japanese have different sentential structures. One of these differences is the use of periods (including question marks and exclamation marks) to end sentence. If we use periods as the end of sentence, there will be quite a number of sentences en masse between both texts at sentence-level alignment. That is, there will be some alignments involving *n*>1 Chinese sentences or *m*>1 Japanese sentences. These kinds of alignment modes are usually more difficult for an algorithm to manage and fallible. They thus propose to allow a sentence in one text to be matched with a part of the sentence in the other text if possible in order to create a finer correspondence pair. But in their works, they only allow a sentence ended with periods to be matched with the sequence of clause and/or phrase ended with break points in the other text, rather than clause to clause alignment. It does reduce the *n*-*m*(*n*>1 or *m*>1) alignment at the level of sentence, but, on the other hand, as we will see in following example in Table 1, it will cause that *n*>1 Chinese clauses are aligned with one Japanese sentence or *m*>1 Japanese clauses are aligned with one Chinese sentence.

Table 1. An example of Chinese-Japanese bi-text: one Chinese sentence is aligned with two Japanese sentences

Chinese text	Japanese text
幸好，刀子小，拇指的骨头硬，所以直到今天指头还连在手掌上，不过那伤痕到死也不会消失。	幸ナイフが小さいのと、親指の骨が堅かったので、今だに親指は手に付いている。然し創痕は死ぬまで消えぬ。

In Table 1, one Chinese sentence is aligned with two Japanese sentences. If we allow Japanese sentence to be aligned with clauses in Chinese, then we can get some new alignments as listed in Table 2.

Table 2. When a Japanese sentence can be aligned with clauses in Chinese

Chinese text	Japanese text
幸好，刀子小，拇指的骨头硬，所以直到今天指头还连在手掌上，	幸ナイフが小さいのと、親指の骨が堅かったので、今だに親指は手に付いている。
不过那伤痕到死也不会消失。	然し創痕は死ぬまで消えぬ。

Since the aligned units in Chinese are clauses which marked by either commas or periods, while the units in Japanese is marked by periods, the first alignment in Table 2 is that 4 Chinese units are aligned with one Japanese unit. Comparing with alignment in Table 1 which includes at most two units in an alignment, this alignment unites more units in Chinese, thus more easily to expose to mismatch. Beside of that, texts on both sides are still a compound, including several propositions.

If we allow clause to clause alignment, where both languages can use break points as the end of a unit, we will have alignments as in Table 3 for the same bi-text. Where, one alignment is 2-1 and others are 1-1. The biggest lumped unit includes 2 clauses, which is same as that in sentence-level, but here we get several alignments in finer grain.

Table 3. Clause to Clause alignment

幸好，刀子小，	幸ナイフが小さいのと、
拇指的骨头硬，	親指の骨が堅かったので、
所以直到今天指头还连在手掌上，	今だに親指は手に付いている。
不过那伤痕到死也不会消失。	然し創痕は死ぬまで消えぬ。

The above example is not a singular situation. A statistics from manually aligned 251 bi-text paragraphs in our corpus shows that although there are more than 4 percents pairs including more than 2 units in either side in both sentence-level and clause-level bi-text, there are only 0.4 percents of pairs including more than 3 units at either side of clause-level alignment, while the percent is nearly 2 in sentence-level alignment. As often noted, *n-m* alignment modes (*n*>3 or *m*>3) are easily exposed to mismatch in alignment. Less such pairs, thus, make it more easily to be matched.

Also, as we have noticed, the clause-level alignment is more manageable for a parser to do further processing. We have some statistics in Table 4 to show a contrast between sentences and clauses in our Chinese-Japanese corpus.

Table 4. Average numbers of characters in sentences and clauses

	Chinese characters per unit	Japanese characters per unit
Sentence	25.5	35
Clause	10	16

Based on above two reason, that is, it is easier to deal with in alignment itself and easier for further processing (such as parsing). We thus think it is more useful and hopeful to do alignment at level of clause.

There is another characteristic we will utilize in our Chinese-Japanese alignment. For some kinds of corpus, such as novels which we currently work on, there are lots of dialogs quoted directly. There are often several clauses even sentences in these direct quotations. To align them at clause-level, we should break the quotations, but in Chinese-Japanese pair, it often involves re-orders of clauses. For example, let we consider the bi-text in Table 5.

If we do not break the quotation marks for texts in Table 5, they can be aligned as in Table 6 including several clauses or sentences in each of the alignments.

Table 5. An example of bi-text includes direct quotation

Chinese Text	Japanese Text
我看她依然带着奇怪的表情，就问：“我买点什么土产回来送你呢？你要什么来着？” 她说：“想吃越后的竹叶糖。”	それでも妙な顔をしているから「何かみやげに買って来てやろう、何が欲しい」と聞いてみたら「越後の笹飴が食べたい」と云った。

When we break the quotations, we should re-order the clauses in order to get correct alignments as in Table 7. We have a special pre-process to manage these re-orders in

Table 6. An alignment when quotations are not broken

Chinese text	Japanese text
我看她依然带着奇怪的表情，就问：“我买点什么土产回来送你呢？你要什么来着？”	それでも妙な顔をしているから「何かみやげに買って来てやろう、何が欲しい」と聞いてみたら
她说：“想吃越后的竹叶糖。”	「越後の笹飴が食べたい」と云った。

Table 7. An alignment when quotations are broken

Chinese text	Japanese text
我看她依然带着奇怪的表情，	それでも妙な顔をしているから
就问：	と聞いてみたら
“我买点什么土产回来送你呢？	「何かみやげに買って来てやろう、
你要什么来着？”	何が欲しい」
她说：	と云った。
“想吃越后的竹叶糖。”	「越後の笹飴が食べたい」

3 The Approach to Clause Alignment

We use dynamic programming to find overall optimal alignment paragraph by paragraph. We combine both length-based information and Hanzi-based information to measure the cost for each possible alignment between Chinese and Japanese strings.

Before we give the measure function, we first define a few notations. Let s_i denote the i th clause in Chinese, $|s_i|$ denote the number of character it includes, s_{ij} denote the string from the i th clause to the j th clause in Chinese text, t_u denote the u th clause in Japanese, t_{uv} denote the string from the u th clause to the v th clause in Japanese text. $j < i$ means that s_{ij} is an empty string, and so does t_{uv} . Let $d(i,u;j,v)$ be the function that computes the cost of aligning s_{ij} with t_{uv} . We divide $d(i,u;j,v)$ into three parts, as shown in (1), to reflect that we utilize three information sources to compute the cost.

$$d(i,u;j,v) = L(i,u;j,v) + M(i,u;j,v) - \alpha H(i,u;j,v) \tag{1}$$

Where $L(i,u;j,v)$ depends on length ratio of the paired two strings s_{ij} and t_{uv}, different length ratios cause different $L(i,u;j,v)$; $M(i,u;j,v)$ depends on the alignment mode in the pair, that is, depending on how many clauses involved in this pair at both sides, different *n-m* modes cause different $M(i,u;j,v)$. $H(i,u;j,v)$ is the contribution from Hanzi characters common in both strings.

We first describe how to use length-based information, and then Hanzi-based information. We give a deliberate measure on multiple alignments which are prone to mismatch.

3.1 Length and Mode

To utilize the length information in Chinese-Japanese alignment, we make an investigation on length correlation between these two languages. We also check if the normal hypothesis of the length ratio is appropriate for this language pair. Following Gale and Church (1991), we use alignment data at paragraph-level to do the correlation and normal test. 4893 paragraphs in our Chinese-Japanese corpus are randomly chosen for these tests.

The correlation test is to examine if the lengths of Chinese texts are related to their Japanese translations. That is, if longer texts in Chinese tend to be translated into longer texts in Japanese, and if shorter texts tend to be translated into shorter texts. Figure 1 shows that the lengths (in characters) of Chinese and Japanese paragraphs are highly correlated.

In Gale and Church (1991), they based their distance measure on an assumption that each character in one language gives rise to a random number of characters in the other language. They assume those random variables are independent and identically distributed with a normal distribution. They check the English-French pair for this assumption, and estimate parameters for the normal distribution. However, this assumption is not checked in Chinese-Japanese pair. For making use of the same kind

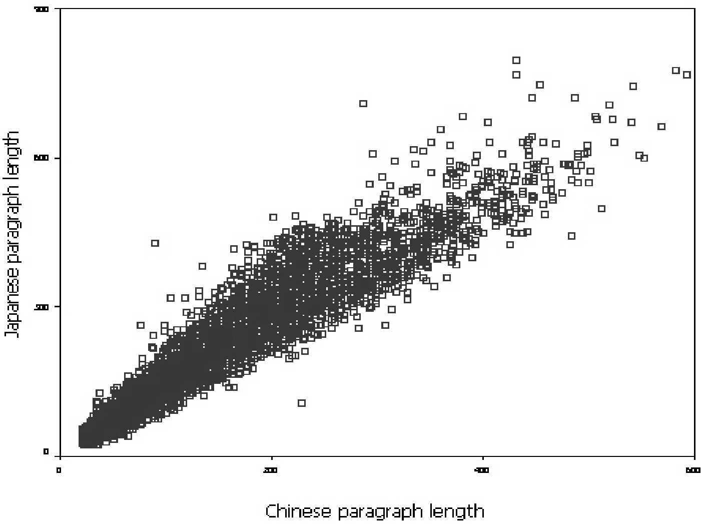

Fig. 1. Paragraph lengths are highly correlated in Chinese-Japanese bi-texts. The horizontal axis shows the length of Chinese paragraphs, while vertical scale shows the lengths of the corresponding Japanese paragraphs. The Pearson correlation (.948) is significant at the 0.01 level (2-tailed)

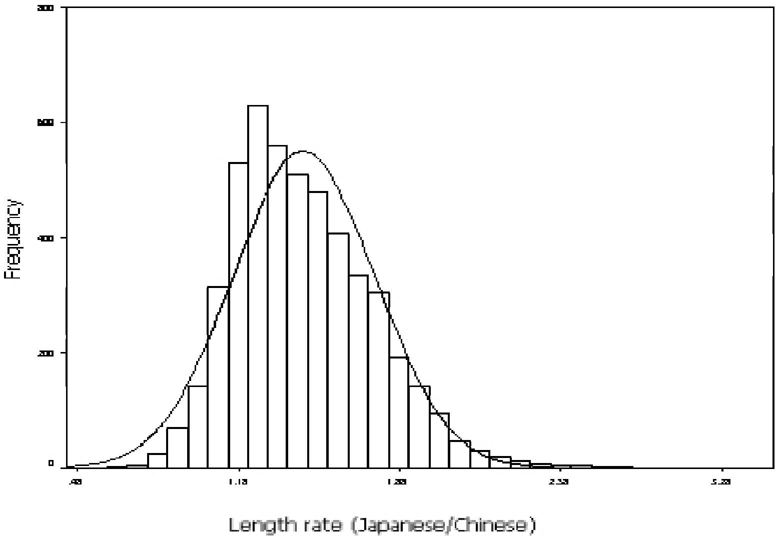

Fig. 2. The length-rate data of Chinese-Japanese pair is approximately normal. The horizontal axis shows the length-ratio (Japanese/Chinese), while vertical scale shows the frequency of each ratio. The histogram fits normal distribution approximately with mean=1.47, Std. Deviation=0.31, kurtosis=3.85, skewness=0.99

of length-based information in alignment, we thus give an investigation on if this assumption can be held in Chinese-Japanese pair. We also use the paragraph-based data to do the test. We use 4893 length ratios from 4893 Chinese-Japanese paragraphs as samples. Figure 2 is the histogram for these ratios.

We can find that length ratio data is approximately normal with mean μ =1.47 and Std. Deviation σ =0.31. Then following Gale and Church (1991), we compute $L(i,u;j,v)$ using equation (2).

$$L(i,u;j,v) = -100 * \log(2 * (1 - \Pr(|(|t_{uv}|/|s_{ij}| - \mu)/\sigma|))) \tag{2}$$

In (2), $L(i,u;j,v)$ is set to 0 when both strings s_{ij} and t_{uv} are empty, $L(i,u;j,v)$ is set to a given maximum value when only one of these two strings is empty. The maximum value is set to 2000 in our experiments.

To compute the cost for different alignment mode, we aligned 928 pairs at the level of clause manually. Then we estimate the probabilities of different alignment modes using frequencies of these modes occurring in data. The result is listed in Table 8.

Table 8. The frequencies of different alignment modes

Category (*n-m*)	Frequency f(*n-m*)	Probability Pr(*n-m*)
0-1 or 1-0	1	0.001
1-1	623	0.671
1-2	74	0.080
2-1	162	0.175
2-2	30	0.032
Others	38	0.041
Total	928	1.000

Suppose that $j-i=n$ means the string s_{ij} including n Chinese clauses, specially, we use $n=-1$ to denote that there is no clause in s_{ij} . Similarly, $v-u=m$ means the string t_{uv} including m Japanese clauses and $m=-1$ means that there is no clause in t_{uv}, the cost for mode $n-m$ can be computed by (3).

$$M(i,u;j,v) = -100 * \log(\Pr(n-m)) \tag{3}$$

3.2 Hanzi Information

Another source of information is Hanzi occurring commonly in both Chinese and Japanese. We include totally 2626 Chinese-Japanese Hanzi character pairs in a dictionary. We construct different contribution measures as in equation (4) and (5) for different alignment mode. The equation (5) is for 2-2 mode, and the equation (4) is for other modes. By including some length-related information in these measures, we can also control the alignment length.

$$H(i,u;j,v)=\begin{cases} 0 & j<i \mid v>u & 0-1or1-0 \\ h_1(i,0;u,0) & j=i \,\&\, v=u & 1-1 \\ h_1(i,0;u,0)+h_2(i,j;u,0) & j=i+1 \,\&\, v=u & 2-1 \\ h_1(i,0;u,0)+h_3(i,0;u,v) & j=i \,\&\, v=u+1 & 1-2 \end{cases} \tag{4}$$

Where $h_1(i,0,u,0)=|s_i \cap t_u|/\min\{s_i,t_u\}$, here we use the relative number of commonly occurring Hanzi to measure the similarity between two strings, $h_2(i,j,u,0)=(|s_j \cap t_u|-|s_i \cap s_j \cap t_u|)/s_j$, which is the relative number of commonly occurring Hanzi only in s_j and t_u but not in s_i, similarly, we use $h_3(i,0,u,v)$ to denote $(|s_i \cap t_v|-|s_i \cap t_u \cap t_v|)/t_v$, which is the relative number of commonly occurring Hanzi only in s_i and t_v but not in t_u. For 2-2 mode, we use (5) as a measure.

$$\begin{aligned} H(i,u;j,v)= \min\{ \; & h_1(i,0;u,0)+h_2(i,j;u,0)+h_1(0,j;0,v), \\ & h_1(i,0;u,0)+h_3(i,0;u,v)+h_1(0,j;0,v), \\ & h_1(i,0;0,v)+h_1(0,j;u;0)\} \end{aligned} \tag{5}$$

The three items in the right of equation (5) are the measures for three different ways of merging s_i and s_j and matching with t_u and t_v in Chinese-Japanese alignment. Comparing with previous measures, such as those in Tan and Nagao (1995), we think these measures are more reasonable. Also, by using these measures, we can not only manage different alignment modes such as contraction, expansion and merger, but also take length information into consideration.

This completes the description of distance measure. We finally use dynamic programming to find the overall least cost in matching. Let $D(i,u)$ be total distance aligning clauses from first one to i th in Chinese and clauses from first one to u th in Japanese, the recurrence then can be described in (6).

$$D(j,v)=\min\{D(i-1,u-1)+d(i,j;u,v)\} \tag{6}$$

4 Experiments and Evaluations

We use 251 Chinese-Japanese paragraph pairs in our experiments, which include total 2760 Chinese clauses and 2482 Japanese clauses respectively. If all the bilingual clauses are aligned correctly, it should produce 2161 pairs of translation examples at the level of clause. 100 paragraphs are originally in Chinese, and others are originally in Japanese. We have got all the parameters in last section except coefficient α in distance measure. To estimate this parameter, we align another 41 paragraphs which include 427 Chinese clauses and 381 Japanese clauses. We use it as training data in our experiments when α is needed.

All language data used in this paper are drawn from a Chinese-Japanese bilingual corpus which includes more than 3 millions of Chinese characters and more than 4 millions of Japanese characters. There are 31 contemporary novels written by either Chinese or Japanese authors, including totally more than 70,000 paragraph-based alignments.

We first implement some experiments to investigate the affects brought from different information sources and different combinations of these sources. The results are listed in Table 9.

There are 86.02% of pairs aligned correctly when all information is used. As we have mentioned, if all the bilingual clauses are aligned correctly, there are more than 4% of pairs including more than 2 units in either side of bi-text. Considering that our current program cannot manage these pairs, the accuracy of 86.02 percents is reasonably good. Also, 8.39% of pairs are partly aligned correctly, while only 5.59% of pairs are completely wrong.

From Table 9, we can find Hanzi information (H) is the most helpful one for alignment. When only one information source is used, H achieves the best accuracy, which at least gains 25% more than other two kinds of information. Under the same conditions, H always achieves more improvements than others. Combining H with mode information(M) improves M's initial accuracy from 34.52 to 78.84, while length information(L) raises it to 71.81. Combining H with L improves L's initial accuracy from 40.36 to 83.18, while M raises it to 71.81. When two of the three kinds of source are +, by combining H, we get an improvement from 71.81 to 86.02, more than 10 points, while combination of L improves the accuracy from 78.84 to 86.02, combination of M only contributes less than 3%, from 83.18 to 86.02.

Table 9. How different factors effect the accuracy (+: used, −: not used)

H	L	M	Accuracy(%)
+	-	-	66.65
+	+	-	83.18
+	-	+	78.84
+	+	+	**86.02**
-	+	-	40.36
-	-	+	34.52
-	+	+	71.81

Table 10. A Chinese-Japanese pair aligned by using a normal measure

Chinese text	Japanese text
当时，勘太郎无路可逃， 拼命扑过来。	その時勘太郎は逃げ路を失って、 一生懸命に飛びかかって来た。

Table 11. A Chinese-Japanese pair aligned correctly by using our measures

Chinese text	Japanese text
当时，勘太郎无路可逃，	その時勘太郎は逃げ路を失って、
拼命扑过来。	一生懸命に飛びかかって来た。

We also implement an experiment where $H(i,u;j,v)$ is just the number of Hanzi characters commonly occurring in both sides of bi-text under consideration. We name this $H(i,u;j,v)$ a normal measure for Hanzi information. When other information sources are same, the best accuracy for normal measure is 76.77%. There is nearly 10% improvement achieved by our $H(i,u;j,v)$. Our proposal shows its advantage in leveraging Hanzi information for different alignment modes. A normal measure tends to lump small units into the long string which may include more common Hanzi characters. For example, bi-text in Table 10 is aligned by using a normal measure. It can be aligned correctly using measure (5) as in Table 11.

Finally, we find that the accuracy is in fact sensitive to parameters of length and mode, but can be improved significantly by adjusting coefficient of Hanzi information in the cost function (1).

Some works have been done for Chinese-Japanese Sentence alignment. Tan and Nagao (1995) achieved 96% accuracy on less than 500 sentences which are mostly selected from text. They also included 20% (20 out of 100 messages) sentences used for parameter estimation in their test data. While in Zaima et al. (2001), the accuracy is less than 60% at the level of sentence. No previous work is reported on clause-level alignment. For comparative, we also implement a sentence-level alignment for the same 251 paragraphs. We achieve 87.28% accuracy on 1189 Chinese sentence and 1158 Japanese sentences. Clause-level alignment based on results of sentence-based alignment has an accuracy of less than 75%.

5 Conclusions

This paper describes a Chinese-Japanese alignment at the level of clause by combining both length information and Hanzi character information. We give a detail description on some characteristics in Chinese-Japanese bilingual texts. We check the correlation between the lengths of Chinese text and Japanese text, and find that length ratio data fits normal hypothesis approximately. We pay a special attention on *n*-*m* alignment where *n* or *m* is greater than 1, and propose a similarity measure based on Hanzi information to get better accuracy than normal one. Experiments show our proposal is very helpful. We believe that the proposed similarity measures can be also helpful for the alignment of other language pairs by using other lexical information instead of Hanzi information in Chinese-Japanese pairs.

We will extend our program to manage *n-m* alignments where *n* or *m* is bigger than 2 in our future works. By analyzing the new alignment modes such as 1-3 , 2-3 or 3-3 etc., we will build Hanzi information based measures to achieve higher accuracy for Chinese-Japanese alignment.

Acknowledgements. This work has been partly supported by the Education Ministry of Japan under Grant-in-Aid for Scientific Research B (No. 14380166). Part of this work was done while the first author was at Nara Institute of Science and Technology (NAIST). He appreciates Professor Yuji Matsumoto and the Matsumoto laboratory in NAIST very much for lots of help he had received, including permitting him to use the Chinese-Japanese corpus.

References

1. Peter F. Brown, Jennifer C. Lai and Robert L. Mercer. Aligning Sentences in Parallel Corpora. In *Proc. of 29th Conf. of Assoc. for Comput. Linguistics*, ACL-1991. pp.169-176.
2. Mona Diab and Philip Resnik. An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics,* ACL-2002, pp.255-262.
3. Yuan Ding and Martha Palmer Automatic Learning of Parallel Dependency Treelet Pairs, In *proceedings of the First International Joint Conference on Natural Language Processing* ,IJCNLP-2004, pp.30-37.
4. William A. Gale and Kenneth W. Church. A Program for Aligning Sentences in Bilingual Corpora. In *Proc. of 29th Conf. of Assoc. for Comput. Linguistics*, ACL-1991, pp.177-184.
5. Hiroyuki Kaji, Yuuko Kida and Yasutusgu. Morimoto. Learning Translation Templates from Bilingual Text. In *Proceedings of the fifteenth International Conference on Computational Linguistics,* COLING-1992, Nantes, France, pp.672-678.
6. Martin Kay and Martin Roscheisen. Text-Translation Alignment. *Computational Linguistics.* Vol. 19, No. 1, pp.121-142, 1993.
7. Chew Lim Tan and Makoto Nagao. Automatic Alignment of Japanese-Chinese Bilingual Texts. *IEICE Trans. Information and System.* Vol, E78-D, No. 1, Jan., 1995.
8. Yuji Matsumoto, Hiroyuki Ishimoto, Takehito Utsuro: Sructural Matching of Parallel Texts. In *Proc. of 31st Conf. of the Association for Computational Linguistics*. ACL-1993 pp.23-30.
9. Chunyu Kit, Jonathan J. Webster, King Kui Sin, Haihua Pan and Heng Li. Clause alignment for Hong Kong legal text. *Intern. Journal of Corpus Linguistics* Vol. 9, No.1, 2004, pp.29-51.
10. I. Dan Melamed. Pattern recognition for mapping bitext correspondence. In *Parallel Text Processing*, J. Veronis (eds.). pp.25-47. 2000, Kluwer Academic Publishers.
11. Jörg Tiedemann. Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing, Doctoral Thesis, *Studia Linguistica Upsaliensia 1*, ISSN 1652-1366, ISBN 91-554-5815-7.
12. Ashish Venugopal , Stephan Vogel and Alex Waibel . Effective Phrase Translation Extraction from alignment model. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, ACL-2003, pp.319-326.
13. Jean Veronis. From the Rosetta stone to the information society—A survey of parallel text processing. In *Parallel Text Processing*. J. Veronis (ed.). 25-47. 2000, Kluwer Academic Publishers.

14. Wu Dekai. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In *Proc. of the 32st Meeting of Association for Comput. Linguistics*, ACL-1994. pp.80-87.
15. Wu Dekai. Stochastic inversion transduction grammars and bilingual parsing of parallel corpara. *Computational Linguistics*, Vol.23, No.3, pp.377-404.
16. Yasumichi Zaiman, Ryoko Yasukawa, Fuji Ren and Teruaki Aizawa. Text alignment using statistical technique and the language feature. *Technical Report of IEICE* TL2000-40, NLC2000-75(2001-03), pp.1-8.