Self-Organizing Chinese and Japanese Semantic Maps

Qing Ma†, Min Zhang‡, Masaki Murata†, Ming Zhou§, Hitoshi Isahara†
†Communications Research Laboratory, Kyoto 619-0289, Japan
‡State Key Lab of Intelligent Tech. and Sys, Tsinghua University, Beijing 100084, China
§Microsoft Research Asia, Beijing 100080, China
E-mail: qma@crl.go.jp

Abstract

This paper describes a corpus-based connectionist approach to the development of self-organizing Chinese and Japanese semantic maps, proposing an improved coding method using TFIDF term-weighting and newly introducing a numerical evaluation for objectively judging the results. The adaption of TFIDF term-weighting is proved to be effective by experimental comparisons with five other coding methods. The effectiveness and necessity of the proposed method for creating semantic maps are clarified by comparisons with a conventional clustering technique and multivariate statistical analysis.

1 Introduction

Computing word similarity in meanings is an important technique that can be applied in 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; Karov and Edelman, 1996; Lin, 1997) 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, computing using these approaches is performed 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, the word similarity is computed using a mathematical measure of vector distance.

In practical applications, e.g., to find the optimal query expansion in information retrieval, words must be further sorted globally based on the prior computing of word similarity. Usually, the sorting is done using various clustering techniques. Word clustering, however, only classifies words into several groups. It is difficult, or at least not intuitive, to recognise the relationships between groups or the relationships between words within groups. To begin with, it may even be a problem 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 technique to word clustering to sort words. We need a technique that maps words from a very large lexicon into a small semantic space, i.e., a visible representation in which words with similar meanings are placed at the same or neighboring points so that the distance between the points represents the semantic similarity of the words. This representation is called a semantic map.

There have been several studies on developing semantic maps for English (Ritter and Kohonen, 1989) and also for Chinese and Japanese (Ma et al., 2000 and Ma et al., 2001) using a corpus-based connectionist approach. In these studies, a self-organizing neural network, which was proposed by Kohonen (1984) and called a self-organizing map (SOM), has been adopted as an unsupervised learning machine. In these self-organizing English maps, however, co-occurrent words were gathered only from three-word windows, and each word was then simply coded (i.e., transformed into vectors) according to randomly generated patterns of co-occurrent words. In studies to develop Chinese and Japanese maps, on the other hand, co-occurrent words were gathered according to their grammatical relationships, i.e., adjective/noun–noun

in Chinese and adjective/nominal adjectival–noun in Japanese. The words then were coded, taking into account the semantic correlation between the words, which was established using a form of word similarity computing. Coding based on word similarity computing used in developing Chinese and Japanese maps has made remarkable progress compared to the random coding used to develop English maps[1]. However, the coding method is still too crude: cooccurence frequency, which is essential information for judging the importance of a word, has not been utilized in word similarity computing. Because the self-organization of semantic maps is actually a form of sorting based on word similarity computing, coding based on word similarity computing significantly affects the map created. It is, therefore, very important in developing semantic maps to establish the optimal coding method. In addition, in previous studies, there has been no numerical evaluation used to objectively judge the maps developed; i.e., they were only evaluted subjectively by intuition.

This paper describes a corpus-based connectionist approach to the development of self-organizing Chinese and Japanese semantic maps. An improved coding method using TFIDF term-weighting is proposed to solve the problems existing in the previous coding method. The evaluations for the maps created are performed by the numerical evaluation using the newly introduced accuracy, recall, and F-measure, as well as by intuition, and by the comparisons with a clustering method and multivariate statistical analysis.

2 SOM

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. It is as if some meaningful nonlinear coordinate system for different input features was created over the network. Such a learning process is competitive and unsupervised and is called a self-organizing process.

When input vector $x \in \Re^n$ is given, it is compared to all reference vectors $m_i \in \Re^n$, which are associated by each node and is gradually modified in the learning process, and the network responses comply with the two different stages, learning and mapping, as follows. In the mapping stage, only the node whose reference vector has the smallest Euclidean distance to the input vector is activated. This node, c, is called a best-matching node or a winner. It can thus be defined by

$$c = argmin_i\{||x - m_i||\}. \quad (1)$$

In the learning stage, on the other hand, not only the best-matching node but also its neighborhood nodes are activated and their reference vectors are changed so that they are closer to the same sample input vector x. This will result in a local relaxation or smoothing effect on the reference vectors of nodes in this neighborhood, which in continued learning leads to global ordering. This gradual adaption of reference vectors can be expressed as

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)], \quad (2)$$

where $h_{ci}(t)$ is the neighborhood function. For convergence it is necessary that $h_{ci}(t) \to 0$ when $t \to \infty$. A widely applied neighborhood function can be written in terms of a Gaussian function,

$$h_{ci}(t) = \alpha(t) \cdot \exp(-\frac{||r_c - r_i||^2}{2\sigma^2(t)}). \quad (3)$$

Here, $r_c \in \Re^2$ and $r_i \in \Re^2$ are the location vectors of nodes c and i, respectively. Term $\alpha(t)$ is the learning rate and $\sigma(t)$ defines the radius of the neighborhood. Both terms are monotonically decreasing functions of time, and their exact forms are not critical.

Usually, the learning process consists of an ordering phase and a fine adjustment phase. In the ordering phase, learning rate $\alpha(t)$ and radius of the neighborhood $\sigma(t)$ should start with larger values. The ordering of m_i occurs during this initial period, while the remaining steps are only needed to finely adjust the map. In the fine adjustment phase, the radius can still contain the nearest neighbors of node c, and $\alpha = \alpha(t)$ should attain a low value over a long period.

3 Data Coding for Self-Organizing Semantic Maps

Suppose there is a set of words w_i $(i = 1, \cdots, n)$ that we are planning to self-organize. Word w_i

[1] A previous study showed that the random coding method is not applicable in creating Japanese maps.

can be defined by a set of its co-occurrent words as

$$w_i = \{a_1^{(i)}, a_2^{(i)}, \cdots, a_{\alpha_i}^{(i)}\}, \quad (4)$$

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

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

$$V(w_i) = [d_{i1}, d_{i2}, \cdots, d_{in}]^T. \quad (5)$$

The $V(w_i) \in \Re^n$ is the input to the SOM. That is, the role of the SOM is to manifest the semantic relationships existing in such high-dimensional vectors and represent them in two-dimensional space. Therefore, the method of computing word similarity d_{ij} is a key point in the coding.

3.1 The previous method

In the previous method, word similarity d_{ij} between word w_i and w_j is measured by

$$d_{ij} = \begin{cases} \frac{(\alpha_i - c_{ij}) + (\alpha_j - c_{ij})}{\alpha_i + \alpha_j - c_{ij}} & \text{if } i \neq j \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

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

3.2 TFIDF term-weighting method

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

In this method, the word similarity d_{ij} between word w_i and w_j is measured by

$$d_{ij} = \begin{cases} \frac{(T_i - T_{ij}) + (T_j - T_{ij})}{T_i + T_j - T_{ij}} & \text{if } i \neq j \\ 0, & \text{otherwise,} \end{cases} \quad (7)$$

where T_i and T_j are, respectively, the expansions of the number, α_i and α_j, of co-occurrent words of w_i and w_j, and T_{ij} is an expansion of the number, c_{ij}, of co-occurrent words that both w_i and w_j have in common. They are calculated by

$$T_i = \sum_{x=1}^{x=\alpha_i} t_x^{(i)} \quad \text{and} \quad T_{ij} = \sum_{x=1}^{x=c_{ij}} t_x^{(ij)}, \quad (8)$$

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

$$t_x^{(i)} = tf(a_x^{(i)}, w_i) \cdot idf(a_x^{(i)}), \quad (9)$$

and

$$t_x^{(ij)} = tf(a_x^{(i)}, w_i, w_j) \cdot idf(a_x^{(i)}). \quad (10)$$

Here, $tf(a_x^{(i)}, w_i)$ is the co-occurrence frequency of co-occurrent 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 $a_x^{(i)}$ appeares in all the headwords, i.e.,

$$idf(a_x^{(i)}) = \log \frac{n}{df(a_x^{(i)})} + 1. \quad (11)$$

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

The TFIDF value $t_x^{(i)}$ (including $t_x^{(ij)}$) is therefore a weight reflecting the importance of co-occurrent word a_x^i for word w_i: the larger the value, the higher the co-occurrence frequency with word w_i and the lower the co-occurrence

frequency with other words, and, therefore, the higher the importance of the co-occurrent word a_x^i for word w_i. If we consider all co-occurrent words as having the same importance to each headword, then we can set all TFIDF values $t_x^{(i)} = 1$. In this case, $T_i = \alpha_i$ and $T_{ij} = c_{ij}$ by Eq. (8). Thus, Eq. (7) becomes the same as Eq.(6). The Eq. (7) is therefore an expansion of Eq. (6).

3.3 Other methods

By replacing the TFIDF values $t_x^{(i)}$ and $t_x^{(ij)}$ with the frequencies $f_x^{(i)}$ and $f_x^{(ij)}$ in Eq. (8), we can easily establish a frequency term-weighting method. In the coding methods described so far, the word similarity computing is performed by using a form of simple set operation as shown in Eq. (6) and Eq. (7). Instead of using such a set operation, we can also consider using vector calculation or information theory. On the basis of these methods, we worked out three further methods: cosine measure (Frakes, 1992) of vectors composed of frequencies, cosine measure of vectors composed of TFIDF values, and entropy calculation of vectors composed of frequencies.

4 Experimental Results

4.1 The data

In Chinese, to evaluate the experimental results more easily and objectively, the headwords (in total, 85 nouns) were selected from six categories in "The Contemporary Chinese Classified Dictionary" (Dong et al., 1998). We also added several Chinese family names which are not in the dictionary but which appear frequently in newspapers as a new category. The co-occurent words were adjectives and nouns which, together with headnouns, form the noun phrases and were gathered by computer from eleven years of "The People's Daily." The total number of co-occurrent words was 69,030 and the number of different co-occurrent words was 22,118.

In Japanese, we used noun phrases composed of nouns and adjectives/nominal adjectivals, gathered by computer from eight years of the Mainichi Shinbun newspaper in order of frequency of co-occurring adjectives/nominal adjectivals. The number of nouns was 100, the total number of co-occurrent words was 33,870, and the number of different co-occurrent words was 4,023.

4.2 The 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.

4.3 Evalution methods

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

$$P = \frac{\sum_{i=1}^{C} p_i}{C}, R = \frac{\sum_{i=1}^{C} r_i}{C}, \tag{12}$$

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

$$p_i = \frac{\#(\text{words correctly classified as class } i)}{\#(\text{words of class } i \text{ in result})}, \tag{13}$$

$$r_i = \frac{\#(\text{words correctly classified as class } i)}{\#(\text{words of class } i)}, \tag{14}$$

where # means number.

(b) Intuitive evaluation: Because the most remarkable features of maps are their visibility and the continuity of the classifications, numerical evalution only of classifications is insufficient. It is therefore very important to judge whether the maps are intuitively meaningful.

(c) Comparison with other methods: To assess the effectiveness of the proposed method in semantic classification, a comparison with a conventional clustering technique was performed. In addition, to assess the usefulness of the proposed method for developing semantic maps, we investigated whether it was feasible to

use multivarite statistical analyses such as principle component analysis to construct such visible representations of semantic classification.

4.4 The Results

Because the Chinese headwords were manually selected from a semantic dictionary, and we know their exact semantic categories, we were able to perform the numerical evalution proposed in this paper for Chinese maps and for classifications obtained using the hierarchical clustering technique. Table 1 shows the results of numerical evalution ranked in order of the F-measure. In the table, the results obtained using two other coding methods of cosine measures are not included, because all the words in the maps using these two coding methods were merged together and the evalution of them does not make sense. This table shows that (i)TFIDF term-weighted coding produces much higher precision than all other coding methods, and the recall using TFIDF term-weighted coding is also higher than for all other coding methods; (ii)the F-measure of the classifications in maps based on both the previous coding method and the TFIDF term-weighted coding method are higher than those for the clustering results produced by these two codings; (iii)the F-measure of the map based on the frequency term-weighted coding method was slightly higher than that of the map based on the previous coding method, which is higher than that of the map based on the entropy coding method in turn. In addition, both the precision and recall of the clustering by adopting the TFIDF term-weighted coding method is much higher than those of the clustering with the previous coding method.

In Figure 1, (a) shows a semantic map of Chinese nouns self-organized using the TFIDF term-weighted coding method, 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, SHOUFA (method), JINGTOU (spirit), DAOYE (broker), DIANZI (idea), JUEQIAO (knack), and ZHENGFU (government) were mapped in incorrect areas in the sense that not only they were different from the definition in the dictionary, but also they were intuitively inconsistent. Even among these nouns, however, the noun JINGTOU (spirit) was mapped near the correct area *emotion*. In addition, one noun, SHIJIE (world), in the area of *sports games* was mapped incorrectly in the sense that it differed from the definition in the dictionary, but its location is 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 *method*, the category *sports* was split into two groups: one for actual *sports* and one for *sports games*. The category *business* was split into two groups: one for *business* and one for *economy*. These classifications clearly do not contradict the original one in essence[2)]. That is, the self-organized map is basically consistent with the definitions found in the Chinese dictionary. Naturally, it is also generally intuitively consistent.

Table 1: Comparative result of various coding methods and clustering.

	Precision	Recall	F-measure
Clustering*[1]	0.936	0.864	0.899
Entropy	0.925	0.874	0.899
Previous	0.926	0.90	0.913
Frequency	0.928	0.90	0.914
Clustering*[2]	0.95	0.896	0.922
TFIDF	0.944	0.907	0.925

*[1]Using the previous coding method
*[2]Using the TFIDF term-weighted coding method

A comparison of the self-organized map with the classifications obtained using hierarchical clustering shows both methods produce similar results. And, as with the map, a total of 85 nouns were also divided into the completely same eight categories.

Principle component analysis of the same Chinese data using TFIDF term-weighted coding showed that the cumulative coefficients of determination for the top two and ten principle components were 8.29% and 24.53%, respectively. In general, if the value is not larger than 80%, the multivariate data cannot be com-

[2)]The precision and recall shown in Table 1 were calculated by taking the categories "sports" and "sports game" as a whole, and "method" and "policy" as a whole.

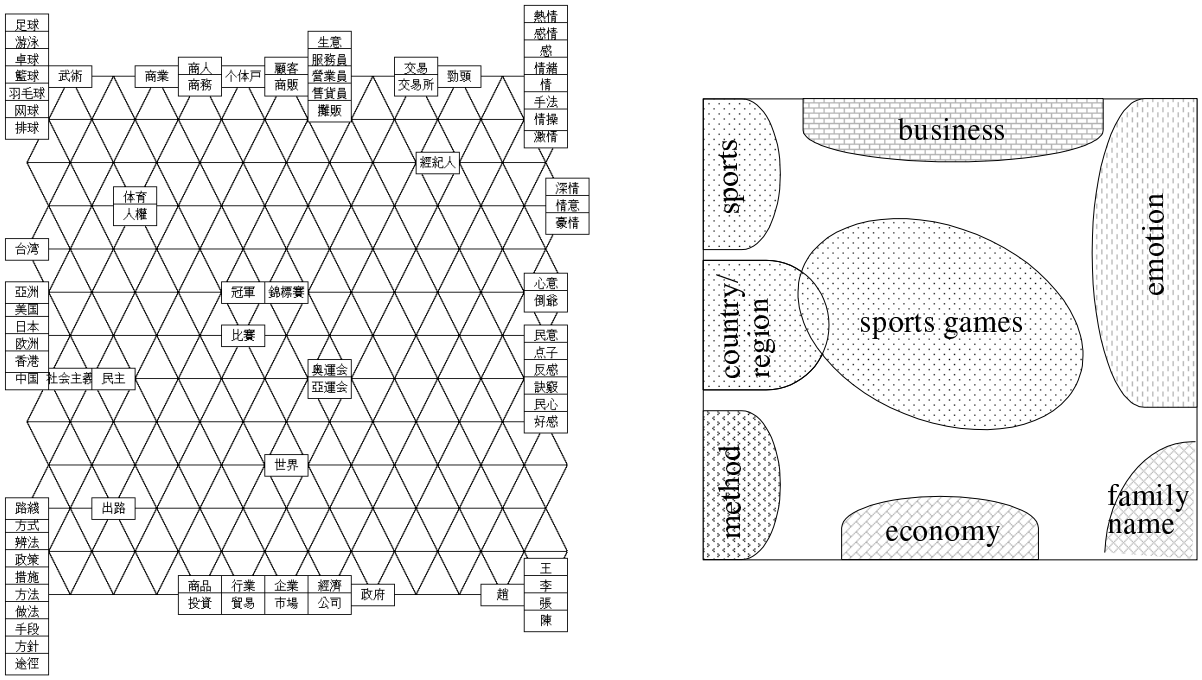

Figure 1: Chinese semantic map based on TFIDF term-weighted coding method.

pressed into a small number of principle components. Therefore, it is difficult to construct good semantic maps using multivariate statistical analysis. An experiment to plot out the data using the first and second principle components was performed. The results are shown in Figure 2. Clearly, almost all of the nouns were concentrated on some right area and this method cannot be used to create a meaningful map.

It was not possible to perform a numerical evalution of the Japanese maps because all the Japanese nouns were gathered using an automatic method from newspaper and their exact meaning was not known. We can, however, judge intutively that the Japanese semantic map (Figure 3) obtained using the TFIDF term-weighting method was better than maps generated using other coding methods and the classification was not inferior to that obtained using hierarchical clustering. In addition, a principle component analysis on the same data with the TFIDF term-weighted coding method showed that the cumulative coefficient of determination for the top two and ten principle components was 7.317% and 22.679%, respectively. The plotted result using the first and second principle shows that all words are merged together and this method cannot be used to create a meaningful Japanese map.

5 Conclusion

This paper proposed a corpus-based connectionist approach to the development of self-organizing Chinese and Japanese semantic maps, proposing an improved coding method using TFIDF term-weighting and introducing a numerical evalution for objectively judging the results. The results of computer experiments showed that meaningful Chinese and Japanese maps can be created and that the adaption of the TFIDF term-weighting was effective. Comparisons with the clustering technique and multivariate statistical analysis showed that the proposed method produces appropriate classification and is suitable for creating visible representation of classification results. In the future work, we aim to enlarge the scale of semantic maps to the level for practical use and then apply them to meaning-based Chinese and Japanese information retrieval tasks.

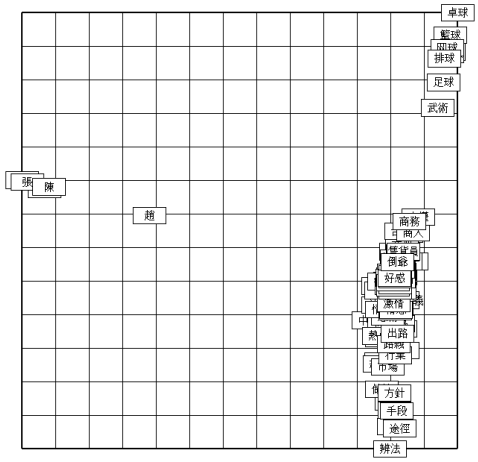

Figure 2: Chinese semantic map using principle component analysis.

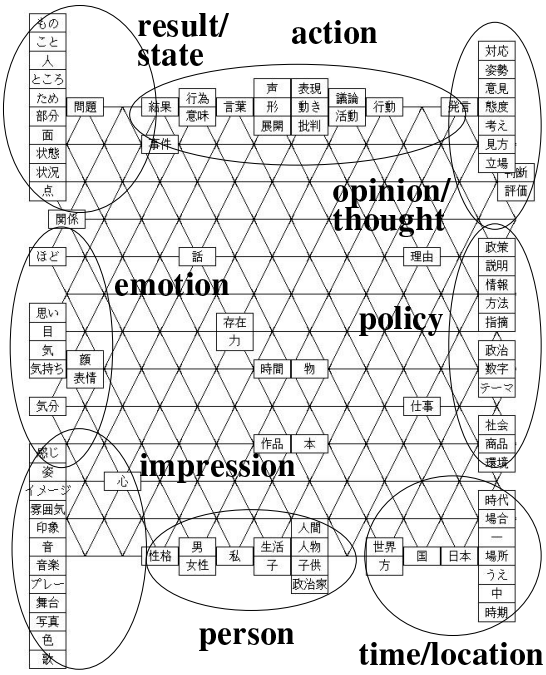

Figure 3: Japanese semantic map based on T-FIDF term-weighted coding method.

References

Dagan, I., Marcus, S., and Markovitch, S.: Contextual word similarrity and estimation from sparse data, *ACL'93*, Columbus, Ohio, pp. 164-171, 1993.

Dagan, I., Pereira, L, and Lee, L.: Similarity-based estimation of word cooccurrence probabilities, *ACL'94*, Las Cruces, NM, pp. 272-278, 1994.

Dong, D. N., et al. (Eds.), Contemporary Chinese Classified Dictionary, *Han-Yu-Da-Ci-Dian Press*, 1998.

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

Harris, ZS.: Mathematical structures of language, New York: Wiley, 1968.

Hindle, D.: Noun classification from predicate argument structures, *ACL'90*, Pittsburgh, PA, pp. 268-275, 1990.

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

Kohonen, T.: Self-organization and associative memory, Springer Series in Information Science, Vol. 8, *Springer*, 1984.

Lin, D.: Using syntactic dependency as local context to resolve word sense ambiguity, *ACL-EACL'97*, Madrid, pp. 64-71, 1997.

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

Ma, Q., Zhang, M., Zhou, M., Huang, C., and Isahara, H.: Emergence of Chinese semantic maps from self-organization, *ICONIP'2001*, Shanghai, China, pp. 681-686, 2001

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

Murata, M., Ma, Q., Uchimoto, K., Ozaku, H., Utiyama, M. and Isahara, H.: Japanese Probabilistic Information Retrieval Using Location and Category Information, *IRAL'2000*, pp. 81-88, Hong Kong, 2000.

Ritter, H. and Kohonen, T.: Self-organizing semantic maps, *Biological Cybernetics*, 61, pp. 241-254, 1989.

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

Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, Vol. 28, No. 1, pp. 11-21, 1972.