Bilingual Text Matching using Bilingual Dictionary and Statistics

Takehito Utsuro Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
Takayama, Ikoma, Nara 630-01, Japan
{utsuro, matsu}@is.aist-nara.ac.jp

Abstract

This paper describes a unified framework for bilingual text matching by combining existing hand-written bilingual dictionaries and statistical techniques. The process of *bilingual text matching* consists of two major steps: *sentence alignment* and *structural matching of bilingual sentences*. Statistical techniques are applied to estimate word correspondences not included in bilingual dictionaries. Estimated word correspondences are useful for improving both sentence alignment and structural matching.

1 Introduction

Bilingual (or parallel) texts are useful as resources of linguistic knowledge as well as in applications such as machine translation.

One of the major approaches to analyzing bilingual texts is the statistical approach. The statistical approach involves the following: alignment of bilingual texts at the sentence level using statistical techniques (e.g. Brown, Lai and Mercer (1991), Gale and Church (1993), Chen (1993), and Kay and Röscheisen (1993)), statistical machine translation models (e.g. Brown, Cocke, Pietra, Pietra et al. (1990)), finding character-level / word-level / phrase-level correspondences from bilingual texts (e.g. Gale and Church (1991), Church (1993), and Kupiec (1993)), and word sense disambiguation for MT (e.g. Dagan, Itai and Schwall (1991)). In general, the statistical approach does not use existing hand-written bilingual dictionaries, and depends solely upon statistics. For example, sentence alignment of bilingual texts are performed just by measuring sentence lengths in words or in characters (Brown et al., 1991; Gale and Church, 1993), or by statistically estimating word level correspondences (Chen, 1993; Kay and Röscheisen, 1993).

The statistical approach analyzes unstructured sentences in bilingual texts, and it is claimed that the results are useful enough in real applications such as machine translation and word sense disambiguation. However, structured bilingual sentences are undoubtedly more informative and important for future natural language researches. Structured bilingual or multilingual corpora serve as richer sources for extracting linguistic knowledge (Klavans and Tzoukermann, 1990; Sadler and Vendelmans, 1990; Kaji, Kida and Morimoto, 1992; Utsuro, Matsumoto and Nagao, 1992; Matsumoto, Ishimoto and Utsuro, 1993; Utsuro, Matsumoto and Nagao, 1993) . Compared with the statistical approach, those works are quite different in that they use word correspondence information available in hand-written bilingual dictionaries and try to extract structured linguistic knowledge such as structured translation patterns and case frames of verbs. For example, in Matsumoto et al. (1993), we proposed a method for finding structural matching of parallel sentences, making use of word level similarities calculated from a bilingual dictionary and a thesaurus. Then, those structurally matched parallel sentences are used as a source for acquiring lexical knowledge such as verbal case frames (Utsuro et al., 1992; Utsuro et al., 1993).

With the aim of acquiring those structured linguistic knowledge, this paper describes a unified framework for bilingual text matching by combining existing hand-written bilingual dictionaries and statistical techniques. The process of *bilingual text matching* consists of two major steps: *sentence alignment* and *structural matching of bilingual sentences*. In those two steps, we use word correspondence information, which is available in hand-written bilingual dictionaries, or not included in bilingual dictionaries but estimated with statistical techniques.

The reasons why we take the approach of combining bilingual dictionaries and statistics are as follows: Statistical techniques are limited since 1) they require bilingual texts to be long enough for extracting useful statistics, while we need to acquire structured linguistic knowledge even from bilingual texts of about 100 sentences, 2) even with bilingual texts long enough for statistical techniques, useful statistics can not be extracted for low frequency words. For the reasons 1) and 2), the use of bilingual dictionaries is inevitable in our application. On the other hand, existing hand-written bilingual dictionaries are limited in that available dictionaries are only for daily words and usually domain specific on-line bilingual dictionaries are not available. Thus, statistical techniques are also inevitable for extracting domain specific word correspondence information not included in existing bilingual dictionaries.

At present, we are at the starting point of combining existing bilingual dictionaries and statistical techniques. Therefore, as statistical techniques for estimating word correspondences not included in bilingual dictionaries, we decided to adopt techniques as simple as possible, rather than techniques based-on complex probabilistic translation models such as in Brown et al. (1990), Brown, Pietra, Pietra and Mercer (1993), and Chen (1993). What we adopt are simple co-occurrence-frequency-based techniques in Gale and Church (1991) and Kay and Röscheisen (1993). As techniques for sentence alignment, we adopt also quite a simple method based-on the number of word correspondences, without any probabilistic translation models.

In the following sections, we illustrate the specifications of our bilingual text matching framework.

2 The Framework of Bilingual Text Matching

The overall framework of bilingual text matching is depicted in **Fig.** 1. Although our framework is implemented for Japanese and English, it is language independent.

First, bilingual texts are aligned at sentence level using word correspondence information which is available in bilingual dictionaries or estimated by statistical techniques. *"Statistical estimation"* at text level indicates that length-based statistical techniques are applied if necessary. (At present, they are not implemented.) *"Statistical estimation"* at sentence level indicates that word-to-word correspondences are estimated by statistical techniques. Then, each monolingual sentence is parsed into a disjunctive dependency structure and structurally matched using word correspondence information. In the course of structural matching, lexical and syntactic ambiguities of monolingual sentences are resolved. Finally, from the matching results, monolingual lexical knowledge and translation patterns are acquired.

So far, we have implemented the following: sentence alignment based-on word corre-

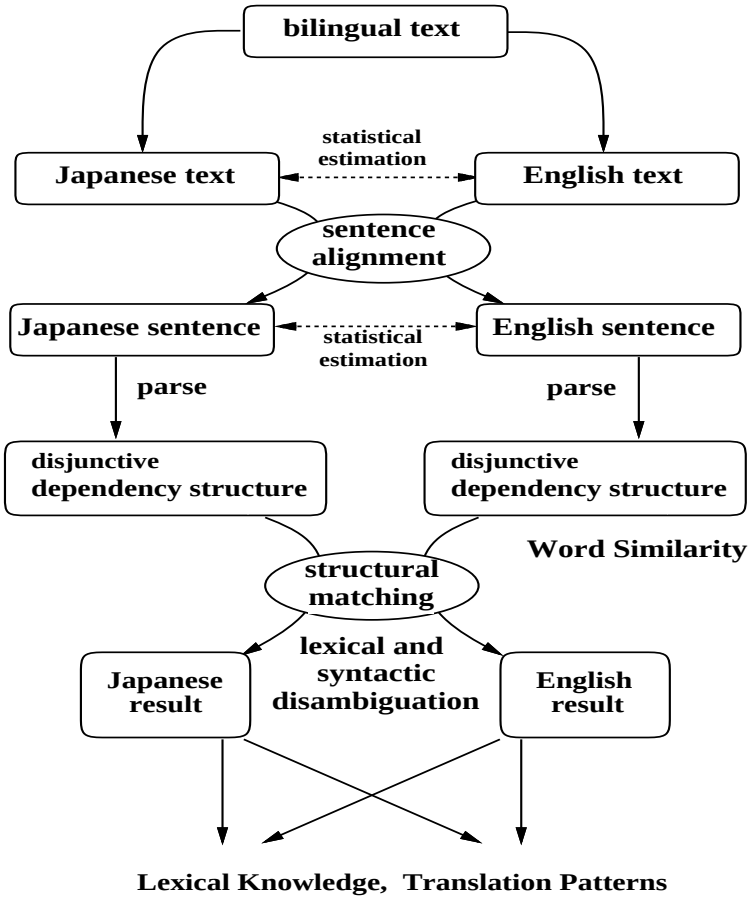

Fig. 1: The Framework of Bilingual Text Matching

spondence information, word correspondence estimation by co-occurrence-frequency-based methods in Gale and Church (1991) and Kay and Röscheisen (1993), structural matching of parallel sentences (Matsumoto et al., 1993), case frame acquisition of Japanese verbs (Utsuro et al., 1993), and acquisition of translation rules between Japanese and English (Matsumoto and Kitamura, 1994). In the remainder of this paper, we describe the specifications of sentence alignment and word correspondence estimation in sections 3 and 4, then report the results of small experiments and evaluate our framework in section 5.

3 Sentence Alignment

3.1 Bilingual Sentence Alignment Problem

In this section, we formally define the problem of bilingual sentence alignment.[1]

[1]In this paper, we do not describe paragraph alignment process. For the moment, our paragraph alignment program is not reliable enough and the results of sentence alignment are better without paragraph

Let S be a text of n sentences of a language, and T be a text of m sentences of another language and suppose that S and T are translation of each other:

$$\begin{aligned} S &= s_1, s_2, \ldots, s_n \\ T &= t_1, t_2, \ldots, t_m \end{aligned}$$

Let p be a pair of minimal corresponding segments in texts S and T. Suppose that p consists of x sentences $s_{a-x+1}, \ldots, s_a$ in S and y sentences $t_{b-y+1}, \ldots, t_b$ in T and is denoted by the following:

$$p = \langle a, x; b, y \rangle$$

Note that x and y could be 0. In this paper, we call the pair of minimal corresponding segments in bilingual texts a sentence *bead*.[2] Then, sentences in bilingual texts of S and T are aligned into a sequence P of sentence beads:

$$P = p_1, p_2, \ldots, p_k$$

We put some restriction on possibilities of sentence alignment. We assume that each sentence belongs to only one sentence bead and order constraints must be preserved in sentence alignment. Supposing $p_i = \langle a_i, x_i; b_i, y_i \rangle$, those constraints are expressed in the following:

$$\begin{aligned} a_0 &= 0 \quad , \quad b_0 = 0 \\ a_i &= a_{i-1} + x_i \quad , \quad b_i = b_{i-1} + y_i \quad (1 \leq i \leq k) \end{aligned}$$

Suppose that a scoring function h can be defined for estimating the validity of each sentence bead p_i. Then, bilingual sentence alignment problem can be defined as an optimization problem that finds a sequence P of sentence beads which optimizes the total score H of the sequence P:

$$H(P) = H_h(h(p_1), \ldots, h(p_k))$$

3.2 Bilingual Sentence Alignment using Word Correspondence Information

In this section, we describe the specification of our sentence alignment method based-on word correspondence information.[3]

alignment than with paragraph alignment. Since bilingual texts in our bilingual corpus are not so long, the computational cost of sentence alignment is not serious problem even without paragraph alignment.

[2]The term *bead* is taken from Brown et al. (1991).

[3]We basically use dictionary-based bilingual sentence alignment method originally reported in Murao (1991). The work in Murao (1991) was done under the supervision of Prof. M. Nagao and Prof. S. Sato (JAIST, East).

3.2.1 Score of Sentence Bead

Before aligning sentences in bilingual texts, content words are extracted from each sentence (after each sentence is morphologically analyzed if necessary), and word correspondences are found using both bilingual dictionaries and statistical information source for word correspondence. Then, using those word correspondence information, the score h of a sentence bead p is calculated as follows.

First, supposing $p = \langle a, x; b, y \rangle$, and let $n_s(a, x)$ and $n_t(b, y)$ be the numbers of content words in the sequences of sentences $s_{a-x+1}, \ldots, s_a$ and $t_{b-y+1}, \ldots, t_b$ respectively, and $n_{st}(p)$ be the number of corresponding word pairs in p. Then, the score h of p is defined as the ratio of $n_{st}(p)$ to the sum of $n_s(a, x)$ and $n_t(b, y)$:

$$h(p) \quad = \quad \frac{n_{st}(p)}{n_s(a, x) + n_t(b, y)}$$

3.2.2 Dynamic Programming Method

Let P_i be the sequence of sentence beads from the beginning of the bilingual text up to the bead p_i:

$$P_i \quad = \quad p_1, p_2, \ldots, p_i$$

Then, we assume that the score $H(P_i)$ of P_i follows the recursion equation below:

$$H(P_i) \quad = \quad H(P_{i-1}) + h(p_i) \tag{1}$$

Let $H_m(a_i, b_i)$ be the maximum score of aligning a part of S (from the beginning up to the $a_i (= a_{i-1} + x_i)$ - th sentence) and a part of T (from the beginning up to $b_i (= b_{i-1} + y_i)$ - th sentence). Then, Equation 1 is transformed into:

$$H_m(a_i, b_i) \quad = \quad \max_{x_i, y_i} \Big\{ H_m(a_i - x_i, b_i - y_i) + h(\langle a_i, x_i; b_i, y_i \rangle) \Big\}$$

where the initial condition is:

$$H_m(a_0, b_0) \quad = \quad H_m(0, 0) \quad = \quad 0$$

We limit the pair (x_i, y_i) of the numbers of sentences in a sentence bead to some probable ones. For the moment, we allow only 1-1, 1-2, 1-3, 1-4, 2-2 as pairs of the numbers of sentences:

$$\begin{aligned} (x_i, y_i) \quad \in \quad & \{(1,1), (1,2), (2,1), (1,3), \\ & (3,1), (1,4), (4,1), (2,2)\} \end{aligned}$$

This optimization problem is solvable as a standard problem in dynamic programming. Dynamic programming is applied to bilingual sentence alignment in most of previous works (Brown et al., 1991; Gale and Church, 1993; Chen, 1993).

4 Word Correspondence Estimation

In this section, first we describe estimation functions based-on co-occurrence frequencies. Then, we show how to incorporate word correspondence information available in bilingual dictionaries and to estimate word correspondences not included in bilingual dictionaries. Finally, we describe the threshold function for extracting corresponding word pairs.

4.1 Estimation Function

In the following, we assume that sentences in the bilingual text are already aligned.

Let w_s and w_t be words in the texts S and T respectively, we define the following frequencies:

$$\begin{aligned} freq(w_s, w_t) &= \text{(frequency of } w_s \text{ and } w_t\text{'s co-occurring in a sentence bead)} \\ freq(w_s) &= \text{(frequency of } w_s\text{)} \\ freq(w_t) &= \text{(frequency of } w_t\text{)} \\ N &= \text{(total number of sentence beads)} \end{aligned}$$

Then, estimation functions of Gale's (Gale and Church, 1991) and Kay's (Kay and Röscheisen, 1993) are given as below.

4.1.1 Gale's Method

Let $a \sim d$ be as follows:

$$\begin{aligned} a &= freq(w_s, w_t) \\ b &= freq(w_s) - freq(w_s, w_t) \\ c &= freq(w_t) - freq(w_s, w_t) \\ d &= N - a - b - c \end{aligned}$$

Then, the validity of word correspondence w_s and w_t is estimated by the following value:

$$\begin{aligned} h_g(w_s, w_t) &= \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \\ &= \frac{(ad - bc)^2}{freq(w_s)freq(w_t)(N - freq(w_s))(N - freq(w_t))} \end{aligned}$$

4.1.2 Kay's Method

The validity of word correspondence w_s and w_t is estimated by the following value:

$$h_k(w_s, w_t) = \frac{2freq(w_s, w_t)}{freq(w_s) + freq(w_t)}$$

4.2 Incorporating Bilingual Dictionary

By incorporating word correspondence information available in bilingual dictionaries, it becomes easier to estimate word correspondences not included in bilingual dictionaries.

Let w_s be a word in the text S and w_t, w_t' be words in the text T. Suppose that the correspondence of w_s and w_t is included in bilingual dictionaries, while the correspondence of w_s and w_t' is not included. Then the problem is to estimate the validity of word correspondence of w_s and w_t'.

Let $freq(w_s, w_t)$, $freq(w_s, w_t')$, $freq(w_s)$, $freq(w_t)$, and $freq(w_t')$ be the same as above, and $freq(w_s, w_t, w_t')$ be the frequency of w_s, w_t, and w_t''s co-occurring in a sentence bead. Then, we solve the problem above by defining $freq'(w_s, w_t')$, $freq'(w_s)$, $freq'(w_t')$, and N' which become the inputs to Gale's method or Kay's method. We describe two different ways of defining those values.

Estimation I

One is to estimate all the word correspondences equally except that the co-occurrence of w_s and w_t is preferred to that of w_s and w_t'. $freq'(w_s, w_t')$, $freq'(w_s)$, $freq'(w_t')$, and N' are given below:[4]

$$
\begin{aligned}
freq'(w_s, w_t') &= freq(w_s, w_t') - \sum_{w_t} freq(w_s, w_t, w_t') \\
freq'(w_s) &= freq(w_s) \\
freq'(w_t') &= freq(w_t') \\
N' &= N \\
(\ freq'(w_s, w_t) &= freq(w_s, w_t)\ )
\end{aligned}
$$

When w_s, w_t, and w_t' are co-occurring in a sentence bead, the co-occurrence of w_s and w_t is preferred and that of w_s and w_t' is ignored. Thus, $freq'(w_s, w_t')$ is obtained by subtracting the frequency of all those cases from the real co-occurrence frequency of w_s and w_t'. But, $freq'(w_s)$ and $freq'(w_t')$ are the same as the real frequencies and the estimated word correspondences reflect the real co-occurrence frequencies in the input text. (Compare with **Estimation II**.) Word correspondences both included and not included in bilingual dictionaries are equally estimated their validities.

[4]It can happen that, within a sentence bead, one word of a language has more than one corresponding words of the opposite language and all the correspondences are included in bilingual dictionaries. In that case, formalizations in this section need some modifications.

Estimation II

The other is to remove from the input text all the co-occurrences of word pairs included in bilingual dictionaries. $freq'(w_s, w'_t), freq'(w_s), freq'(w'_t)$, and N' are given below:

$$\begin{aligned} freq'(w_s, w'_t) &= freq(w_s, w'_t) - \sum_{w_t} freq(w_s, w_t, w'_t) \\ freq'(w_s) &= freq(w_s) - \sum_{w_t} freq(w_s, w_t) \\ freq'(w'_t) &= freq(w'_t) - \sum_{w'_s} freq(w'_s, w'_t) \\ & \quad \text{(the correspondence of } w'_s \text{ and } w'_t \\ & \quad \text{is included in bilingual dictionaries)} \\ N' &= N \end{aligned}$$

With this option, after all the co-occurrences of word pairs included in bilingual dictionaries are removed from the input text, word correspondences not included in bilingual dictionaries are estimated their validities.

In the following sections, we temporarily adopt **Estimation I** for estimating word correspondences not included in bilingual dictionaries. It is necessary to further investigate and compare the two estimation methods with large-scale experiments.

4.3 Threshold Function

As a threshold function for extracting appropriate corresponding word pairs, we use a hyperbolic function of word frequency and estimated value for word correspondence.

At first, we define the following variables and constants:[5]

$$\begin{aligned} x &= \text{(co-occurrence frequency)} \\ y &= \text{(estimated value for word correspondence)} \\ a &= \text{(constant for eliminating low frequency words)} \\ & \quad \left(\text{ 1.0 for both } h_g \text{ and } h_k \right) \\ b &= \text{(constant for eliminating words with low estimated value)} \\ & \quad \left(\text{ 0.1 for } h_g \text{ and 0.3 for } h_k \right) \\ c &= \text{(lower bound of word frequency)} \\ & \quad \left(\text{ 2.5 for both } h_g \text{ and } h_k \right) \end{aligned}$$

[5]Note that values for constants are determined temporarily and need further investigation with large-scale experiments. Especially, constants related to word frequency have to be tuned to the length of texts.

Then, the threshold function $g(x, y)$ is defined as below:

$$g(x, y) = \frac{x(y - b)}{a} \quad (x > c)$$

And the condition for extracting corresponding word pairs is given below:

$$g(x, y) > 1 \quad , \quad x > c$$

When using extracted word correspondences in sentence alignment and structural matching, at present we ignore the estimated values and use estimated word correspondences and word correspondences in bilingual dictionaries equally.

5 Experiment and Evaluation

In this section, we report the results of a small experiment on aligning sentences in bilingual texts and statistically estimating word correspondences.

The sentence alignment program and the word correspondence estimation program are named *AlignCO*. The processing steps of *AlignCO* are as follows:

1. Given a bilingual text, content words are extracted from each sentence.
2. A Japanese-English dictionary of about 50,000 entries is consulted and word correspondence information is extracted for content words of each sentence.
3. The sentence alignment program named *AlignCO/A* aligns sentences in the input text by the method stated in section 3.2.
4. Given the aligned sentences in the bilingual text, the word correspondence estimation program named *AlignCO/C* estimates word correspondences which are not included in the Japanese-English dictionary with option **Estimation I** in section 4.2.
5. Combining word correspondence information available in the Japanese-English dictionary and estimated by *AlignCO/C*, sentences in the input text are realigned.

As input Japanese-English bilingual texts, we use two short texts of different length — 1) *"The Dilemma of National Development and Democracy"* (305 Japanese sentences and 300 English sentences, henceforth *"dilemma"*), 2) *"Pacific Asia in the Post-Cold-War World"* (134 Japanese sentences and 123 English sentences, henceforth *"cold-war"*). Since the results of Gale's method and Kay's method did not differ so much, we show the result of Gale's method only.

5.1 Sentence Alignment

The followings are five best results of sentence alignment before and after estimating word correspondences not included in the Japanese-English dictionary. The results are improved after estimating word correspondences not included in the bilingual dictionary.

"dilemma"

	number of errors (five best solutions)					average error rate
1st trial	18	18	19	19	16	6.3%
2nd trial	13	14	14	15	13	4.8%

"cold-war"

	number of errors (five best solutions)					average error rate
1st trial	5	6	4	7	8	4.9%
2nd trial	4	2	2	5	0	2.1%

5.2 Word Correspondence Estimation

We classify the estimated word correspondences into three categories, "correct", "part of phrase", and "wrong". "part of phrase" means that the estimated word correspondence can be considered as part of corresponding phrases. "error rate" is the ratio of the number of "wrong" word correspondences to the total number.

"dilemma"

total	correct	part of phrase	wrong	error rate
87	53	30	4	4.6%

"cold-war"

total	correct	part of phrase	wrong	error rate
37	19	10	8	21.6%

The result of *"dilemma"* is better than that of *"cold-war"*. This is because the former is longer than the latter.

The followings are example word correspondences of each category where f_s, f_t, and f_{st} are $freq(w_s)$, $freq(w_t)$, and $freq(w_s, w_t)$ respectively. The parenthesized correspondence is not extracted by the threshold function.

correct

w_s	w_t	h_g	f_s	f_t	f_{st}
スルタン	sultan	0.75	4	3	3
報道	press	0.80	5	4	4
自由	liberal	0.64	20	15	14
経済	economic	0.32	33	19	15

part of phrase

w_s	w_t	h_g	f_s	f_t	f_{st}
文	civilian	1.00	6	6	6
文	supremacy	0.83	6	5	5
民	civilian	0.69	6	6	5
民	supremacy	0.83	6	5	5
優先	supremacy	0.44	4	5	3
(優先	civilian	0.37	4	6	3)

wrong

w_s	w_t	h_g	f_s	f_t	f_{st}
意味	does	0.49	6	3	3
太平洋	and	0.41	47	62	5

Most of "correct" correspondences are proper names like "**スルタン** – sultan", or those which have different parts of speech, like "**自由** (noun) – liberal (adjective)" and "**経済** (noun) – economic (adjective)", or those which can be considered as translation equivalents but not included in the Japanese-English dictionary, like "**報道** (news) – press".

The examples of "part of phrase" form a phrase correspondence "**文民優先** – civilian supremacy".

The former "wrong" correspondence "**意味** (meaning) – does" comes from the correspondence of long distance dependent phrases "**意味**, **する** – does ∼ mean". The latter "wrong" correspondence "**太平洋** (pacific ocean) – and" is extracted by Gale's method because both $freq$(**太平洋**) and $freq$(and) are high and close to the total number of sentence beads. This correspondence is not extracted by Kay's method.

Then, in **Fig.** 2, we illustrate the relation between the estimated value $h_g(w_s, w_t)$ of Gale's method and the co-occurrence frequency $freq(w_s, w_t)$ for the text *"dilemma"*. The threshold function seems optimized so that it extracts as many word correspondences of the category "correct" and "part of phrase" as possible, and extracts as few word correspondences of the category "wrong" as possible.

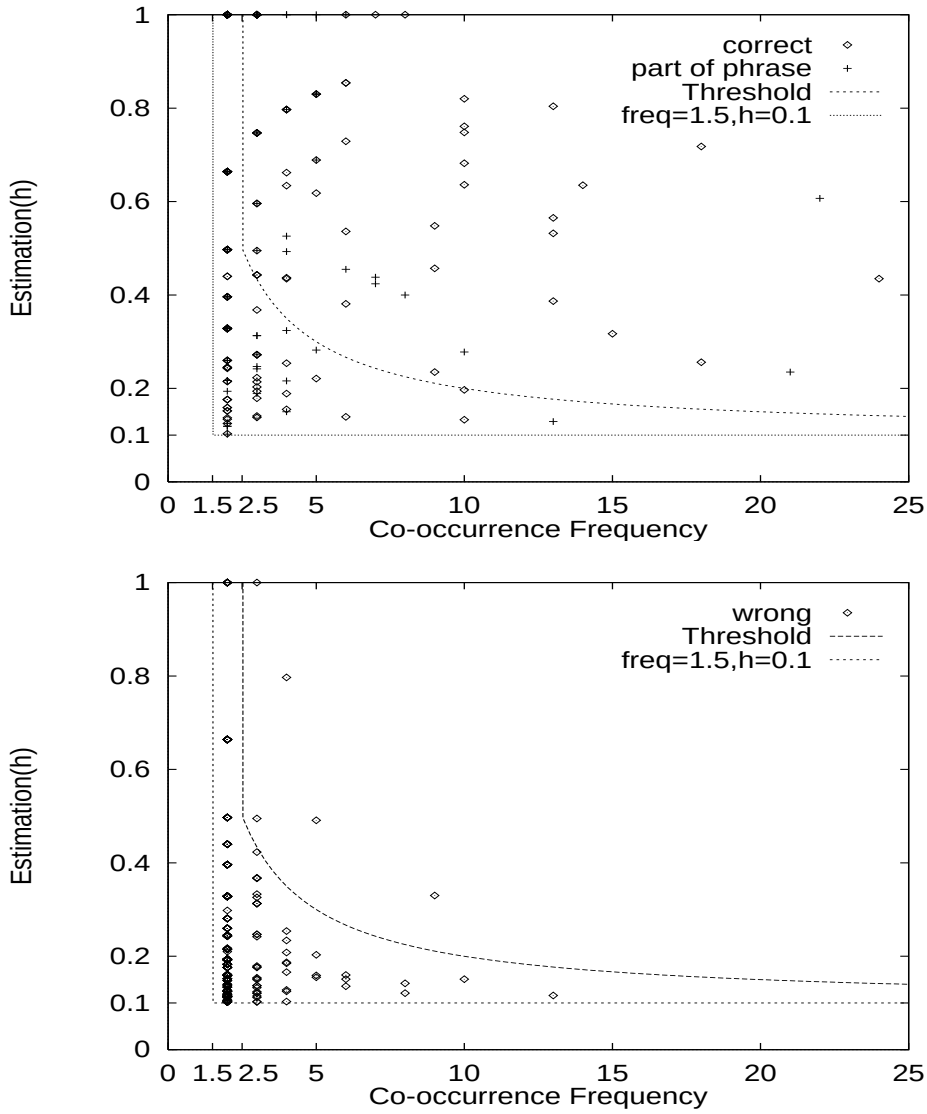

Fig. 2: Estimation per Co-occurrence Frequency of Word Correspondences (*"dilemma"*)

6 Concluding Remarks

This paper described a unified framework for bilingual text matching by combining existing hand-written bilingual dictionaries and statistical techniques. Especially, we described a method for aligning sentences using word correspondence information, and a method for estimating word correspondences not included in bilingual dictionaries.

Estimated word correspondence information will improve the results of structural matching of bilingual sentences. It will be reported in the future. With the same techniques as those for estimating word correspondences, it is quite easy to estimate correspondences of phrases such as noun phrases and idiomatic expressions. Then, the results of structural matching will be much more improved.

In order to improve the accuracy of sentence alignment, we need to combine our word-correspondence-based method with those length-based methods in Brown et al. (1991) and Gale and Church (1993). In the case of Japanese-English texts, the word-based method in Brown et al. (1991) might be better than the character-based method in Gale and Church (1993).

References

Brown, P. F., Cocke, J., Pietra, S. A., Pietra, V. J. D. et al. (1990). A statistical approach to machine translation, *Computational Linguistics* **16**(2): 79–85.

Brown, P. F., Lai, J. C. and Mercer, R. L. (1991). Aligning sentences in bilingual corpora, *Proceedings of the 29th Annual Meeting of ACL*, pp. 169–176.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D. and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics* **19**(2): 263–311.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information, *Proceedings of the 31th Annual Meeting of ACL*, pp. 9–16.

Church, K. W. (1993). Char_align: A program for aligning parallel texts at the character level, *Proceedings of the 31th Annual Meeting of ACL*, pp. 1–8.

Dagan, I., Itai, A. and Schwall, U. (1991). Two languages are more informative than one, *Proceedings of the 29th Annual Meeting of ACL*, pp. 130–137.

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

Gale, W. and Church, K. (1991). Identifying word correspondences in parallel texts, *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pp. 152–157.

Kaji, H., Kida, Y. and Morimoto, Y. (1992). Learning translation templates from bilingual text, *Proceedings of the 14th COLING*, pp. 672–678.

Kay, M. and Röscheisen, M. (1993). Text-translation alignment, *Computational Linguistics* **19**(1): 121–142.

Klavans, J. and Tzoukermann, E. (1990). The BICORD System: Combining lexical information from bilingual corpora and machine readable dictionaries, *Proceedings of the 13th COLING*, Vol. 3, pp. 174–179.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora, *Proceedings of the 31th Annual Meeting of ACL*, pp. 17–22.

Matsumoto, Y. and Kitamura, M. (1994). Acquisition of translation rules from parallel corpus, *Information Science Technical Report 94014*, Graduate School of Information Science, Nara Institute of Science and Technology.

Matsumoto, Y., Ishimoto, H. and Utsuro, T. (1993). Structural matching of bilingual texts, *Proceedings of the 31th Annual Meeting of ACL*, pp. 23 – 30.

Murao, H. (1991). Studies on bilingual text alignment, *Bachelor Thesis*, Kyoto University. (in Japanese).

Sadler, V. and Vendelmans, R. (1990). Pilot implementation of a bilingual knowledge bank, *Proceedings of the 13th COLING*, Vol. 3, pp. 449–451.

Utsuro, T., Matsumoto, Y. and Nagao, M. (1992). Lexical knowledge acquisition from bilingual corpora, *Proceedings of the 14th COLING*, pp. 581–587.

Utsuro, T., Matsumoto, Y. and Nagao, M. (1993). Verbal case frame acquisition from bilingual corpora, *Proceedings of the 13th IJCAI*, pp. 1150–1156.