Natural Language Engineering **3** (1): 1–14. Printed in the United Kingdom

© 1997 Cambridge University Press

High-performance bilingual text alignment using statistical and dictionary information

MASAHIKO HARUNO and TAKEFUMI YAMAZAKI

NTT Communication Science Laboratories
1-1 Hikarinooka Yokosuka Kanagawa 239, Japan

(*Received 14 February 1996; revised 24 February 1997*)

Abstract

This paper describes an accurate and robust text alignment system for structurally different languages. Among structurally different languages such as Japanese and English, there is a limitation on the amount of word correspondences that can be statistically acquired. The main reason for this is the systems of functional (closed) words are quite different in the two languages. The proposed method makes use of two kinds of word correspondences in aligning bilingual texts. One is a bilingual dictionary of general use. The other is the word correspondences that are statistically acquired in the alignment process. Our method gradually determines sentence pairs (anchors) that correspond to each other by relaxing parameters. The method, by combining two kinds of word correspondences, achieves adequate word correspondences for complete alignment. As a result, texts of various length and of various genres in structurally different languages can be aligned with high precision. Experimental results show our system outperforms conventional methods for various kinds of Japanese–English texts.

1 Introduction

Corpus-based approaches based on bilingual texts are promising for various applications (i.e. lexical knowledge extraction (Kupiec 1993; Matsumoto *et al.*, 1993; Smadja and McKeown 1994; Dagan and Church 1994; Kumano and Hirakawa 1994), machine translation (Brown *et al.* 1993; Sato and Nagao 1990; Kaji *et al.* 1992) and information retrieval (Sato 1992)). Most of these works assume voluminous aligned corpora.

Many methods have been proposed to align bilingual corpora. One of the major approaches is based on the statistics of simple features such as sentence length in words (Brown *et al.* 1991) or in characters (Gale and Church 1993), and character sequences themselves (Church 1993). These techniques are widely used because they can be implemented in an efficient and simple way through dynamic programing. However, their main targets are rigid translations (i.e. Hansard corpus) that are almost literal translations. In addition, the texts being aligned were structurally similar European languages (i.e. English–French, English–German).

The simple-feature based approaches don't work in flexible translations for

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

structurally different languages such as Japanese and English, mainly for the following two reasons. One is the difference in the character systems of the two languages. Japanese has three types of characters (Hiragana, Katakana and Kanji), each of which has different amounts of information. In contrast, English has only one type of character. The other is the grammatical difference of the two languages. In particular, the systems of functional (closed) words are quite different from language to language. Among European languages, articles and prepositions offer strong clues for alignment. Japanese has a quite different system of closed words. Thus, it is impossible in general to apply the simple feature-based methods to Japanese–English translations.

One alternative alignment method is the lexicon-based approach that makes use of the word-correspondence knowledge of the two languages. Kay and Roscheisen (1993) proposed a relaxation method to iteratively align bilingual texts using the word correspondences acquired during the alignment process. Although the lexicon-based method works well among European languages, the method does not work in aligning structurally different languages. In Japanese–English translations, the method does not capture enough word correspondences to permit alignment. As a result, it can align only some of the two texts. This is mainly because the method cannot use clues of functional words such as articles and prepositions. The number of confident word correspondences of content words is not enough for complete alignment. The problem cannot be addressed as long as the method relies only on statistics. Other methods in the lexicon-based approach embed lexical knowledge into stochastic models (Chen 1993; Wu 1994), but these methods were tested using just rigid translations (i.e. Hansard corpus).

To tackle the problem, we describe in this paper a text alignment system that uses both statistics and bilingual dictionaries at the same time. Bilingual dictionaries are now widely available on-line due to advances in CD-ROM technologies. For example, English–Spanish, English–French, English–German, English–Japanese, Japanese–French, Japanese–Chinese and other dictionaries are now commercially available. It is reasonable to make use of these dictionaries in bilingual text alignment. The pros and cons of statistics and online dictionaries are discussed below. They show that statistics and on-line dictionaries are complementary in terms of bilingual text alignment.

Statistics Merit Statistics is robust in the sense that it can extract context-dependent usage of words and that it works well even if word segregation[1] is not correct.

Statistics Demerit The number of word correspondences acquired by statistics is not enough for complete alignment.

Dictionaries Merit They can contain the information about words that appear only once in the corpus.

Dictionaries Demerit They cannot capture context-dependent keywords in the corpus and are weak against incorrect word segregation. Entries in the dictionaries differ from author to author and are not always the same as those in the corpus.

[1] In Japanese, there are no explicit delimiters between words. The first task for alignment is, therefore, to divide the text stream into words.

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

Our system iteratively aligns sentences by using statistical and on-line dictionary word correspondences. The characteristics of the system are as follows:

- The system performs well and is robust for various lengths (especially short) and various genres of texts.
- The system is very economical because it assumes only online-dictionaries of general use and doesn't require the labour-intensive construction of domain-specific dictionaries.
- The system is extendable by registering statistically acquired word correspondences into user dictionaries.

We will hereafter deal with Japanese–English translations, although the proposed method is language independent.

The proposed alignment method is incorporated in a graphical alignment environment BACCS (Bilingual Aligned Corpus Construction System). BACCS offers a user interface that enables the user to easily confirm and modify alignment results and register word correspondences created by the alignment program into a user dictionary.

The construction of the paper is as follows. First, section 2 offers an overview of our alignment system. Section 3 describes the entire alignment algorithm in detail. Section 4 reports experimental results for various kinds of Japanese–English texts, including newspaper editorials, scientific papers and critiques on economics. The evaluation is performed from two points of view: precision-recall of alignment and word correspondences acquired during alignment. Section 5 describes our graphical alignment environment BACCS in which user can see the alignment results, and easily modify the results and the user dictionary. Section 6 concerns related work, and section 7 concludes the paper.

2 System overview

Figure 1 overviews our alignment system BACCS. The input to the system is a pair of Japanese and English texts, one the translation of the other. First, sentence boundaries are found in both texts using finite state transducers. The texts are then part-of-speech (POS) tagged and separated into original form words.[2] Original forms of English words are determined by 80 rules using the POS information. From the word sequences, we extract only nouns, adjectives, adverbs, verbs and unknown words (only in Japanese) because Japanese and English closed words are different and impede text alignment. This pre-processing operation can be easily implemented with regular expressions.

The initial state of the algorithm is a set of already known anchors (sentence pairs). These are determined by article boundaries, section boundaries and paragraph boundaries. In the most general case, initial anchors are only the first and final

[2] We use in this phase the JUMAN morphological analysing system (Kurohashi *et al.*, 1994) for tagging Japanese texts and Brill's transformation-based tagger (Brill, 1992, 1994) for tagging English texts. We would like to thank all people concerned for providing us with the tools.

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

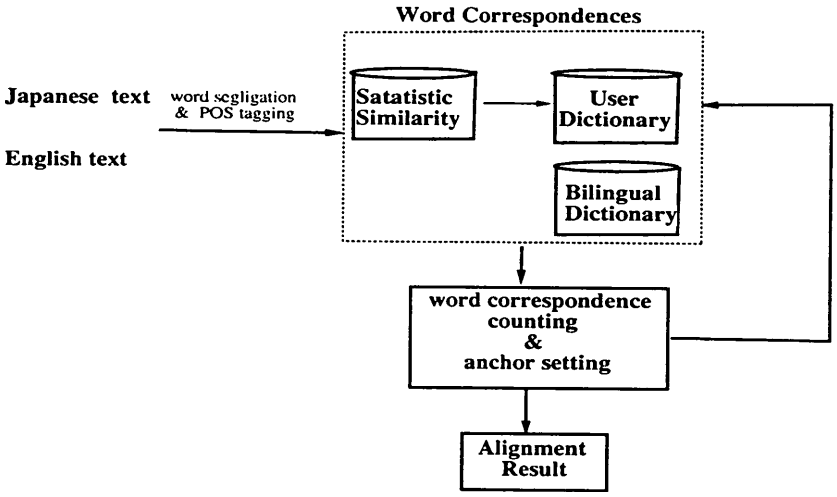


Fig. 1. Overview of the alignment system BACCS.

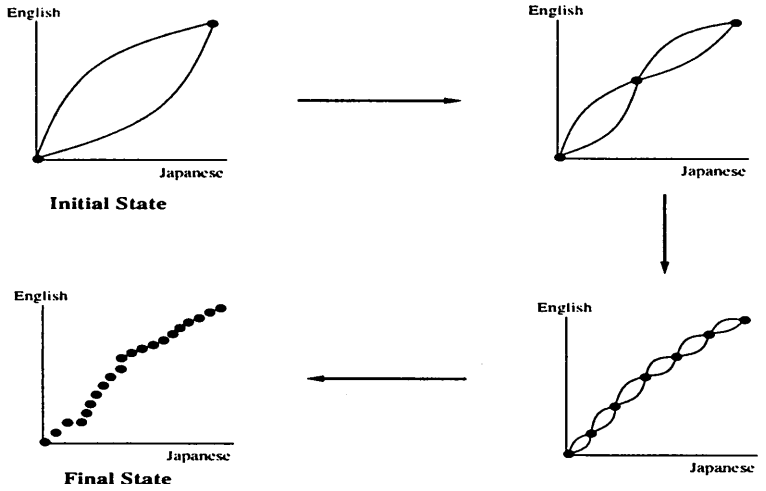


Fig. 2. Alignment process.

sentence pairs of both texts, as depicted in Figure 2. Possible sentence correspondences are determined from the anchors. Intuitively, the number of possible correspondences for a sentence is small near anchors, while large between the anchors. In this phase, the most important point is that each set of possible sentence correspondences should include the correct correspondence.

The main task of the system is to find anchors from the possible sentence correspondences by using two kinds of word correspondences: statistical word correspondences and word correspondences as held in a bilingual dictionary.[3] By

[3] Adding to the bilingual dictionary of general use, users can reuse their own dictionaries created in previous sessions.

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

using both correspondences, the sentence pair whose correspondences exceeds a predefined threshold is judged as an anchor. These newly found anchors make word correspondences more precise in the subsequent sessions. By repeating this anchor setting process with threshold reduction, sentence correspondences are gradually determined from confident pairs to non-confident pairs. The gradualism of the algorithm makes it robust because anchor-setting errors in the last stage of the algorithm have little effect on overall performance. The output of the algorithm is the alignment result (a sequence of anchors) and word correspondences as by-products.

By using BACCS, the user can see and modify sentence correspondences graphically and easily extend the system by registering word correspondences into user dictionaries.

3 Algorithms

3.1 Statistics used

In this section, we describe the statistics used to decide word correspondences. From many similarity metrics applicable to the task, we choose mutual information and *t-score* because the relaxation of parameters can be controlled in a sophisticated manner. Mutual information represents the similarity on the occurrence distribution and *t-score* represents the confidence of the similarity. These two parameters permit more effective relaxation than the single parameter used in conventional methods (Kay and Roscheisen 1993).

Our basic data structure is the Alignable Sentence Matrix (ASM) and the Anchor Matrix (AM). ASM represents possible sentence correspondences and consists of ones and zeros. A one in ASM indicates the intersection of the column and row constitutes a possible sentence correspondence. On the contrary, AM is introduced to represent how a sentence pair is supported by word correspondences. The *i-j* element of AM indicates how many times the corresponding words appear in the *i-j* sentence pair. As alignment proceeds, the number of ones in ASM reduces, while the elements of AM increase.

Let p_i be a sentence set comprising the ith Japanese sentence and its possible English correspondences as depicted in Figure 3. For example, p_2 is the set comprising

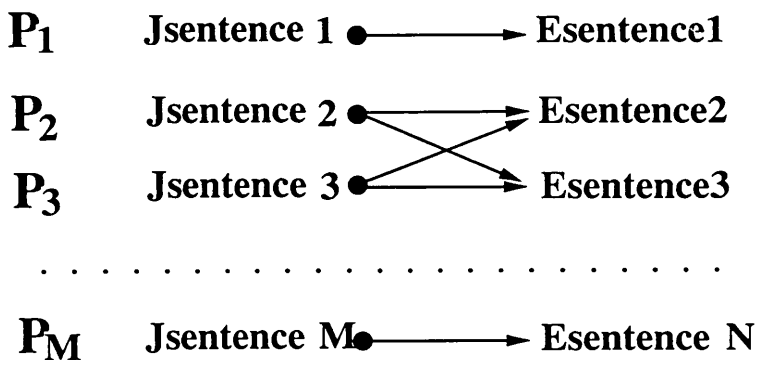


Fig. 3. Possible sentence correspondences.

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

J sentence$_2$, *Esentence*$_2$ and *Esentence*$_3$, which means *Jsentence*$_2$ has the possibility of aligning with *Esentence*$_2$ or *Esentence*$_3$. The p_is can be directly derived from ASM.

We introduce the contingency matrix (Fung and Church 1994) to evaluate the similarity of word occurrences. Consider the contingency matrix shown Table 1, between Japanese word w_{jpn} and English word w_{eng}. The contingency matrix shows (a) the number of p_is in which both w_{jpn} and w_{eng} were found, (b) the number of p_is in which just w_{eng} was found, (c) the number of p_is in which just w_{jpn} was found, and (d) the number of p_is in which neither word was found. Note here that p_is overlap each other and w_{eng} may be double counted in the contingency matrix. We count each w_{eng} only once, even if it occurs more than twice in p_is.

If w_{jpn} and w_{eng} are good translations of one another, a should be large, and b and c should be small. In contrast, if the two are not good translations of each other, a should be small, and b and c should be large. To make this argument more precise, we introduce mutual information:

$$\log \frac{prob(w_{jpn}, w_{eng})}{prob(w_{jpn})\, prob(w_{eng})}$$

The probabilities are:

$$prob(w_{jpn}) = \frac{a+c}{a+b+c+d} = \frac{a+c}{M}$$

$$prob(w_{eng}) = \frac{a+b}{a+b+c+d} = \frac{a+b}{M}$$

$$prob(w_{jpn}, w_{eng}) = \frac{a}{a+b+c+d} = \frac{a}{M}$$

Unfortunately, mutual information is not reliable when the number of occurrences is small. Many words occur just once which weakens the statistics approach. To avoid this, we employ *t-score*, defined below, where M is the number of Japanese sentences. Insignificant mutual information values are filtered out by thresholding *t-score*. For example, *t-scores* above 1·65 are significant at the $p > 0{\cdot}95$ confidence level:

$$t \approx = \frac{prob(w_{jpn}, w_{eng}) - prob(w_{jpn})\, prob(w_{eng})}{\sqrt{\frac{1}{M} prob(w_{jpn}, w_{eng})}}$$

3.2 Basic alignment algorithm

Our basic algorithm is an iterative adjustment of the Anchor Matrix (AM) using the Alignable Sentence Matrix (ASM). Given an ASM, mutual information and *t-score* are computed for all word pairs in possible sentence correspondences. A word combination exceeding a predefined threshold is judged as a word correspondence. To find new anchors, we combine these statistical word correspondences with the word correspondences in a bilingual dictionary. Each element of AM, which represents a sentence pair, is updated by adding the number of word correspondences in the sentence pair. A sentence pair containing more than a predefined number of

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

Table 1. *Contingency matrix*

	w_{jpn}	
w_{eng}	a	b
	c	d

corresponding words is determined to be a new anchor. The detailed algorithm is as follows.

3.2.1 Constructing initial ASM

This step constructs the initial ASM. If the texts contain M and N sentences, respectively, the ASM is an $M \times N$ matrix. First, we decide a set of anchors using article boundaries, section boundaries and so on. In the most general case, initial anchors are the first and last sentences of both texts as depicted in Figure 2. Next, possible sentence correspondences are generated. Intuitively, true correspondences are close to diagonal linking of the two anchors. We construct the initial ASM using such a function that pairs sentences near the middle of the two anchors with as many as $O(^2\sqrt{\bar{L}})$ (L is the number of sentences existing between two anchors) sentences in the other text because the maximum deviation can be stochastically modeled as $O(^2\sqrt{\bar{L}})$ (Kay and Roscheisen 1993). The initial ASM has little effect on the alignment performance so long as it contains all correct sentence correspondences.

3.2.2 Constructing AM

This step constructs an AM when given an ASM and a bilingual dictionary. Let t_{high}, t_{low}, I_{high} and I_{low} be two thresholds for *t-score* and two thresholds for mutual information, respectively. Let *ANC* be the minimal number of corresponding words for a sentence pair to be judged as an anchor.

First, mutual information and *t-score* are computed for all word pairs appearing in a possible sentence correspondence in ASM. We use hereafter the word correspondences whose mutual information exceeds I_{low} and whose *t-score* exceeds t_{low}. For all possible sentence correspondences $Jsentence_i$ and $Esentence_j$ (any pair in ASM), the following operations are applied in order:

1. If the following three conditions hold, add 3 to the *i-j* element of AM. (1) $Jsentence_i$ and $Esentence_j$ contain a bilingual dictionary word correspondence (w_{jpn} and w_{eng}). (2) w_{eng} does not occur in any other English sentence that is a possible translation of $Jsentence_i$. (3) $Jsentence_i$ and $Esentence_j$ do not cross any sentence pair that has more than *ANC* word correspondences.
2. If the following three conditions hold, add 3 to the *i-j* element of AM. (1) $Jsentence_i$ and $Esentence_j$ contain a stochastic word correspondence ($_{jpn}$ and w_{eng}) that has mutual information I_{high} and whose *t-score* exceeds t_{high}. (2) w_{eng} does not occur in any other English sentence that is a possible translation of $Jsentence_i$. (3) $Jsentence_i$ and $Esentence_j$ do not cross any sentence pair that has more than *ANC* word correspondences.

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

3. If the following three conditions hold, add 1 to the *i-j* element of AM. (1) $Jsentence_i$ and $Esentence_j$ contain a stochastic word correspondence (w_{jpn} and w_{eng}) that has mutual information I_{low} and whose *t-score* exceeds t_{low}. (2) w_{eng} does not occur in any other English sentence that is a possible translation of $Jsentence_i$. (3) $Jsentence_i$ and $Esentence_j$ does not cross any sentence pair that has more than *ANC* word correspondences.

The first operation deals with word correspondences in the bilingual dictionary. The second operation deals with stochastic word correspondences which are highly confident and in many cases involve domain specific keywords. These word correspondences are given the value of 3. The third operation is introduced because the number of highly confident corresponding words are too small to align all sentences. Although word correspondences acquired by this step are sometimes false translations of each other, these words appear in corresponding sentences with high probability. Thus, they can play a crucial role mainly in the final iterations phase. They are given one point.

3.2.3 *Adjusting ASM*

This step adjusts ASM using the AM constructed by the above operations. The sentence pairs that have at least *ANC* word correspondences are determined to be new anchors. By using the new set of anchors, a new ASM is constructed using the same method as used for initial ASM construction.

Our algorithm implements a kind of relaxation by gradually reducing t_{low}, I_{low} and *ANC*, which enables us to find confident sentence correspondences first. As a result, our method is more robust than dynamic programming techniques against the shortage of word-correspondence knowledge.

In this section, we report the result of experiments on aligning sentences in bilingual texts and on statistically acquired word correspondences. The texts for the experiment varied in length and genres as summarized in Table 2. Texts 1 and 2 are editorials

Table 2. *Test texts*

No.	Text name	Jpn	Eng	1-1	1-2	2-1	3-1
1	Root out guns at all costs	26	28	24	2	0	0
2	Economy facing last hurdle	36	41	25	7	2	0
3	Pacific Asia in the Post-Cold-War World	134	124	114	0	10	0
4	Visualizing the Mind	225	214	186	6	15	1

taken from *Yomiuri Shinbun* and its English version *Daily Yomiuri*. This data was distributed electrically via a WWW server.[4] The first two texts clarify the system's performance on shorter texts. Text 3 is an essay on economics taken from a quarterly publication of The International House of Japan. Text 4 is a scientific survey on brain

[4] We would like to thank Yomiuri Shinbun Co. for permitting us to use the data.

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

Table 3. *Result of sentence alignment*

	Proposed		**Statistics**		**Dictionary**	
Text	PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL
1	96·4%	96·3%	65·0%	48·5%	89·3%	88·9%
2	95·3%	93·1%	61·3%	49·6%	87·2%	75·1%
3	96·5%	97·1%	87·3%	85·1%	86·3%	88·2%
4	91·6%	93·8%	82·2%	79·3%	74·3%	63·8%

science taken from *Scientific American* and its Japanese version *Nikkei Science*. **Jpn** and **Eng** in Table 2 represent the number of sentences in the Japanese and English texts, respectively. The remaining table entries show categories of matches by manual alignment and indicate the difficulty of the task.

We briefly report here the computation time of our method. Let us consider Text 4 as an example. After 15 seconds for full preprocessing, the first iteration took 25 seconds with $t_{low} = 1{\cdot}55$ and $I_{low} = 1{\cdot}8$. The rest of the algorithm took 20 seconds in all. This experiment was performed on a SPARC Station 20. From the result, we may safely say that our method can be applied to voluminous corpora.

4.1 Sentence alignment

Table 3 shows the performance on sentence alignments for the texts in Table 2. **Proposed**, **Statistics** and **Dictionary** represent the methods using both statistics and dictionary, only statistics and only dictionary, respectively. Both **Proposed** and **Dictionary** use a CD-ROM version of a Japanese–English dictionary containing 40,000 entries. **Statistics** repeats the iteration by using statistical corresponding words only. This is identical to Kay's method (Kay and Roscheisen 1993), except for the statistics used. **Dictionary** performs the first interaction of the algorithm by using corresponding words of the bilingual dictionary. This delineates the coverage of the dictionary. The parameter setting used for each method was the optimum as determined by empirical tests.

In Table 3, PRECISION delineates how many of the aligned pairs are correct and RECALL delineates how many of the manual alignments we included in systems output. Unlike conventional sentence-chunk based evaluations, our result is measured on the sentence-sentence basis. Let us consider a 3-1 matching (sentence chink). Although conventional evaluations can make at most one error from the chunk because they count errors base on the number of chunks. On the other hand, three errors may arise by our sentence-sentence based evaluation. Note that our evaluation is more strict than the conventional one, especially for difficult texts, because they contain more complex matches.

For Texts 1 and 2, both the proposed method and the dictionary method perform much better than the statistical method. This is obviously because statistics cannot capture word-correspondences in the case of short texts.

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

Table 4. *Statistically acquired keywords*

Japanese	English	Mutual Information
記録	recording	3.58
リアルタイム	real	3.51
ニューロン	neuron	3.51
フィルム	film	3.51
グルコース	glucose	3.51
増加	increase	3.51
磁	MEG	3.51
解像度	resolution	3.43
電気	electrical	3.43
グループ	group	3.39
電気	recording	3.39
記録	electrical	3.39
言う	generate	3.33
提供	provide	3.33
図	MEG	3.33
NMR	NMR	3.17
ファンクショナル	functional	3.17
機器	equipment	3.17
臓器	organ	3.15
注射	compound	3.10
水	water	3.10
標識	radioactive	3.10
PET	PET	3.10
解像度	spatial	3.10
そのようだ	such	3.10
代謝	metabolism	3.06
言う	verb	3.04
科学者	scientist	2.95
地図	mapping	2.92
大学	university	2.92
思考	thought	2.90
化合物	compound	2.82
標識	label	2.82
オートラジオ	radioactivity	2.77
視覚	visual	2.77
信号	signal	2.77
リアルタイム	time	2.69
オートラジオ	autoradiography	2.67
能力	ability	2.63
CT	CT	2.63
…	…	…
聴覚	auditory	2.15
心	mental	2.05
MRI	MRI	1.8

Text 3 is easy to align in terms of both the complexity of the alignment and the vocabularies used. All methods performed well on this text.

For Text 4, **Proposed** and **Statistics** perform much better than **Dictionary**. The reason for this is that Text 4 concerns brain science and the bilingual dictionaries of general use did not contain domain specific keywords. On the other hand, the

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

proposed and statistical methods well capture the keywords as described in the next section. Note here that **Proposed** performs better than **Statistics** in the case of longer texts, too. There is clearly a limitation in the amount of word correspondences that can be captured by statistics. The precision of the proposed method for Text 4 is not as high as for the other texts due to the following. First, Text 4 contains more technical terms than others because it is on brain science. Bilingual dictionary of general use cannot capture these technical terms. Second, Text 4 contains more multiple alignments (such as 1-2 and 2-1). Multiple alignments are harmful for alignment performance.

In summary, the performance of **Proposed** is better than either **Statistics** or **Dictionary** for all texts, regardless of text length and the domain.

4.2 Word correspondence

In this section, we will demonstrate how well the proposed method captured domain specific word correspondences by using Text 4 as an example. Table 4 shows the word correspondences that have high mutual information. These are typical keywords concerning the non-invasive approach to human brain analysis. For example, NMR, MEG, PET, CT, MRI and functional MRI are devices for measuring brain activity from outside the head. These technical terms are the subjects of the text and are essential for alignment. However, none of them have their own entry in the bilingual dictionary, which would strongly obstruct the dictionary method.

It is interesting to note that the correct Japanese translation of 'MEG' is '脳磁図'. The Japanese morphological analyser we used does not contain an entry for '脳磁図' and split it into a sequence of three characters '脳', '磁' and '図'. Our system skillfully combined '磁' and '図' with 'MEG', as a result of statistical acquisition. These word correspondences greatly improved the performance for Text 4. Thus, the statistical method well captures the domain specific keywords that are not included in general-use bilingual dictionaries. The dictionary method would yield false alignments if statistically acquired word correspondences were not used.

5 Graphical alignment environment

This section describes a graphical alignment BACCS[5] (Bilingual Aligned Corpus Construction System) that incorporates our alignment program. The experimental results in the previous section showed that our alignment method attained high accuracy, however, this method can not always achieve 100% correct alignment. Thus it is necessary for humans to modify and confirm the alignment results given by our alignment program to build an accurately aligned corpus. Such a corpus would be useful resources for extracting various kinds of expressions and their translations in corpus-based learning methods.

BACCS offers a user interface that enables the user to easily confirm and modify alignment results and register word correspondences created by the alignment program into a user dictionary. Figure 4 shows a window image of BACCS. The rear

[5] BACCS was implemented in Tcl/Tk on X window.

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

window shows alignment results, while the front window shows the word correspondence candidates.

The process of making a directly aligned corpus with BACCS is as follows:

1. Select the target bilingual texts.
2. Run the alignment program and display the alignment results.
3. Confirm and modify these results by mouse actions.
4. Choose the proper word correspondences from among the candidates created by the alignment program and register these into the user dictionary.[6]

BACCS not only finds sentence alignment but also extracts bilingual collocations from aligned sentences by using the word-level sorting method (Haruno *et al.* 1996). The annotated bilingual corpus is used in our translation aid system that integrates bilingual corpus, mono-lingual corpus and hand-compiled dictionaries.

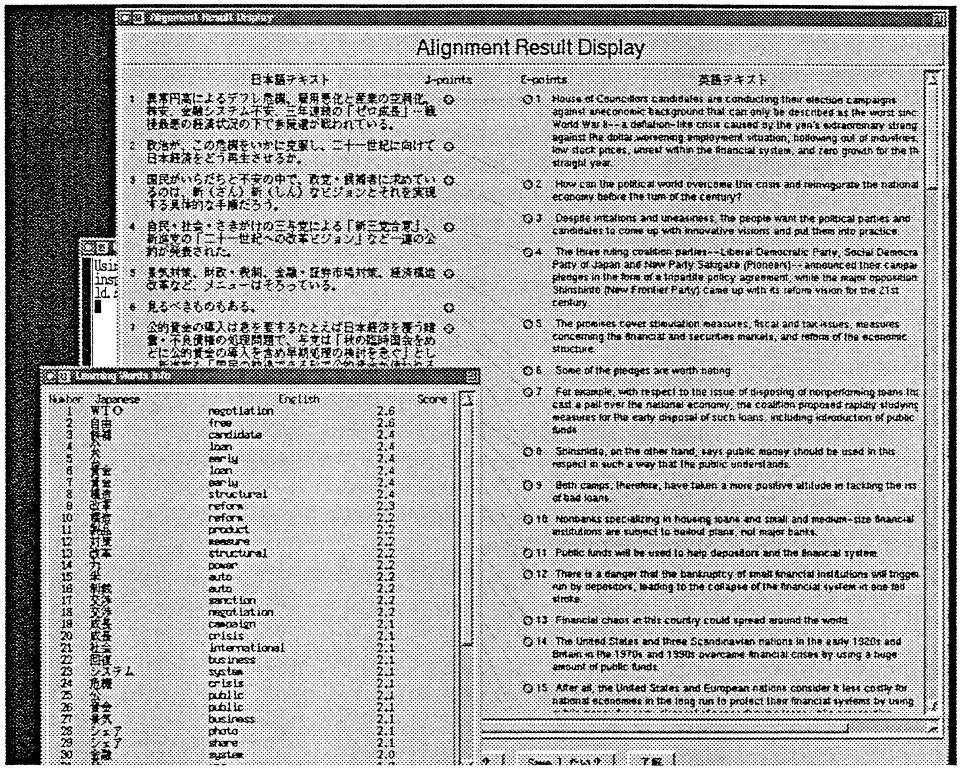


Fig. 4. BACCS.

6 Related work

Sentence alignment between Japanese and English was first explored by Sato and Murao (1991). They found (character or word) length-based approaches were not appropriate due to the structural difference of the two languages. They devised a

[6] In the experiments reported in the previous section, no user dictionaries were used.

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

dynamic programming method based on the number of corresponding words in a hand-crafted bilingual dictionary. Their approach is similar to a bilingual phrase extracting method (Debili and Sammouda 1992). Although some results were promising, the method's performance strongly depended on the domain of the texts and the dictionary entries. Utsuro *et al*. (1994) introduced a statistical post-processing step to tackle the problem. They first applied Sato's method and extracted statistical word correspondences from the result of the first path. Sato's method was then reiterated using both the acquired word correspondences and the hand-crafted dictionary. His method involves the same problem as Sato's method. That is, unless the hand-crafted dictionary contains domain specific key words, the first path yields false alignment, which in turn leads to false statistical correspondences. Because it is impossible in general to cover key words in all domains, it is inevitable that statistics and hand-crafted bilingual dictionaries must be used at the same time. The proposed method involves iterative alignment which simultaneously uses both statistics and a bilingual dictionary.

Fung and Church (1994) and Fung (1995) proposed methods to find Chinese–English word correspondences without aligning parallel texts. Their motivation is that structurally different languages such as Chinese–English and Japanese–English are difficult to align in general. Their methods gave up aligning sentences and only acquired word correspondences. Although their approaches are robust and do not require any information source, aligned sentences are indispensable for higher applications such as translation template acquisition (Matsumoto *et al*. 1993; Kaji *et al*., 1992) and example-based translation (Sato and Nagao 1990). Our method performs accurate alignment for such use by combining the detailed word correspondences: statistically acquired word correspondences and those from a bilingual dictionary of general use.

7 Conclusion

We have described a bilingual text alignment system for structurally different languages. The proposed method uses two kinds of word correspondences at the same time: word correspondences acquired by statistics and those of a bilingual dictionary. By combining these two types of word correspondences, the method covers both domain specific keywords not included in the dictionary and the infrequent words not detected by statistics. As a result, our method outperforms conventional methods for texts of different lengths and different domains.

References

Brill, E. (1992) A simple rule-based part of speech tagger. *Proc. Third Conference on Applied Natural Language Processing*. Pp. 152–155.

Brill, E. (1994) Some advances in transformation-based part of speech tagging. *Proc. 12th AAAI*. Pp. 722–727.

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

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

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press

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

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

Dagan, I. and Church, K. (1994) *Termight*: identifying and translating technical terminology. *Proc. Fourth Conference on Applied Natural Language Processing*. Pp. 34–40.

Debili, F. and Sammouda, E. (1992) Appariement des phrases de textes bilingues. *Proc 14th COLING*. Pp. 518–524.

Fung, P. and Church, K. W. (1994) K-vec: A new approach for aligning parallel texts. *Proc. 15th COLING*. Pp. 1096–1102.

Fung, P. (1995) A pattern matching method for finding noun and proper nouns translations from noisy parallel corpora. *Proc. 33rd ACL*. Pp. 236–243.

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

Haruno, M., Ikehara, S. and Yamazaki, T. (1996) Learning bilingual collocations by word-level sorting. *Proc. 16th COLING*. Pp. 525–530.

Kaji, H., Kida, Y. and Morimoto, Y. (1992) Learning translation templates from bilingual texts. *Proc. 14th COLING*. Pp. 672–678.

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

Kumano, A. and Hirakawa, H. (1994) Building an MT dictionary from parallel texts based on linguistic and statistical information. *Proc. 15th COLING*. Pp. 76–81.

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

Kurohashi, S., Nakamura, T., Matsumoto, Y. and Nagao, M. (1994) Improvements of Japanese morphological analyzer Juman. *Proc. 12th AAAI*. Pp. 722–727.

Matsumoto, Y., Ishimoti, H. and Utsuro, T. (1993) Structural matching of parallel texts. *31st Annual Meeting of ACL*. pp. 23–30.

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

Sato, S. and Nagao, M. (1990) Toward memory-based translation. *Proc. 13th COLING*. Pp. 247–252.

Sato, S. (1992) CTM: an example-based translation aid system. *Proc. 14th COLING*. Pp. 1259–1263.

Smadja, F. and McKeown, K. (1994) Translating collocations for use in bilingual lexicons. *ARPA Human Language Technology Workshop 94*. Pp. 152–156.

Utsuro, T., Ikeda Masaya Yamane, H., Matsumoto, Y. and Nagao, M. (1994) Bilingual text matching using bilingual dictionary and statistics. *Proc. 15th COLING*. Pp. 1076–1082.

Wu, D. (1994) Aligning a parallel English–Chinese corpus statistically with lexical criteria. *32nd Annual Meeting of ACL*. Pp. 80–87.

https://doi.org/10.1017/S1351324997001459 Published online by Cambridge University Press