Self-organizing semantic maps and its application to word alignment in Japanese-Chinese parallel corpora

Qing Ma[a*], Kyoko Kanzaki[b], Yujie Zhang[b], Masaki Murata[b], Hitoshi Isahara[b]

[a] *Department of Applied Mathematics and Informatics, Faculty of Science and Technology, Ryukoku University, Seta, Otsu 520-2194, Japan*

[b] *Keihanna Human Info-Communication Research Center, National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho, Kyoto 619-0289, Japan*

* *Corresponding author.* Tel.: +81-77-543-7505; fax: +81-77-543-7524

E-mail address: qma@math.ryukoku.ac.jp

Abstract

This paper presents a method involving self-organizing monolingual semantic maps that are visible and continuous representations where Chinese or Japanese words with similar meanings are placed at the same or neighboring points so that the distance between them represents the semantic similarity. We used the self-organizing map, SOM, as a self-organizing device. The words to be self-organized are defined by sets of co-occurring words collected from Chinese or Japanese newspapers, according to their grammatical relationships. The words are then coded into vectors to be forwarded to the SOM, taking into account the semantic correlation between them, which is established using a form of word-similarity computation. The self-organized monolingual semantic maps are assessed by numerical evaluations of accuracy, recall, and the F-measure, as well as by intuition, and by the comparisons with a clustering method and with multivariate statistical analysis. This paper further discusses the possibility that the method we propose can be extended to constructing Japanese-Chinese bilingual semantic maps, with the aim of providing a semantics-based approach to word alignment in Japanese-Chinese parallel corpora. We also show the effectiveness of this extended method through small-scale comparative experiments with a baseline method, where the alignment of Japanese and Chinese words is directly determined through the Euclidean distance of vectors representing the words, with a clustering method, and with multivariate statistical analysis.

Keywords: semantic map, word alignment, corpus, parallel corpus, Japanese, Chinese, monolingual, bilingual, SOM

---

1 Introduction

Computing word similarity in meanings is an important technique that can be applied not only to many natural language processing fields such as query expansion in information retrieval (Frakes and Baeza-Yates, 1992) and reasoning in word sense disambiguation (Dagan et al., 1993-1; Karov and Edelman, 1996; Lin, 1997), but also to studies on the lexical databases and/or thesauri (Hindle, 1990; Hatzivassiloglou and McKeown. 1993; Kanzaki et al., 2002; Kanzaki et al., 2003; Kanzaki et al., 2004).

A number of corpus-based statistical approaches have been used to compute word similarity (Hindle, 1990; Dagan et al., 1994; Mori and Nagao, 1998). In general, computation using these approaches is done as follows. First, words are represented by sets of their word co-occurrence statistics, relying on the assumption that the meaning of words is related to their patterns of occurrence with other words in the text (Harris, 1968). Second, all word representations are transformed into vectors. Finally, word similarity is computed using a mathematical measure to determine vector distance.

In practical applications, words must further be sorted globally based on prior computation of word similarities. This sorting is usually done using various clustering techniques. Word clustering, however, only classifies words into several groups. It is difficult to recognize the relationships between groups or the relationships between words within groups. To begin with, it may even be problematic to simply classify words into specific groups because some words can be classified into more than one group. To solve these problems, we need an alternative to word clustering to sort words. We need a technique that can map words from a very large lexicon into a small semantic space, i.e., a visible representation where words with similar meanings are placed at the same or neighboring points so that the distance between the points represents the semantic similarity in the words. This representation is called a semantic map.

Semantic maps can be automatically constructed with self-organization. To construct semantic maps, we need to use data that reflects semantic relationships between words. The data might be individual ones such as the co-occurring words appearing before and/or after the target words, the nouns or verbs in case structures, or the words in dependency relationships, according to the applications of semantic maps. Integrated data composed of the above individual information might also be used to create multipurpose semantic maps. The semantic maps created with individual information might be optimal for a specific purpose, but such data cannot be applied in other cases. The semantic maps created using general information, on the other hand, might not be optimal from the viewpoint of various specific purposes, but can be used for a number of different purposes. Therefore, for tasks where we know what kind of information is important, it would be better to construct specific semantic maps, and for tasks where it is unclear what information is important, it would be better to construct multipurpose semantic maps so that undesirable biases can be avoided. In this way, to construct semantic maps, the first priority is to select the type — i.e., specific maps or multipurpose maps — based on the analyses of tasks.

There have been several studies on constructing semantic maps for English (Ritter and Kohonen, 1989). In these, the self-organizing map, or SOM, proposed by Kohonen (1984) has been adopted as an unsupervised learning machine. In these self-organizing English maps, however, the data used for self-organization was prepared neither for specific purposes nor for multipurpose applications. The data was constructed from triplets of words which were simply gathered from three-word windows. As stated above, to construct practical semantic maps, we first have to determine which type of semantic map we want to construct, and then we need to use data that has a linguistic structure suitable for the selected map type. In addition, the window methods for gathering data adopted in English maps are too crude; the data may incorporate a great deal of noise (i.e., unrelated words) in the sets of co-occurring words, which would significantly degrade the quality of the maps created. The random coding adopted in these self-organizing English maps also left plenty of room for improvement. In random coding, all co-occurring words are first coded using n-dimensional, unit-norm random vectors (Press et al., 1994) and then they are composited into a single vector for coding the target word. Naturally, the larger the n is, the higher the possibility that all vectors for co-occurring word types (the different co-occurring words) will be independent of each other. In other words, the independence of the vectors for co-occurring word types cannot be guaranteed perfectly, because the n cannot be set to a value that will be high enough compared with the number of word types. A previous study (Ma et al., 2000) found that random coding cannot be applied to Japanese maps. What kind of co-occurring words should be used and how to effectively code them therefore remain as important issues in developing semantic maps.

This paper presents a method of self-organizing monolingual semantic maps for Chinese and Japanese using SOM for specific purpose; i.e., to construct semantic maps of nouns from the point of view of the adnominal constituents. The words to be self-organized are therefore limited to nouns and the co-occurring words used to define them are gathered together according to their grammatical relationships, i.e., adjective/noun–noun in Chinese and adjective/nominal adjective–noun in Japanese. These maps are helpful in our study of the semantic behaviors of adnominal constituents through an investigation of the relationships between the adnominal constituents and their modified nouns (Kanzaki et al., 2000-1, 2000-2, 2002, 2003, 2004). A vast quantity of co-occurring words were obtained from eleven years of "The People's Daily" for Chinese and eight years of the Mainichi Shinbun for Japanese. Instead of using random coding method, which Ma et al. (2000) found could not be applied to self-organizing Japanese maps, we coded the nouns defined by sets of co-occurring words, where the semantic correlation (word similarity) between words was taken into account beforehand. Further, term-weighting factors, such as the TFIDF (term frequency and inverse document frequency, a well-known term weighting method in the information retrieval field of NLP for selecting important keywords) and co-occurring frequency information, were adopted for coding so that co-occurring words could be weighted according to their importance to the nouns they modify. We assessed the self-organized monolingual semantic maps in three ways — through numerical evaluations of accuracy, recall, and the F-measure; through intuition (analyses based on concrete instances); and through comparison with clustering and multivariate statistical analysis.

In this paper, we also want to demonstrate the possibility that the proposed method can be extended to the construction of Japanese-Chinese bilingual semantic maps, with the aim of providing a semantics-based approach to word alignment in the parallel corpora of both languages. In this extended work, we used Kyoto University's Japanese corpus (Kurohashi and Nagao, 1997) and their translated Chinese corpus as the parallel corpus. Word alignment is an important and fundamental task in NLP and the research related to this includes a series of statistical models (e.g., Brown, et al., 1988; Brown, et al., 1993; Macklovitch and Hanna, 1996), a method involving dynamic programming (Dagan, 1993-2), a statistical approach introducing contextual information (Varea, et al., 2002), and methods of structural alignment (Kaji, et al., 1992; Matsumoto, et al., 1993; Wu, 1995; Imamura, 2001). All of these approaches, however, are either based on statistical information or grammatical structure, but not on meaning. If a bilingual semantic map could be automatically constructed by accepting translation pairs of sentences as inputs, word alignment could easily be obtained from the map. Since a bilingual semantic map, like a monolingual semantic map, would provide results that were visible and continuous, it would be easy to handle one-to-many or many-to-one correspondences. Furthermore, bilingual maps could perhaps be applied to foreign language learning or foreign language writing through the use of bilingual parallel corpora. The most important factor is that the translations should be free in most cases. Existing alignment methods that rely on statistical or grammatical information have evident limitations, and these suggest the necessity for developing an approach based on semantics. Through small-scale experiments, we also showed that the extended method was effective compared with a baseline method, where the alignment of Japanese and Chinese words was directly determined through the Euclidean distance of vectors representing the words, with a clustering method, and with multivariate statistical analysis.

The paper is organized as follows. Section 2 provides a brief introduction on

SOM. Section 3 describes the method for obtaining self-organizing monolingual semantic maps and presents the results of computer experiments. Section 4 describes the extension made to self-organizing bilingual semantic maps to align words and presents the results of computer experiments. Section 5 has the conclusion and points to the future directions this research will take.

2 Self-organizing map (SOM)

A SOM can be visualized as a two-dimensional array of nodes on which a high-dimensional input vector can be mapped in an orderly manner through a learning process. After the learning, a meaningful nonlinear coordinate system for different input features is created over the network. Such a learning process is competitive and unsupervised and is called a self-organizing process.

Suppose input $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T \in \Re^n$, where $\Re^n$ is an n-dimensional space. Each node i is then associated with a parametric reference vector $\mathbf{m_i}$, which equals $[m_{i1}, m_{i2}, \cdots, m_{in}]^T \in \Re^n$, whose element m_{ij} is a scalar weight between node i and input element x_j and is gradually modified during the learning process. When input vector $\mathbf{x}$ is given, it is compared to all reference vectors $\mathbf{m_i}$, which are associated with each node and is gradually modified during the learning process. Here, the network responses comply with two different stages, learning and mapping, as follows. Only a node whose reference vector has the smallest Euclidean distance to the input vector is activated during the mapping stages. This node, c, is called the best-matching node or *winner*. It can thus be defined by

$$c = \underset{i}{\operatorname{argmin}}\{||\mathbf{x} - \mathbf{m_i}||\}. \tag{1}$$

In the learning stage, on the other hand, not only the best-matching node but also its neighboring nodes are activated and their reference vectors are changed so that they are closer to the same input vector $\mathbf{x}$. This results in a local relaxation or smoothing effect on the reference vectors of the nodes in the neighborhood, which leads to global ordering during continued learning. This gradual adaptation of reference vectors can be expressed as

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)], \tag{2}$$

where $h_{ci}(t)$ is the neighborhood function. It is necessary that

$$\lim_{t \to \infty} h_{ci}(t) = 0 \tag{3}$$

for convergence. A widely applied neighborhood function can be written in terms of a Gaussian function:

$$h_{ci}(t) = \alpha(t) \cdot \exp(-\frac{||\mathbf{r}_c - \mathbf{r}_i||^2}{2\sigma^2(t)}). \tag{4}$$

Here, $\mathbf{r}_c \in \Re^2$ is the location vector of node c and $\mathbf{r}_i \in \Re^2$ is that of i. Term $||\mathbf{r}_c - \mathbf{r}_i||$ indicates that the farther node i is from node c, the smaller the h_{ci} and therefore the less the adaption of $m_i(t)$ will be. Term $\alpha(t)$ is the learning rate and $\sigma(t)$ defines the radius of the neighborhood. Both the latter terms are monotonically decreasing functions of time, and their exact forms are not critical. They can thus be defined linearly as

$$\alpha(t) = \alpha(0)\frac{T-t}{T}, \tag{5}$$

and

$$\sigma(t+1) = 1 + (\sigma(t) - 1)\frac{T-t}{T}, \tag{6}$$

where $\alpha(0)$ is an initial value and T is the total number of learning steps.

The learning process usually consists of an ordering phase and a fine adjustment phase. In the ordering phase, $\alpha(t)$ should start with a value that is close to unity, and the initial radius of the neighborhood can be more than half the diameter of the network. The terms $\alpha(t)$ and $\sigma(t+1)$ then decrease monotonically according to Eqs. (5) and (6). The ordering of m_i occurs during this initial phase, while the remaining steps are only needed to finely adjust the map. After the ordering phase, the radius may still contain the nearest neighbors of node c, and $\alpha(t)$ should attain a low value over a long period.

3 Self-organizing monolingual semantic maps

The monolingual semantic maps for self-organization are ones where Chinese or Japanese nouns are mapped in semantic order; i.e., nouns with similar meanings are mapped on (i.e., best-matched by) nodes that are topographically close to one another, and words with meanings that are dissimilar are mapped on nodes that are topographically far apart.

3.1 Learning data

To self-organize these kinds of semantic maps with a SOM, it is first necessary to have some kind of unsupervised learning data that reflects the semantic relations between nouns. Nouns usually need to be modified by something, which is called a modifier in linguistics, so that sentences will have a clearer meaning in their respective contexts. These modifiers are also called adnominal constituents. An adnominal constituent plus a noun make up a noun phrase, e.g., うれしい思い (happy thought), うれしい気持ち (happy feeling), and アカデミックな観点 (academic viewpoint) in Japanese and 重点企業 (key enterprises), 重点経済 (key economy), and 小小的心意 (small feeling) in Chinese. From these examples, we can see that both 思い (thought) and 気持ち (feeling) can be modified by *ureshii* (happy), but 観点 (viewpoint) cannot be modified by this in Japanese; likewise, both 企業 (enterprise) and 経済 (economy) can be modified by 重点 (key), but 心意 (feeling) cannot be modified by this in Chinese. That is, the semantic relationship between nouns in a sense depends on how many common adnominal constituents they have. Nouns are therefore defined as the sets of their co-occurring adnominal constituents, which are then coded into vectors forming the learning data to be passed to the SOM.

3.2 Data coding

Suppose there is a set of words w_i $(i = 1, \cdots, n)$ that we are planning to self-organize. Word w_i can be defined by a set of its co-occurring words as

$$w_i = \{a_1^{(i)}, a_2^{(i)}, \cdots, a_{\alpha_i}^{(i)}\}, \tag{7}$$

where $a_j^{(i)}$ is the jth co-occurring word of w_i and α_i is the number of co-occurring words of w_i.

Suppose we have correlative matrix D whose element d_{ij} is the word similarity [1] between words w_i and w_j. between words w_i and w_j, and word similarity is the reverse.). We can then code word w_i with the elements in the i-th row of correlative matrix D as

$$V(w_i) = [d_{i1}, d_{i2}, \cdots, d_{in}]^T. \tag{8}$$

The $V(w_i) \in \Re^n$ is the input to the SOM. Note that the individual d_{ij} of vector $V(w_i)$ only reflects the relationships between a pair of words when they are considered independently. To establish the relationships between word w_i and all other words, representations like vector $V(w_i)$ become necessary. Even if we have all such high-dimensional vectors for all words, it is still difficult to establish their global relationships. We therefore need to use SOM to reveal the semantic relationships in such high-dimensional vectors and represent them in two-dimensional space. In other words, the role of the SOM is merely to self-organize vectors and the quality of the maps created essentially depends on the

[1] Strictly speaking, element d_{ij} used here should be called the semantic distance between words w_i and w_j, and word similarity is the reverse.

vectors given. Therefore, the method by which word similarity d_{ij} is computed is a key to creating maps. A number of effective methods for computing word similarity have been proposed as follows.

3.2.1 Baseline method

Word similarity d_{ij} between words w_i and w_j with the baseline method is determined by

$$d_{ij} = \begin{cases} \frac{(\alpha_i - c_{ij}) + (\alpha_j - c_{ij})}{\alpha_i + \alpha_j - c_{ij}} & \text{if } i \neq j \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where α_i and α_j are, respectively, the numbers of co-occurring words of w_i and w_j, and c_{ij} is the number of co-occurring words that both w_i and w_j have in common. Word similarity d_{ij} is therefore the normalized distance between w_i and w_j in the context of the number of co-occurring words that they have in common; i.e., the smaller the d_{ij}, the closer w_i and w_j are in meaning.

3.2.2 Frequency term-weighting method

The frequency term-weighting method is based on the assumption that for a noun, the higher the co-occurrence frequency, the more important the co-occurring word is.

By regarding the frequency of each co-occurring word as an important weight to the headword (the target word), word similarity d_{ij} between word w_i and w_j can be measured by

$$d_{ij} = \begin{cases} \frac{(F_i - F_{ij}) + (F_j - F_{ij})}{F_i + F_j - F_{ij}} & \text{if } i \neq j \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

where F_i and F_j are respective frequency weighted values, expansions of the numbers (i.e., α_i and α_j) of co-occurring words of w_i and w_j, and F_{ij} is a frequency weighted value, an expansion of the number (i.e., c_{ij}) of common co-occurring words that both w_i and w_j have. They can be calculated as

$$F_i = \sum_{x=1}^{\alpha_i} f_x^{(i)} \quad \text{and} \quad F_{ij} = \sum_{x=1}^{c_{ij}} f_x^{(ij)}, \tag{11}$$

where $f_x^{(i)}$ and $f_x^{(ij)}$ are the co-occurring frequences of co-occurring word $a_x^{(i)}$ $(x = 1, \cdots, \alpha_i)$ of word w_i in Eq. (5) and the common co-occurring word $a_x^{(i)}$ that both words w_i and w_j $(x = 1, \cdots, c_{ij})$ have.

3.2.3 TFIDF term-weighting method

TFIDF calculation is a well-known term-weighting method (Sparck Jones, 1972), which has mainly been used to select important keywords in document classification and information retrieval (e.g., Robertson and Walker, 1994 and Murata et al., 2000). Weighting the importance of each co-occurring word with TFIDF is based on the assumption that only words that frequently co-occur with a particular headword, but rarely co-occur with other headwords, are really important. This is based on the idea that each headword is regarded as a document and its co-occurring words are regarded as keywords.

Word similarity d_{ij} between word w_i and w_j with TFIDF is measured by

$$d_{ij} = \begin{cases} \frac{(T_i - T_{ij}) + (T_j - T_{ij})}{T_i + T_j - T_{ij}} & \text{if } i \neq j \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

where T_i and T_j are the expansions of the numbers, α_i and α_j, of co-occurring words for w_i and w_j, respectively, and T_{ij} is an expansion of the number, c_{ij}, of co-occurring words that w_i and w_j have in common. These are calculated with

$$T_i = \sum_{x=1}^{\alpha_i} t_x^{(i)} \quad \text{and} \quad T_{ij} = \sum_{x=1}^{c_{ij}} t_x^{(ij)}, \tag{13}$$

where $t_x^{(i)}$ and $t_x^{(ij)}$ are the TFIDF values of co-occurring words $a_x^{(i)}$ $(x = 1, \cdots, \alpha_i)$ of w_i and co-occurring words $a_x^{(i)}$ that w_i and w_j $(x = 1, \cdots, c_{ij})$ have in common. Their respective values can be calculated as

$$t_x^{(i)} = tf(a_x^{(i)}, w_i) \cdot idf(a_x^{(i)}), \tag{14}$$

and

$$t_x^{(ij)} = tf(a_x^{(i)}, w_i, w_j) \cdot idf(a_x^{(i)}). \tag{15}$$

Here, $tf(a_x^{(i)}, w_i)$ is the co-occurrence frequency of co-occurring word $a_x^{(i)}$ and word w_i, $tf(a_x^{(i)}, w_i, w_j)$ is the co-occurrence frequency of $a_x^{(i)}$, w_i, and w_j, and $idf(a_x^{(i)})$ is the inverse frequency with which $a_x^{(i)}$ appears in all headwords:

$$idf(a_x^{(i)}) = \log \frac{n}{df(a_x^{(i)})} + 1, \tag{16}$$

where, n is the total number of headwords and $df(a_x^{(i)})$ is the number of headwords co-occurring with $a_x^{(i)}$.

TFIDF value $t_x^{(i)}$ (including $t_x^{(ij)}$) is therefore a weight reflecting the importance of co-occurring word a_x^i for word w_i. If we consider that all co-occurring words have the same importance to each headword, then Eq. (12) is the same as Eq. (9).

Note that this method is not only theoretically effective, but also provided the best experimental result as shown in Table 1 in the next section.

3.3 Experimental Results

3.3.1 Data

To evaluate the experimental results in Chinese more easily and objectively, we manually selected headwords (a total of 85 nouns) from six categories in "The Contemporary Chinese Classified Dictionary" (Dong et al., 1998) using the criterion that they could be classified easily by ourselves (i.e., we knew their exact meaning, or there was little ambiguity) so that they could be numerically evaluated. We also added several Chinese family names as a new category, which were not in the dictionary but appear frequently in newspapers. The co-occurring words were adjectives and nouns that together with headwords formed noun phrases and these were collected by computer from eleven years of "The People's Daily." There were a total of 69,030 co-occurring word tokens and 22,118 co-occurring word types.

We used noun phrases composed of nouns and adjectives/nominal adjectives in Japanese, collected by computer from eight years of the Mainichi Shinbun newspaper in order of the frequency of co-occurring adjectives/nominal adjectives. There were 100 nouns, a total of 33,870 co-occurring word tokens, and 4,023 co-occurring word types[2].

3.3.2 SOM

We used a SOM of a 13×13 two-dimensional array. The number of dimensions of input, n, was 85 in Chinese and 100 in Japanese. In the ordering phase, the number of learning steps T was set at 10,000, the initial value of the learning rate $\alpha(0)$ was set at 0.1, and the initial radius of the neighborhood $\sigma(0)$ was set at 13, a value equal to the diameter of the SOM. In the fine adjustment phase, the T was set at 100,000, $\alpha(0)$ was set at 0.01, and the $\sigma(0)$ was set at 7. The initial reference vectors $m_i(0)$ consisted of random values between 0 and 1.0.

[2] The main reason for the huge discrepancy between tokens and types for Japanese words as compared to Chinese words is that both the Japanese adjectives and nominal adjectives have many transformations for tense[e.g., 美しい (is beautiful) → 美しかった (was beautiful), きれいな (is clear) → きれいだった (was clear)].

3.3.3 Evaluation methods

(a) Numerical evaluation: Because numerical evaluation is always the most objective, we worked out precision and recall, which are defined as follows, as numerical measures.

$$P = \frac{\sum_{i=1}^{C} p_i}{C}, \qquad R = \frac{\sum_{i=1}^{C} r_i}{C}, \tag{17}$$

where C is the total number of classes and p_i is the precision and r_i is the recall for one class i. These are defined by

$$p_i = \frac{\#\text{ words correctly classified as class } i}{\#\text{ words of class } i \text{ in result}}, \tag{18}$$

$$r_i = \frac{\#\text{ words correctly classified as class } i}{\#\text{ words of class } i}, \tag{19}$$

where # means number.

We also used the F-measure to obtain a general score for precision and recall:

$$\text{F-measure} = \frac{2P \times R}{P + R} \tag{20}$$

(b) Intuitive evaluation: Because the most notable features of maps are their visibility and the continuity of classifications, it is insufficient only to evaluate classifications numerically. It is therefore important to judge whether the maps are intuitively meaningful; i.e., whether the nouns were mapped correctly according to our "common sense" through concrete analyses of instances.

(c) Comparison with other methods: To assess the effectiveness of our method in semantic classification, we compared it with a conventional clustering technique. In addition, to assess its usefulness in developing semantic maps, we investigated whether it was feasible to use multivariate statistical analyses such as principal component analysis to construct visible representations of semantic classification.

3.3.4 Results

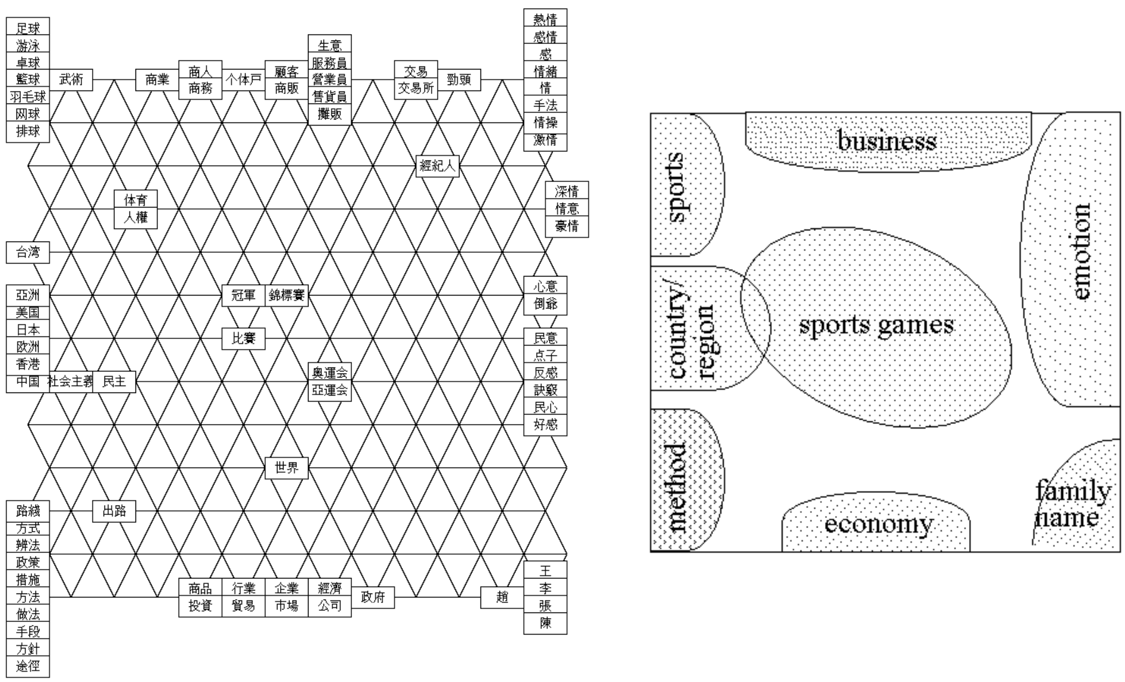
Because the Chinese headwords were manually selected from a semantic dictionary, and we knew their exact semantic categories, we were able to numerically evaluate the Chinese maps and classifications obtained using the hierarchical clustering technique. Table 1 lists the results of numerical evaluation ranked in order of the F-measure. The table shows that the F-measures for the classifications in maps based on both baseline coding and TFIDF term-weighted coding were higher than those for clustering results produced by these two coding methods. It also shows that the F-measure for the map based on TFIDF term-weighted coding was higher than that for the map based on frequency term-weighting, which itself was slightly higher than that for the map based on baseline coding.

Figure 1 (a) has a semantic map of Chinese nouns that have been self-organized with TFIDF term-weighted coding, and (b) shows that the map can be divided into eight groups according to their meanings, so that nouns in the same group have similar meanings. Of the total of 85 nouns, only six nouns, 手法 (method), 勁頭 (spirit), 倒爺 (broker), 点子 (idea), 訣竅 (knack), and 政府 (government) were mapped in incorrect areas in the sense that not only were they different from the definition in the dictionary, but also they were intuitively inconsistent. However, even in these nouns, 勁頭 (spirit) was mapped near the correct area *emotion*. In addition, although the noun 世界 (world) in the area of *sports games* was mapped incorrectly in the sense that it differed from the definition in the dictionary, its location was intuitively reasonable. The remaining 78 nouns, which were originally distributed into seven semantic categories in the Chinese dictionary, were therefore correctly divided into eight semantic categories in the semantic map. The category *politics* was merged with category *method* and the category *sports* was split into two groups: for actual *sports* and for *sports games*. The category *business* was split into two groups: one for *business* and another for *economy*. These classifications clearly do not contradict the original in essence (note that the precision and recall in Table 1 were calculated taking the categories *sports* and *sports game* as a whole, and *method* and *policy* as a whole.). That is, the self-organized map is basically consistent with the definitions found in the Chinese dictionary. Naturally, it is also intuitively consistent in most cases.

Table 2 lists the classification results obtained through the hierarchical clustering with TFIDF coding, where the 85 nouns were divided into eight categories. As with the semantic map, these eight categories were obtained manually from the hierarchical clustering results. Also as with the semantic map, the category *politics* was also merged into the category *method*. Moreover, the category *business* was split into two groups: one for *business* and another for *economy*. The category *sports* was also divided into two groups: one for *sports* and another for *sports game*. If we compare these classification results to the map obtained with TFIDF coding, we find that both results are very similar and the map is slightly better than that obtained with clustering, noting that there are eight nouns, 政府 (government), 社会主義 (socialism), 民主 (democracy), 人権 (human rights), 体育 (sports), 手法 (technique), 点子 (idea), and 訣竅 (knack), classified into incorrect areas (underlined in the table) in the sense that they are not only different from the definitions in the dictionary, but also inconsistent with our intuition. This agrees with the numerical evaluations discussed above.

Table 1
Comparative results for various coding methods and clustering.

	Precision	Recall	F-measure
Clustering with baseline coding	0.936	0.864	0.899
Baseline	0.926	0.90	0.913
Frequency	0.928	0.90	0.914
Clustering with TFIDF term-weighted coding	0.95	0.896	0.922
TFIDF	0.944	0.907	0.925



(a) (b)

Fig. 1. Chinese semantic map based on TFIDF term-weighted coding.

Principal component analysis of the same Chinese data with TFIDF term-weighted coding showed that the cumulative coefficients of determination for the top two and ten principal components were 8.29% and 24.53%, respectively. In general, if the value is not larger than 80%, the multivariate data cannot be compressed into a small number of principal components. Therefore, it is difficult to construct good semantic maps with multivariate statistical analysis. We conducted an experiment to plot the data using the first and second principal components. The results are shown in Figure 2. Clearly, as almost all nouns were concentrated in some area to the right this method could not be used to create a meaningful map.

It was not possible to numerically evaluate the Japanese maps because all the

Table 2
Clustering results with TFIDF term-weighted coding.

Class	**Noun**	**Corresponding to the map**
1	李 張 王 陳 趙	*person's name*
2	籃球 排球 网球 卓球 羽毛球 游泳 足球 武術	*sports*
3	方針 路線 途経 出路 政策 措施 辨法 方法 方式 手段 做法	*method*
4	公司 市場 企業 行業 経済 <u>政府</u> <u>社会主義</u> <u>民主人権</u> <u>体育</u> 商業 商品 投資 貿易	*economy*
5	比賽 冠軍 錦標賽 奥運会 亜運会	*sports game*
6	美国 日本 中国 世界 欧州 亜州 台湾 香港	*country/region*
7	民意 民心 反感 好感 感 <u>手法</u> 感情 情感 激情 熱情 情緒 情操 深情 情意心意 情 豪情 勁頭	*emotion*
8	營業員 售貨員 商販 服務員 攤販 个体戸 顧客 生意 商人 交易所 經紀人 交易 <u>点子</u> <u>訣竅</u> 倒爺 商務	*business*

(The underlined words are those classified into incorrect areas.)

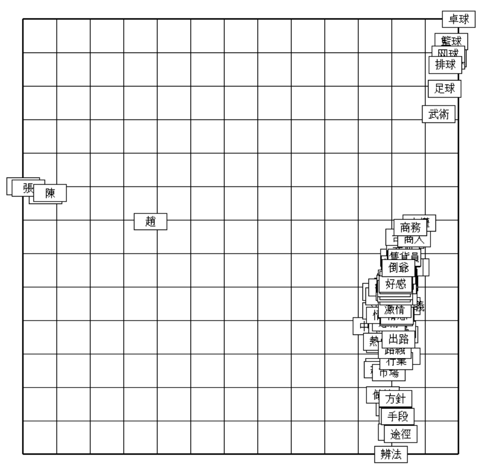
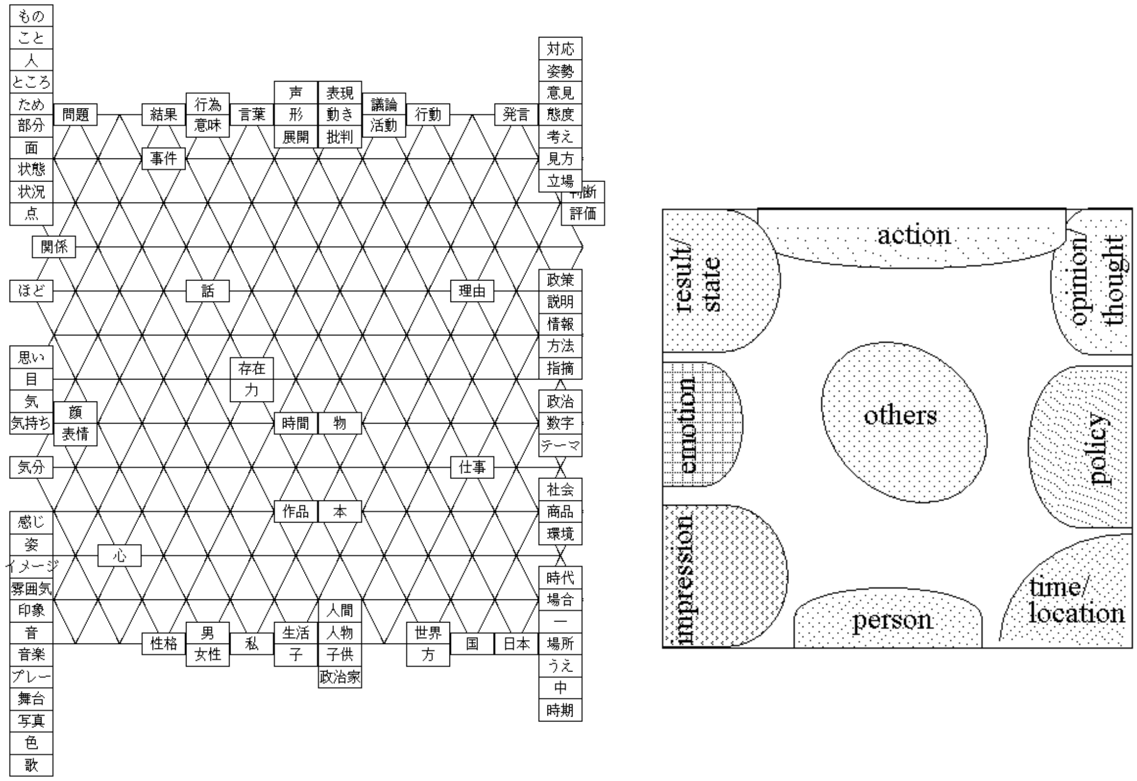


Fig. 2. Chinese semantic map using principal component analysis.

Japanese nouns were automatically collected from newspapers and we could not know their exact meaning. We could, however, assess that the Japanese self-organized semantic map (Figure 3) was generally intuitively consistent and the classification was not inferior to that obtained by hierarchical clustering. In addition, principal component analysis on the same data revealed that the cumulative coefficient of determination for the top two and ten principal components was 7.317% and 22.679%, respectively. The results with the first and second principal components revealed that all words were merged together



(a) (b)

Fig. 3. Japanese semantic map based on the TFIDF term-weighted coding method.

and this method could not be used to create a meaningful Japanese map (Figure 4).

The main reason that principal component analysis could not be applied to this task was the strong nonlinearity of the data. That is, while the SOM is capable of nonlinear processing, principal component analysis can only perform linear processing because the principal components are obtained through a linear combination of original variables: thus, principal component analysis cannot be applied to data with strong nonlinearity.

3.4 Summary

Section 3 described a method for self-organizing monolingual semantic maps. The results of computer experiments indicated that meaningful Chinese and Japanese maps can be created and adapting TFIDF or frequency term-weighting is effective for achieving this. Comparisons with clustering and multivariate statistical analysis revealed that the proposed method produces appropriate classifications and is suitable for creating visible representations of classification results.

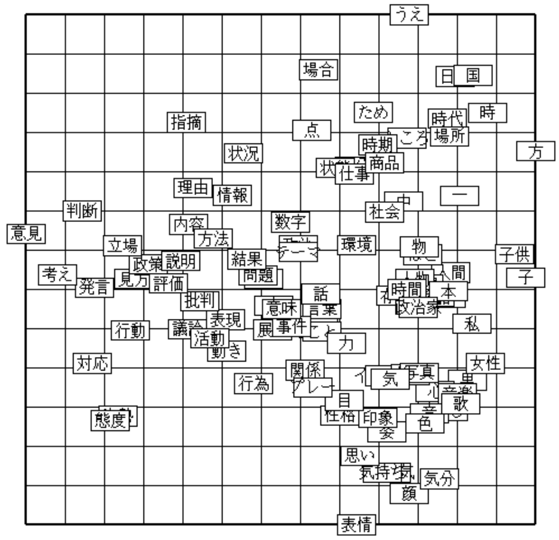


Fig. 4. Japanese semantic map using principal component analysis.

4 Self-organizing bilingual semantic maps

When a translation pair of sentences like

(Japanese) 経営 トップ が 低 成長 時代 定着 を 実感 して いる こと を うかがわ せた 。

(Chinese) 由此 可以 看出 ， 最高 経営者 深感 経済 仍 停留 在 低速 増長 時代 。

(English translation: We can see that upper management has realized that the economy is fixed in an eras of slow growth.)

is given, the bilingual semantic map for self-organization is one where all Japanese and Chinese words appearing in the two sentences are mapped in semantic order. Each Japanese word can therefore be automatically aligned to a Chinese word from this map by measuring its distance: if the Chinese word 可以 (can) is closest to the Japanese word せた (can), for example, then the Japanese word せた (can) is regarded as being aligned to the Chinese word 可以 (can).

As a part of the Japanese-Chinese machine translation project, we are constructing a parallel bilingual corpus based on the Kyoto University Japanese corpus (Kurohashi and Nagao, 1997). The translation pairs of sentences used in this paper were therefore obtained from the corpus. As the corpus is still being manually constructed by linguistics experts, only a very small part was available when this work was done. Since this corpus has already been morphologically analyzed, the Japanese sentences were used directly, while the translated Chinese sentences were segmented and part-of-speech tagged with a morphological analysis tool developed by Peking University (Zhou and Duan, 1994).

The same as with the case for the monolingual semantic map, it is also necessary to have learning data that can reflect the semantic relations between Japanese and Chinese words to self-organize bilingual semantic maps when translation pairs of sentences are given. To do so, we first have to deal with the problem of evaluating two different languages with the same gauge, or self-organizing words of different languages in the same map. One way that this can easily be achieved is to unify them into one language through a bilingual dictionary. In this paper, the words appearing in a translated Chinese sentence were given up to five translated Japanese candidates, and these were used instead of the original Chinese words. The candidates were obtained manually [3] from two Chinese-Japanese dictionaries: "Han Ri Ci Dian", published by Jilin Education Publisher, and the "Chunichi Daijiten", published by Taishukan Publishing Co., Ltd. The later dictionary was only used when a word had no entry in the former dictionary. The candidates were selected according to the following order of priority: (i) a word that is also in the Japanese original sentence; (ii) a word with the same part of speech (POS) as the original Chinese word; (iii) a word chosen based on the order listed in the dictionary; and (iv) a word appearing in the Kyoto University corpus. Thus, all the words in the translated Chinese sentence in the pair given above can be rendered in terms of Japanese candidates as follows:

(Chinese) 由此:これによって 可以:ことができる/てよい 看出:見抜く/看破 ，:， 最高:最高/最も高い 経営者:経営者 深感:実感 経済:経済/生活/経済的 仍:依然として/いまなお 停留:滞在/止まる 在:で/に/している/しつつある 低速:低 増長:増長/ふえる 時代:期/時代 。:。

In this way, we can express a translation pair of sentences only in terms of Japanese words. As this example demonstrates, we can recognize translated

[3] If there are Chinese-Japanese electronic dictionaries avaiable in computer, then we can obtain the data automatically. However, we have no such dictionaries at present.

Japanese candidates, such as “これによって (from this)” or “ことができる/てよい (can/may)”, that do not exist in the original Japanese sentence. This means that it is virtually impossible to align words by only using surface representations, even if the translation pair of sentences has been unified by a single language.

The actual learning data used in self-organization was obtained as follows. Each Japanese word appearing in a Japanese sentence was defined in terms of its co-occurring words (the target word itself and the words to its immediate left and right). They were obtained from eight years (1991-1998) of the Japanese newspaper, the Mainichi Shinbun, and used as learning data. Each Chinese word appearing in a translated Chinese sentence was defined in terms of the co-occurring words of its Japanese translation candidates and the Chinese words defined in this way were used as learning data. In the next section, we provide a detailed description of how the learning data was constructed and of the coding method used to transform it into inputs for the SOM.

4.2 Data coding

Suppose we are given a Japanese-Chinese translation pair of sentences:

$$J_1, J_2, \cdots, J_m$$

$$C_1 : J_{11}/\cdots/J_{1,n_1}, \cdots, C_n : J_{n1}/\cdots/J_{n,n_n}$$

, where J_i $(i = 1, \cdots, m)$ are Japanese words forming the Japanese sentence, C_i $(i = 1, \cdots, n)$ are Chinese words forming the translated Chinese sentence, J_{ij} $(i = 1, \cdots, n,\ j = 1, \cdots, n_i)$ is the jth translated Japanese candidate for C_i, $n_i (1 \leq n_i \leq t)$ is the number of candidates for C_i, and t is the maximum number of candidates ($t = 5$ in this paper).

Word w_i $(= J_i)$ of a Japanese sentence is defined by a set of co-occurring information:

$$w_i = J_i = \{a_1^{(i)}, f_1^{(i)}, \cdots, a_{\alpha_i}^{(i)}, f_{\alpha_i}^{(i)}\}, \tag{21}$$

where $a_j^{(i)}$ is a co-occurring word of J_i, $f_j^{(i)}$ is the normalized (i.e., $\sum_{j=1}^{\alpha_i} f_j^{(i)} = 1$) co-occurrence frequency, and α_i is the number of words co-occurring with J_i. Word w_j $(= C_j)$ of a translated Chinese sentence is also defined by a set of co-occurring information:

$$w_j = C_j = \{J_{j1}, \cdots, J_{j,n_j}\} = \{a_1^{(j)}, f_1^{(j)}, \cdots, a_{\alpha_j}^{(j)}, f_{\alpha_j}^{(j)}\}, \tag{22}$$

where $a_i^{(j)}$ is a co-occurring word of either or severals of $J_{j1}, \cdots, J_{j,n_i}$, $f_i^{(j)}$ is the normalized co-occurrence frequency (sum of frequencies when occurring with severals), and α_i is the number of words co-occuring with J_i.

Since the Chinese words are also defined in terms of co-occurring Japanese words, there is no need to distinguish between them, and it thus becomes possible to apply all existing coding methods to self-organizing monolingual semantic maps. Here, the semantic distance d_{ij} between any two words w_i and w_j appearing in a translation pair of sentences is calculated by frequency term-weighting as we can see from Eq. (10). Note that we also tried to use TFIDF term-weighting, but the map obtained with this method was worse than that obtained with frequency term-weighting. The reasons for this will be studied in our future work.

4.3 Experimental Results

4.3.1 Data

Word-alignment experiments were conducted for ten translation pairs of sentences. The learning data was obtained in the manner described in Section 4.1. If we considering the translation pair of sentences given at the begining of Section 4 as an example, there were $N = m + n$=16+15=31 words, a total of 62,627 co-occurring word tokens, and 22,077 co-occurring word types. Of the 31 words, the Japanese period symbol ("。")[4] had the largest number of co-occurring words (4,180), while the word "うかがわ" (see) in the Japanese sentence and the comma "，" in the translated Chinese sentence had the smallest numbers of co-occurring words (5 each).

4.3.2 SOM

Except for the number of input dimensions which should be the same as the number of words to be mapped, all other SOM parameters were entirely the same as those used in the self-organizing monolingual semantic maps.

4.4 Results

Figure 5 shows the map for the translation pair given in at the begining of Section 4 as an example. Here, the words tagged "J" are Japanese words from the Japanese sentence and the words tagged "C" are Chinese words from the translated Chinese sentence. We could obtain the word-alignment results listed

[4] Although there is actually no need to align Japanese period symbols between sentences, this step was not omitted because the sentences were processed mechanically.

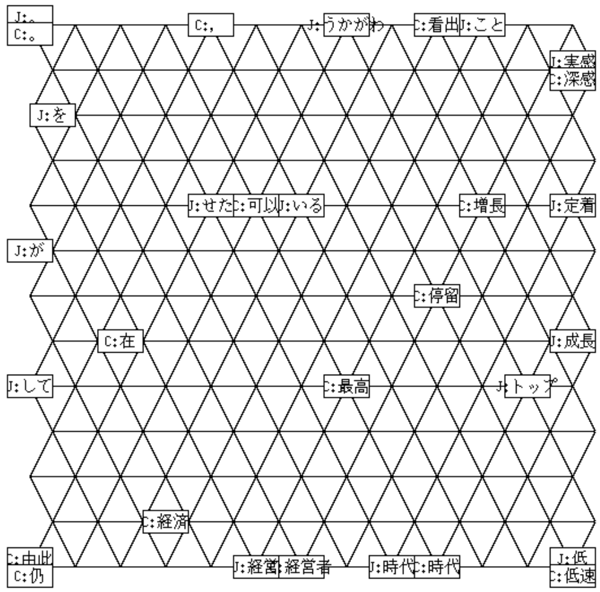


Fig. 5. Bilingual semantic map obtained through self-organization.

in Table 3 from the map by focusing on each Japanese word and choosing the closest Chinese word to it. The correct answers are also given in the table where we can see that [J： 低, C：低速] (low speed),[J：時代, C：時代] (age), [J：実感, C：深 感] (realize), [J： うかがわ, C：看 出] (see), [J：せた, C：可以] (can), and [J:。, C:。] were aligned correctly. The other alignment results are incorrect in the strict sense of the word. If we consider the second Chinese candidate, however, we can see that [J:トップ, C:最高] (top), and [J:成長, C:増長] (growth) were aligned correctly. What was also interesting was that although the Japanese word "J:定着" (fixed) was incorrectly aligned with "C:増長" (growth) as the first candidate, they were somewhat similar in the sense that both were words of tendency. Of the incorrect alignment results, [J:こと (thing)，C:看出 (see)] and [J:を，C:。] were due to the fact that there were no Chinese words (or at least none appearing in the sentence) that corresponded to these Japanese words. Another problem was incorrect alignment caused by the inconsistency of word segmentation between Japanese and Chinese sentences, as for [J:経営 (management)，C:経営者 (manager)]. None of these problems can be resolved by only applying the word alignment technique.

Table 4 lists the baseline word alignment results that were obtained by focusing on each Japanese word and choosing the Chinese word with the smallest semantic distance d_{ij} calculated with Eq. (10). From this table we can see that [J：うかがわ (see), C：深感 (realize)] was incorrect, while "J：うかがわ" (see) was correctly aligned by using the semantic map. Although incorrect results were obtained both for the semantic map, such as [J:成長 (growth), C:停留 (stop)] or [J:定着 (fixed), C:増長 (growth)] and for the baseline such as, [J:成長 (growth), C:時代 (age)] or [J:定着 (fixed), C:深感 (realize)], the results for the semantic map were somewhat correct in meaning, whereas those for the

Table 3
Word alignment results obtained from semantic map

Japanese	1st Chinese candidate	2nd Chinese candidate	Correct answer
J:経営:	C:経営者	C:経済	-
J:トップ	C:停留	C:最高	C:最高
J:が	C:在	C:。	-
J:低	C:低速	C:時代	C:低速
J:成長	C:停留	C:増長	C:増長
J:時代	C:時代	C:経営者	C:時代
J:定着	C:増長	C:深感	C:停留
J:を	C:。	C:,	-
J:実感	C:深感	C:看出	C:深感
J:して	C:在	C:仍	-
J:いる	C:可以	C:停留	-
J:こと	C:看出	C:深感	-
J:を	C:。	C:,	-
J:うかがわ	C:看出	C:,	C:看出
J:せた	C:可以	C:在	C:可以
J:。	C:。	C:,	C:。

baseline were totally wrong. If we check the second candidates, we find that the second candidates for "J:成長" (growth) and "J: トップ" (top) were "C:深感" (realize) and "C:時代" (age) which are incorrect, while they were aligned correctly with the second candidates through the semantic map. We can thus say that the method using semantic map performed better than the baseline method.

Figure 6 shows the word-alignment semantic map obtained by principal component analysis (PCA). By comparing it with Figure 5, we can see that the results obtained through PCA would be worse than those obtained through self-organization. For example, the pair [J：うかがわ，C：看出] could not be obtained through PCA. "J：成長" also could not be correctly aligned even if the second closest candidate was included. In addition, words tend to cluster together in certain areas and the total disposition of the words is thus imbalanced, which detracts from the semantic map's features of visibility and continuity. We also tried to use hierarchical clustering to align word. The results obtained were slightly worse than those obtained with the self-organizing semantic map. For example, [J：うかがわ，C：看出] could also not be correctly

Table 4
Baseline word alignment results

Japanese	1st Chinese candidate	2nd Chinese candidate	Correct answer
J:経営:	C:経営者	C:経済	-
J:トップ	C:経営者	C:時代	C:最高
J:が	C:在	C:。	-
J:低	C:低速	C:最高	C:低速
J:成長	C:時代	C:深感	C:増長
J:時代	C:時代	C:経営者	C:時代
J:定着	C:深感	C:増長	C:停留
J:を	C:在	C:。	-
J:実感	C:深感	C:増長	C:深感
J:して	C:在	C:看出	-
J:いる	C:可以	C:在	-
J:こと	C:深感	C:増長	-
J:を	C:在	C:。	-
J:うかがわ	C:深感	C:看出	C:看出
J:せた	C:可以	C:在	C:可以
J:。	C:。	C:在	C:。

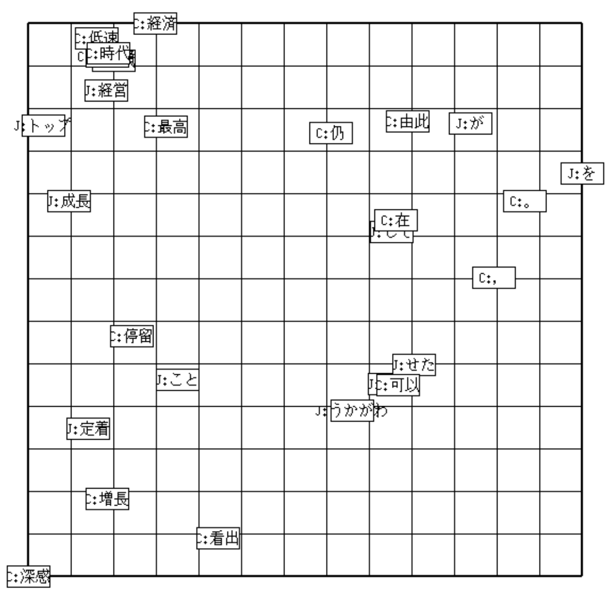


Fig. 6. Semantic map by PCA.

obtained with clustering. Moreover, because we could not know the semantic distance between words within a group, we could not as easily obtain second closest candidates as we could with the semantic map.

Here, note that although we only provided the results for one translation pair because of limited space, we obtained similar results for the other nine translation pairs.

4.5 Summary

Section 4 described a method of self-organizing bilingual semantic maps for word alignment. Its effectiveness was confirmed through a small-scale (ten translation pairs) experimental comparison with the baseline method where the alignment of Japanese and Chinese words was determined directly by the Euclidean distance of vectors representing the words, and by a comparison with hierarchical clustering and multivariate statistical analysis.

5 Conclusion

This paper proposed a method of self-organizing monolingual semantic maps for Japanese and Chinese. Computer experimental results proved that these maps were generally consistent with our intuition. Our comparison demonstrated that the hierarchical clustering technique is inferior to SOM in terms of classifying ability. Furthermore, it has been clarified that multivariate statistical analysis such as principal component analysis and factor analysis gave worse results and therefore cannot be used to create meaningful maps instead of SOM, which re-confirmed the necessity of using our method for this task.

The paper also discussed the possibility of an extension to the automatic construction of bilingual semantic maps of Japanese and Chinese, with the aim of providing a semantics-based approach to word alignment in the parallel corpora of these languages. We showed the effectiveness of word alignment through bilingual semantic maps in small-scale experiments by comparing it with a baseline method where the alignment of Japanese and Chinese words was directly determined through the Euclidean distance of vectors representing the words, and by comparing it with conventional clustering technique and multivariate statistical analysis.

In our future work, we first plan to develop an automatic method of transforming both Japanese and Chinese words in a given translation pair into representations with no distinction between languages after the entire Japanese-

Chinese parallel corpus is made available. We then plan to conduct large-scale word-alignment experiments on all translated pairs appearing in the parallel corpus, to prove the effectiveness of using semantic maps. We finally plan to develop a practical, high-performance, semantics-based word-alignment system by integrating various existing methods into a new version of our system.

References

[1] Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Mercer, R. L., Roossin, P. (1988). A statistical approach to language translation, *COLING'88*, 71-76.

[2] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation, *Computational Linguistics*, 19(2), 263-311.

[3] Brown, R. D. (1997). Automated dictionary example-based translation. *Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation*, 111-118.

[4] Dagan, I., Marcus, S., and Markovitch, S. (1993-1). Contextual word similarity and estimation from sparse data. *ACL'93*, Columbus, Ohio, 164-171.

[5] Dagan, I., Church, K. W., Gale, W. A. (1993-2). Robust bilingual word alignment for mach ine aided translation. *Proceedings of the Workshop on Very Large Corpora*, 1-8.

[6] Dagan, I., Pereira, L., and Lee, L. (1994). Similarity-based estimation of word cooccurrence probabilities. *ACL'94*, Las Cruces, NM, 272-278.

[7] Dong, D. N., et al. (Eds.) (1998). Contemporary Chinese classified dictionary. *Han-Yu-Da-Ci-Dian Press.*

[8] Frakes, William B. and Baeza-Yates, R. (Eds.) (1992). *Information retrieval: data structures & algorithms.* New Jersey, Prentice-Hall.

[9] Harris, Z. S. (1968). Mathematical structures of language, New York: Wiley.

[10] Hatzivassiloglou, V. and McKeown, R.K. (1993). Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. *ACL'93*, 172-182.

[11] Hindle, D. (1990). Noun classification from predicate argument structures. *ACL'90*, Pittsburgh, PA, 268-275.

[12] Imamura, K. (2001). Hierarchical phrase alignment harmonized with parsing. *NLPRS2001*, 377-384.

[13] Kaji, H., Kida, Y., Morimoto Y. (1992). Learning translation templates from bilingual text, *COLING'92*, 672-678.

[14] Karov, Y. and Edelman, S. (1996). Learning similarity-based word sense disambiguation from sparse data. *Proc. the Fourth Workshop on Very Large Corpora*, Copenhagen, pp. 42-55.

[15] Kanzaki, K., Ma, Q., Isahara, H. (2000-1). Similarities and differences among semantic behaviors of Japanese adnominal constituents, ANLP/NAACL2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems, 59-68.

[16] Kanzaki, K., Ma, Q., Isahara, H. (2000-2). Distribution of nouns in a semantic map based on the semantic relations between adnominal constituents and their head nouns, Second International Conference on the Mental Lexicon, October.

[17] Kanzaki, K., Ma, Q., Murata, M., and Isahara, H. (2002). Classification of adjectival and non-adjectival nouns based on their semantic behavior by using a self-organizing semantic map, *COLING'2002 Workshop on Building and Using Semantic Networks*, Taiwan, September, 87-93.

[18] Kanzaki, K., Ma, Q., Yamamoto, E., Murata, M., and Isahara, H. (2003). Adjectives and their Abstract concepts — Toward an objective thesaurus from semantic map, Proceedings of the Second International Workshop on Generative Approaches to the Lexicon, 177-183.

[19] Kanzaki, K., Ma, Q., Yamamoto, E., Murata, M., Isahara, H. (2004). Extraction of hyperonymy of adjectives from large corpora by using the neural network model, to appear in the Fourth International Conference on Language Resources and Evaluation (LREC2004), Lisbon.

[20] Kohonen, T. (1984). Self-organization and associative memory. Springer Series in Information Science. Vol. 8, *Springer*.

[21] Kurohashi, S., Nagao, M. (1997). Kyoto University text corpus project. *Proc. 3rd Annual Meeting of the Association for Natural Language Processing*, 115-118 (in Japanese).

[22] Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. *ACL-EACL'97*, Madrid, 64-71.

[23] Ma, Q., Kanzaiki, K., Murata, M., Uchimoto, K., and Isahara, H. (2000). Self-organizing semantic maps of Japanese nouns in terms of adnominal constituents. *IJCNN'2000*, Como, Italy, Vol. VI, 91-96.

[24] Macklovitch, E., Hannan, M. L. (1996). Line 'em up: advances in alignment technology and their impact on translation support tools. *in Expanding MT Horizons, Second Conference of the Association for Machine Translation in the Americas*. Montreal, 145-156.

[25] Matsumoto, Y., Ishimoto, H, Utsuro, T. (1993). Structural matching of parallel texts, *ACL'93*, 23-30.

[26] Mori, S. and Nagao, M. (1998). A stochastic language model using dependency and its improvement by word clustering. *COLING-ACL'98*, Vol. 2, pp. 898-904.

[27] Murata, M., Ma, Q., Uchimoto, K., Ozaku, H., Utiyama, M. and Isahara, H. (2000). Japanese probabilistic information retrieval using location and category information. *IRAL'2000*, Hong Kong, 81-88.

[28] Press, W. H., Teukolsky, S. A., Vetterling W. T., and Flannery, B. P. (1994). Numerical recipes in C, Cambridge University Press, 2nd Edition.

[29] Ritter, H. and Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61, 241-254.

[30] Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. *ACM SIGIR'94*, Dublin, Ireland.

[31] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Vol. 28, No. 1, 11-21.

[32] Varea, I. G., Och, F. J., Casacuberta, F. (2002). Improving alignment quality in statistical machine translation using context-dependent maximum entropy models, *COLING2002*, 1051- 1057.

[33] Wu, D. (1995). An algorithm for simultaneously bracketing parallel texts by aligning words. *ACL-95*, 244-251.

[34] Zhou, Q. Duan, H. (1994). Segmentation and tagging in modern Chinese corpus. *Chinese Journal of Computers*, Vol. 85.