Semantic Maps for Word Alignment in Bilingual Parallel Corpora

Qing Ma
Ryukoku University
Otsu 520-2194, Japan
qma@math.ryukoku.ac.jp

Yujie Zhang
Communications Research Laboratory
Kyoto 619-0289, Japan
yujie@crl.go.jp

Masaki Murata
Communications Research Laboratory
Kyoto 619-0289, Japan
murata@crl.go.jp

Hitoshi Isahara
Communications Research Laboratory
Kyoto 619-0289, Japan
isahara@crl.go.jp

Abstract

Effective self-organizing techniques for constructing monolingual semantic maps of Japanese and Chinese have already been developed. By extending the monolingual map to a bilingual semantic map, we have proposed a semantics-based approach for word alignment in a Japanese/Chinese bilingual corpus.

1 Introduction

Acquiring translation knowledge from a bilingual parallel corpus requires alignment not only at the sentence level but also at the word level. If a bilingual corpus is aligned at the word level, translation words that are not in a dictionary, such as those depending on domain or time, might be obtained, or, multiple translation candidates might be scored. Furthermore, translation patterns based on the relations of words at the phrase or clause level might be automatically acquired (Brown, 1997). Thus, alignment is a very important, fundamental task in natural language processing (NLP). The research related to this topic includes a series of statistical models (e.g., Brown, et al., 1988; Brown, et al., 1993; Macklovitch and Hanna, 1996), a method using dynamic programming (Dagan, 1993), a statistical approach introducing contextual information (Varea, et al., 2002), and structure alignment methods (Kaji, et al., 1992; Matsumoto, et al., 1993; Wu, 1995; Imamura, 2001). All of these approaches, however, are based on either statistical information or grammatical structure, but not on meaning.

Automatic methods for constructing monoligual semantic maps of Japanese or Chinese have already been proposed (Ma, et al., 2002). In a monoligual semantic map, 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. If a bilingual semantic map could be automatically constructed by accepting translation pairs of sentences as inputs, word alignment would be easily obtained from the map. Since the bilingual semantic map, like the monolingual semantic map, would provide results with visibility and continuity, it would be easy to handle one-to-many or many-to-one alignment. Furthermore, bilingual maps can perhaps be expected to be applied in foreign language learning or foreign language writing by using bilingual parallel corpora. The most important factor is that the translations should usually be free. There is an evident limitation of existing alignment methods that rely on statistical or grammatical information, which suggests the necessity to develop an approach based on meaning.

This paper proposes a new method for automatically constructing bilingual semantic maps of Japanese and Chinese. the method accepts translation pairs of Japanese and Chinese sentences as inputs, with the aiming of providing word alignment based on meaning[1] . We used the Kyoto University Japanese corpus and its translated Chinese corpus to conduct an experiment and confirm the effectiveness of the proposed method. The necessary training data for automatically constructing semantic maps was obtained from eight years of a Japanese newspaper, Mainichi Shinbun.

[1] Since present semantic maps are constructed basically on co-occurrent information, the method proposed is not strictly a semantic approach. However, since our final goal is to develop true semantic maps, and alignment based on meaning itself is a very important idea, we use the expression "based on meaning", without fear of misunderstanding.

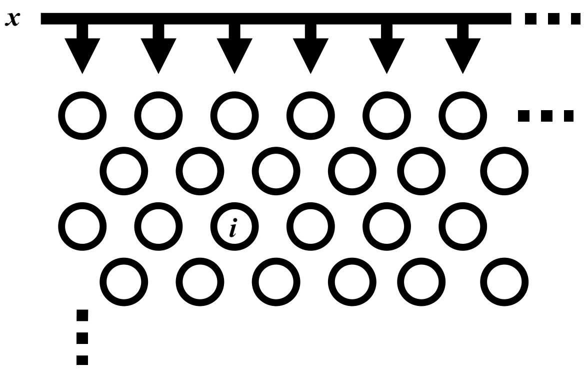

Figure 1: Two-dimensional SOM.

2 Self-organizing Neural Networks

As an automatic method for constructing semantic maps, we have adopted a self-organizing neural network, which was proposed by Kohonen in the early 1980s (see details in Kohonen, 1997) and is called a self-organizing map (SOM). An SOM can be visualized as a two-dimensional array (Figure 1) 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.

Suppose input $x = [\xi_1, \xi_2, \cdots, \xi_n]^T \in \Re^n$, where $\Re^n$ is an n-dimensional space. Each node i is then associated with a parametric reference vector m_i, which equals $[\mu_{i1}, \mu_{i2}, \cdots, \mu_{in}]^T \in \Re^n$, whose element μ_{ij} is a scalar weight between node i and input element ξ_j and is gradually modified in the learning 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 the best-matching node or winner. It can thus be defined by

$$c = argmin_i \{ ||x - m_i|| \}. \tag{1}$$

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 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 in continued learning. This gradual adapttaion of the reference vectors can be expressed as

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

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:

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

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 of the latter terms are monotonically decreasing functions of time, and their exact forms are not critical. They can thus be defined linearly as

$$\alpha(t) = \alpha(0) \frac{T - t}{T}, \tag{4}$$

and

$$\sigma(t+1) = 1 + (\sigma(t) - 1) \frac{T - t}{T}, \tag{5}$$

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. (4) and (5). The ordering of m_i occurs during this initial phase, while the remaining steps are only needed for finely adjusting the map. After the ordering phase, the radius may still contain the nearest neighbors of node c, and $\alpha = \alpha(t)$ should attain a low value over a long period.

3 Self-organizing Semantic Map for Word Alignment

3.1 Purpose

When a translation pair of sentences like

(Japanses) 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 。

is given, to self-organize a semantic map for word alignment means to automatically map all the words in the given sentences by employing some kind of unsupervised learning data.

3.2 Learning data

As part of the Japanese-Chinese machine translation project, we are constructing a bilingual parallel corpus based on the Kyoto University Japanese corpus (Kurohashi and Nagao, 1997). The translation pair of sentences were obtained from the corpus. Since the Kyoto University corpus has already been morphologically analyzed, the Japanese sentences were used directly without any analysis, while the translated Chinese sentences were segmented and part-of-speech tagged by using the morphological analysis tool developed by Beijing University (Zhou and Duan, 1994).

To evaluate the two different languages with the same measure, the words appearing in a translated Chinese sentence were given at most five translated Japanese candidate words, which were used instead of the original Chinese words. The candidates were obtained manually[2] from two Chinese-Japanese dictionaries: "Han Ri Ci Dian", published by Jilin Education Publisher, and "Chunichi Daijiten"[3] , published by Taishukan Publishing Co., Ltd. The candidates were selected according to the following order of priority: (1) a word that also appears in the Japanese original sentence; (2) a word that has the same POS as the original Chinese word; (3) a word chosen according to the order listed in the dictionary; and (4) a word that appears in the Kyoto University corpus. Thus, all the words in the translated Chinese sentence in the pair shown above can be rendered in terms of Japanese candidate words 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 in terms of only Japanese words. As this example shows, however, we can recognize translated Japanese candidate words, such as "これによって" or "ことができる/てよい", that do not exist in the original Japanese sentence. This means that it is virtually impossible to perform word alignment 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 were obtained in the following way. Each Japanese word appearing in a Japanese sentence was defined in terms of its co-occurrent words (the targeted word itself and the words to its immediate left and right). They were obtained from eight years (1991-1998) of the Japanese newspaper, Mainichi Shinbun, and used as learning data. Each Chinese word appearing in a translated Chinese sentence was defined in terms of the co-occurrent words of its Japanese translation candidates and the Chinese words defined in this way were used as learning data. In the next section, we explicitly describe the construction of the learning data and the coding method used to transform it into inputs for the SOM.

3.3 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 the J_i $(i = 1, \cdots, m)$ are the Japanese words forming the Japanese sentence, the C_i $(i = 1, \cdots, n)$ are the 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 the paper).

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

$$w_i = J_i = \{a_1^{(i)}, f_1^{(i)}, \cdots, a_{\alpha_i}^{(i)}, f_{\alpha_i}^{(i)}\}, \qquad (6)$$

where $a_j^{(i)}$ is a co-occurrent 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-occuring with J_i. Word w_j $(= C_j)$ of a translated Chinese sentence is also defined by a set of co-occurrent information:

$$w_j = C_j = \{J_{j1}, \cdots, J_{j,n_j}\} = \{a_1^{(j)}, f_1^{(j)}, \cdots, a_{\alpha_j}^{(j)}, f_{\alpha_j}^{(j)}\}, \qquad (7)$$

where $a_i^{(j)}$ is a co-occurrent word of either or severals of $J_{j1}, \cdots, J_{j,n_i}$, $f_i^{(j)}$ is the normalized co-occurrence frequnecy (it will be the summation of the frequencies when occuring 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-occurrent Japanese words, there is no need to distinguish between Chinese and Japanese, and it thus becomes possible to apply all existing coding methods for self-organizing monolingual semantic maps. In this paper, the semantic distance d_{ij} between any two words w_i and w_j appearing in a

[2] Since there are no online electronic dictionaries, we have to obtain the data manually at present.

[3] This dictionary was used only when a word had no entry in the former dictionary.

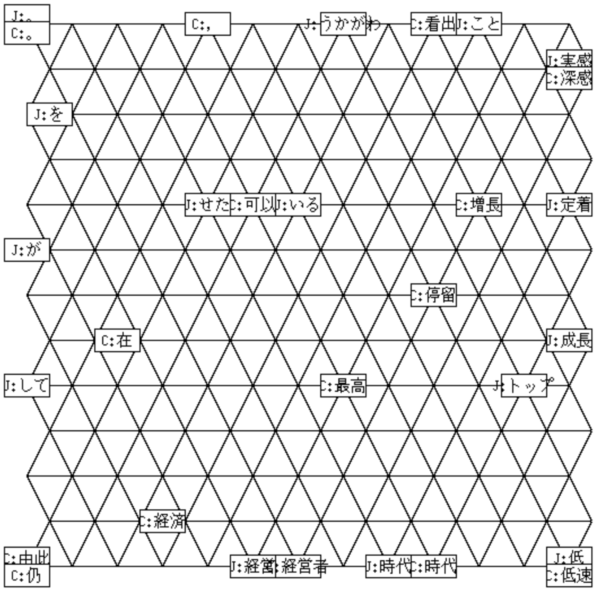

Figure 2: Semantic map obtained by self-organization.

translation pair of sentences is calculated by the following frequency term-weighting method:

$$d_{ij} = \begin{cases} \frac{(F_i - F_{ij}) + (F_j - F_{ij})}{F_i + F_j - F_{ij}} & \text{if } i \neq j \\ 0, & \text{otherwise,} \end{cases} \qquad (8)$$

where F_i and F_j are expansions of α_i and α_j, the numbers of co-occurrent words of w_i and w_j, respectively, and F_{ij} is an expansion of c_{ij}, the number of co-occurrent words that both w_i and w_j have in common. The expansion are obtained as follows:

$$F_i = \sum_{x=1}^{x=\alpha_i} f_x^{(i)} \quad \text{and} \quad F_{ij} = \sum_{x=1}^{x=c_{ij}} f_x^{(ij)}, \qquad (9)$$

where $f_x^{(i)}$ is the co-occurrence frequency of word w_i and its co-occurrent word $a_x^{(i)}$, and $f_x^{(ij)}$ is the co-occurrence frequency of words w_i and w_j and their co-occurrent word $a_x^{(i)}$ $(x = 1, \cdots, \alpha_i)$. As a result, we can define a correlative matrix D with the distance d_{ij} as its element. Each word w_i is thus coded with the elements in the i-th row of the correlative matrix D as

$$V(w_i) = [d_{i1}, d_{i2}, \cdots, d_{iN}]^T, \qquad (10)$$

where N is the total number of words appearing in the translation pair of sentences (i.e., $N = m + n$), and $V(w_i) \in \Re^N$ is the input to the SOM.

Table 1: Word alignment results obtained from the semantic map

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

4 Experimental Results

4.1 Data

Word alignment experiments were performed for ten translation pairs of sentences. The learning data was obtained in the way described in Sec. 3.2. Considering the translation pair of sentences given in Sec. 3.1 as an example, the number of words was $N = m + n$=16+15=31, the total number of co-occurrent words was 62,627, and the number of different co-occurrent words was 22,077. Among the 31 words, the period symbol ("。")[4] had the largest number of co-occurrent words (4,180), while the word "うかがわ" in the Japanese sentence and the comma "，" in the translated Chinese sentence had the smallest numbers of co-occurrent words (5 each).

4.2 SOM

We used an SOM consisting of a 13×13 two-dimensional array. The number of input dimensions, N, was 31, the same as the number of words to be mapped. In the ordering phase, the number of learning steps, T, was set to 10,000, the initial value of the learning rate, $\alpha(0)$, was set to 0.1, and the initial radius of the neighborhood, $\sigma(0)$, was set to 13, equal to the diameter of the SOM. In the fine adjustment phase, T was set to 100,000, $\alpha(0)$ was set to 0.01, and $\sigma(0)$ was set to 7. The initial reference vectors $m_i(0)$ consisted of random values between 0 and 1.0.

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

Table 2: Baseline word alignment results

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

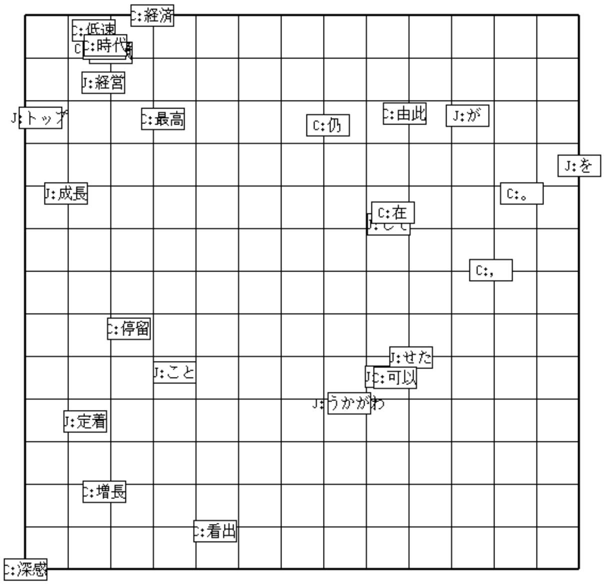

Figure 3: Semantic map by PCA.

4.3 Results

Figure 2 shows the semantic map for word alignment of the translation pair given in Sec. 3.1 as an example. Here, the words tagged with "J" are Japanese words from the Japanese sentence and the words with "C" are Chinese words from the translated Chinese sentence[5] . From the map, we could obtain the word alignment results listed in Table 1 by focusing on each Japanese word and choosing the closest Chinese word to it[6] . The correct answers are also given in the table. From this table, we can see that [J：低, C：低速],[J：時代, C：時代], [J：実感, C：深感],[J：うかがわ, C：看出],[J：せた, C：可以], [J:。, C:。] were aligned correctly. Among these pairs, in the case of [J:うかがわ，C:看出] and [J:せた, C:可以], the Japanese word and the Japanese translation candidates for the Chinese words have different surface representations. The other alignment results are incorrect in the strict sense of the word. Among these apparent mistakes, however, there are some interesting results. For example, for the Japanese word "J:成長", although "C:停留" was aligned as the closest Chinese word, the semantic map shows that the second closest Chinese word is actually "C:増長". That is, if we had included the second closest candidate, we would have obtained the correct answer. Similarly, for "J：トップ", its second candidates is "C：最高", a correct answer. Also, the incorrect alignment results, [J:こと，C:看出] and [J:を，C:。] were due to the fact that there are no Chinese words (or at least none appearing in the sentence) that correspond to these Japanese words. Another problem was incorrect alignment caused by the inconsistency of word segmentation between the Japanese and Chinese sentence, as in the case of [J:経営:，C:経営者]. None of these problems can be resolved by only applying the word alignment technique.

Table 2 lists the baseline word alignment results, which were obtained by focusing on each Japanese word and choosing the Chinese word with the smallest semantic distance d_{ij} which was calculated by Eq. (8). From this table we can see that [J：うかがわ, C：深感] was incorrect, while "J：うかがわ" was correctly aligned by using the semantic map. Also, although incorrect results were obtained both for the semantic map, such as [J:成長, C:停留] or [J:停留, C:増長] and for the baseline such as, [J:成長, C:時代] or [J:停留, C:深感], the results for the semantic map were somewhat correct in meaning, whereas those for the baseline were totally wrong. If we see the second candidates, we can know that the second candidates of "J:成長" and "J: トップ" are "C:深感" and "C:時代" which are incorrect. We can thus say that the method using the semantic map performed better than the baseline method.

Figure 3 shows the word-alignment semantic map obtained by principle component analysis (PCA). By

[5] Because both Japanese and Chinese use Chinese characters, it is necessary to use such symbols to distinguish between languages.

[6] These results were obtained by only selecting the Chinese word closest the Japanese word. If the second closest or third closest words are also used, the word alignment results can include multiple candidates.

comparing it with Figure 2, we can see that the results obtained by PCA would be worse than those obtained by self-organization. For example, the pair [J：うかがわ，C：看出], which has different surface representations, could not be obtained by PCA. "J：成長" also could not be correctly aligned even if the second closest candidate were 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 for word alignment. 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] also could not be correctly obtained with the clustering method. Moreover, because we could not know the semantic distance between words within a group, we could not easily obtain second closest candidates as we could with the semantic map.

5 Conclusion

This paper has proposed a novel word alignment method designed to provide a meaning-based approach. The effectiveness of the proposed method was confirmed through small-scale experiments. In our future work, we are going to conduct numerical evalution through large scale experimental comparison with existing methods. We also plan to develop a word alignment technique for practical use by integrating the proposed method into existing systems.

References

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

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Mercer, R. L.: The mathematics of statistical machine translation: parameter estimation, *Computational Linguistics*, Vol. 19, No. 2, pp.263-311, 1993.

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

Dagan, I., Church, K. W., Gale, W. A.: Robust bilingual word alignment for machine aided translation, *Proceedings of the Workshop on Very Large Corpora*, pp. 1-8, 1993.

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

Kaji, H., Kida, Y., Morimoto Y.: Learning translation templates from bilingual text, *COLING'92*, pp. 672-678, 1992.

Ma, Q., Zhang, M., Murata, M., Zhou, M., Isahara, H.: Self-Organizing Chinese and Japanese Semantic Maps, The 19th International Conference on Computational Linguistics (COLING'2002), Taiwan, pp. 605-611, August, 2002.

Matsumoto, Y., Ishimoto, H, Utsuro, T.: Structural matching of parallel texts, *ACL'93*, pp. 23-30, 1993.

Macklovitch, E., Hannan, M. L.: 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, pp. 145-156, 1996.

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

Wu, D.: An algorithm for simultaneously bracketing parallel texts by aligning words, *ACL-95*, pp. 244-251, 1995.

Imamura, K.: Hierarchical phrase alignment harmonized with parsing, *NLPRS2001*, pp. 377-384, 2001.

Kohonen, T.: Self-organizing maps, Springer, 2nd Edition, 1997.

Zhou, Q. Duan, H.: Segmentation and tagging in modern Chinese corpus, *Chinese Journal of Computers*, Vol. 85, 1994.