Homographic Ideogram Understanding Using Contextual Dynamic Network

Jun Okamoto[†], Shun Ishizaki[§]

[†]Keio Research Institute at SFC
[§]Graduate School of Media and Governance, Keio University
5322 Endo, Fujisawa-shi, Kanagawa, 252-8520, Japan

E-mail: [†]juno@sfc.keio.ac.jp, [§]ishizaki@sfc.keio.ac.jp

Abstract

Conventional methods for disambiguation problems have been using statistical methods with co-occurrence of words in their contexts. It seems that human-beings assign appropriate word senses to the given ambiguous word in the sentence depending on the words which followed the ambiguous word when they could not disambiguate by using the previous contextual information. In this research, Contextual Dynamic Network Model is developed using the Associative Concept Dictionary which includes semantic relations among concepts/words and the relations can be represented with quantitative distances among them. In this model, an interactive activation method is used to identify a word's meaning on the Contextual Semantic Network where the activation values on the network are calculated using the distances. The proposed method constructs dynamically the Contextual Semantic Network according to the input words sequentially that appear in the sentence including an ambiguous word. Therefore, in this research, after the model calculates the activation values, if there is little difference between the activation values, it reconstructs the network depending on the next words in input sentence. The evaluation of proposed method showed that the accuracy rates are high when Contextual Semantic Network has high density whose node are extended using around the ambiguous word.

1. Introduction

Word sense disambiguation is one of the difficult problems in natural language understanding by computers because it needs contextual meanings. A lot of previous works for such disambiguation have been using co-occurrence of words in their context. Several machine learning algorithms, such as Naive Bayes methods or Support Vector Machine (Murata *et.al.*, 2003), have been used based on co-occurrence information among words. Effectiveness of neural network approaches to the word sense disambiguation has been suggested. Not only the neural network architecture but also large-scale machine readable dictionaries were exploited (Veronis & Ide, 1990).

Many of the Japanese ideographs (Chinese characters) have a few meanings. They should be disambiguated by using their contextual information.

In our previous works, we proposed a Contextual Dynamic Network Model (hereinafter referred to as CDNM), where the Contextual Semantic Network had a structure which changes dynamically depending on each word sequentially in input sentences (Okamoto *et.al.*, 2008). In that model, the contextual semantic network architecture is based on the Associative Concept Dictionary (Okamoto & Ishizaki, 2001) including semantic relation and distance information among the concepts. By using the dynamic network, this method could disambiguate word senses based on words located near the ambiguous words.

However, the network was not rich enough for word sense disambiguation when the beginning of the input words was a homographic ideogram. In this paper, we developed the CDNM to be able to disambiguate word senses by using the neighbour words in the input sentence even if the network is not rich enough.

2. Associative Concept Dictionary

Background knowledge is crucial for computers to understand the contents of the text as well as its syntactic or shallow semantic information from input texts. The Associative Concept Dictionary (hereinafter referred to as ACD) has been built based on the results of large-scale online association experiments, which many subjects can use simultaneously in a campus network at a campus of Keio University (Okamoto & Ishizaki, 2001). In these experiments, the stimulus words were fundamental ones chosen from Japanese elementary school textbooks and were presented to human subjects. The subjects were requested to associate words from the stimulus words with a given set of semantic relations, hypernym, hyponym, part/material, attribute, synonym, action and situation. All of the associated concepts are, in the ACD, connected to the stimulus words with distances calculated by a linear programming method. The distance *D(x,y)* between concepts, *x* and *y*, is shown by the following formula:

$$D(x, y) = 0.81F(x, y) + 0.27S(x, y)\,, \qquad (1)$$

where $F(x,y) = \frac{N_x}{n_{xy}+\delta}$, $\delta = \frac{N_x}{10} - 1$, $(N_x \geq 10)$,

and $S(x,y) = \frac{1}{n_{xy}} \sum_{i=1}^{i=n_{xy}} s_{xyi}$.

N_x denotes a number of the subjects who joined the experiments of stimulus word *x*, and n_{xy} denotes a number of subjects who input the associated word *y* with the same semantic relation for a given stimulus word *x*. Furthermore, δ denotes a factor introduced to limit the maximum value of *F(x, y)* to 10, and s_{xyi} denotes an order of the associated word *y* by a subject *i* for a given stimulus word *x*.

The ACD is built using the quantified distances and is organized in a hierarchical structure in terms of the hypernym and hyponym. Attribute information is used to explain the features of the given word. In the association experiment, each stimulus word had 50 subjects who were students at SFC of Keio University. The number of stimulus words is currently 1100. Total number of associated words is about 280,000. And the number of associated words, when the overlapping words are not counted, is about 64,000 words. In Figure1, "chair" is a stimulus word for the association. "Furniture" is a higher-level concept of "chair". The numbers below <1> express frequencies of subjects who gave a same associated word, <2> an average of order of association and <3> a conceptual distances.

	<1>	<2>	<3>
(chair			
(hypernym	↓	↓	↓
(furniture	0.92	1.02	1.09)
(object	0.04	2.50	7.43))
(hyponym			
(sofa	0.48	1.92	1.96)
(rocking-chair	0.28	1.43	2.64))
(part/material			
(wood	0.60	1.20	1.52))
(attribute			
(hard	0.46	1.17	1.82))
(synonym			
(seat	0.02	1.00	8.37))
(action			
(sit down	0.70	1.03	8.37))
(situation			
(school	0.30	2.40	2.78)))

Figure1 Concept dictionary description for a stimulus word "chair" (a part of associated concepts are presented. The stimulus word and associated words are originally in Japanese)

3. Word Sense Disambiguation by Modified Contextual Dynamic Network Model

Several of Japanese ideographs in the stimulus words in the ACD have a few meanings and different pronunciations. In the association experiments, such ideographs were presented as stimulus words followed with their pronunciations to avoid ambiguities.

A CDNM disambiguates word senses by using an interactive activation method in Contextual Semantic Network (hereinafter referred to as CSN). The network is constructed by using the ACD which includes semantic relations and distance information among the words in the context (Okamoto *et.al.*, 2008). In addition, this network is not a static one but dynamic where the network structure changes depending on the context of the words in the sentences.

In the proposed new model, when the network is not rich enough to disambiguate word senses and when the model cannot decide the appropriate meaning of homographic ideograms, it has a structure where input words are added on the network dynamically and sequentially. By using an interactive activation method, we can assign appropriate word senses to the given ambiguous word by comparing activation values and choosing the best for the homographic ideograms.

3.1 Construction of Contextual Semantic Network (CSN) and its Reconstruction

We can use not only information obtained from word co-occurrence in their context but also that from comparatively rich network with quantitative distances and contextual information for the word sense disambiguation. The following steps show a procedure in detail for this network construction.

- Part of speech information for words (nouns, adjectives, adverbs, verbs and so on) and dependency information in an input sentence is obtained by using dependency structure analysis by using Cabocha, Japanese dependency structure analysis software (Kudo & Matsumoto, 2002; Matsumoto *et.al.*, 1999).
- A CSN is constructed by extracting semantic relations among the words from the ACD by using the information obtained from the dependency structure analysis.
  - When an input word, w_i, is included in the ACD as a stimulus word, a network around the w_i is constructed and added to the CSN. The new network starts from the stimulus word by tracing semantic relation paths until the distance accumulated becomes a certain numerical level.
  - When an input word is a homographic ideograph, we use all stimulus words which correspond to homographic ideogram to construct the network.
  - When a stimulus word is a homographic ideograph, inhibitory links between the stimulus words and the associated words from another homographic ideograph are added in the network
  - Inhibitory links among stimulus words are added in the network when the stimulus words are included in a homographic ideograph.
  - Several links are added in the network based on a practical dependency structure among words corresponding to the input word.
- When the network is reconstructed depending on the next input word, the distances of links are expanded depending on the previous input words.

Let an input sentence be "The picture frame of Picasso's picture dropped from the wall, it struck my head and the forehead bled." In this sentence, "picture-frame" and "forehead" are English expressions which correspond to the homographic ideographs "額" in Japanese. This ideograph has two pronunciations /gaku/ and /hitai/.

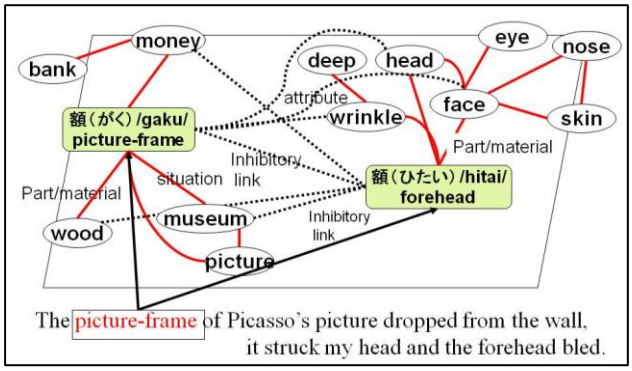

Figure 2 Example of CSN constructed from input word, homographic ideograph "picture frame".

Figure 2 shows an example of a CSN. "Picture-frame" is an input word to construct the network. The square shape nodes are input words in the sentence. The two nodes are those of homological ideographs and are connected with an inhibitory link. Because we cannot assign a word sense to the homographic ideogram, two or more nodes corresponding to the senses of the homographic ideogram are added to the network at the same time. Oval shape nodes are added by using ACD and are connected with the excitatory links. The thick lines connect oval nodes with the word in input sentence. The dotted lines mean inhibitory links. The associated word nodes of a homographic ideograph are connected with other homographic ideographs with inhibitory links. "Museum" is a situation concept of "picture frame". "Picture" and "face" doesn't exist in the sentence but is obtained from ACD.

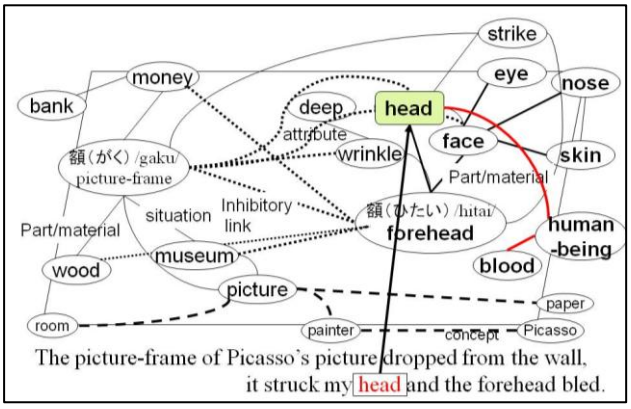

Figure 3 Example of CSN for "head"

Figure 3 is an example of CSN constructed from input word "head". "Human-being" and "blood" are links added in the network.

The broken lines mean excitatory links connected with nodes added from previous input words. The distances of their links are set comparatively long.

3.2 Activation Value Calculation in the Network

The activation value of each node is calculated by an interactive activation model on the CSN. We define the maximum activation level as 1.0. An initial value ($a_i(0)$) and each initial value ($a_i(t)$) at sequential input words are calculated by the following equation.

$$a_i(0) = {(1.0 \times S_{ki})}/{2},$$

$$a_i(t) = {(1.0 \times S_{ki} + a_i(t-1))}/{2}, \quad (2)$$

where 1.0 is a normalization value for S_{ki}, which is a number of node N_i that appears in sentence k. Next, the new activation value ($a_i(t+1)$) of each node N_i at time t+1 is calculated by the following equation (3).

$$a_i(t+1) = a_i(t) - \theta \cdot a_i(t) + \varepsilon_i(t) \ , \quad (3)$$

where the decay parameter θ is assumed to be 0.1 and $\varepsilon_i(t)$ expresses influence of its neighbors at time t. When the neighbors of a node are active, they affect the activation value of the node by excitatory or inhibitory connections, depending on a link between two nodes. Those excitatory and inhibitory influences are combined by a simple equation (4) to yield a net input to the node. Thus, $n_i(t)$ represent the net input to the node by the following the equation.

$$n_i(t) = \sum {a_j(t)}/{\alpha D_{ij}} \ , \quad (4)$$

where $a_j(t)$ denotes an activation value of the node N_j connected with node N_i. α is a constant weight, given by the total number of links of the CSN. D_{ij} denotes a distance between two nodes N_i and N_j. In this paper, The CSN is constructed by tracing semantic relation paths with accumulating the distance before exceeding the value of 5.0. Therefore, the value of D_{ij} with an inhibitory link is assumed to be -5.0. When the net input is excitatory, $n_i(t) > 0$, the effect on the node, $\varepsilon_i(t)$, is given by the following equation.

$$\varepsilon_i(t) = n_i(t)[M - a_i(t)] \ , \quad (5)$$

where M is the maximum activation level of the node and set to 1.0. When the net input is inhibitory, $n_i(t) \leq 0$, the effect of the input on the node is given by the following equation.

$$\varepsilon_i(t) = n_i(t)[a_i(t) - m] \ , \quad (6)$$

where m is the minimum activation level of the node and set to be 0.

3.3 Modified Contextual Dynamic Network Model

It seems that human-beings assign appropriate word senses to the given ambiguous word in the sentence depending on the words which followed the ambiguous word when they could not disambiguate by using the contextual information located before the word. Therefore, in this research, after the model calculates the activation values, if there is little difference between activation values, it reconstructs the network depending on the next word in input sentence.

The system calculates activation values for the two nodes, for example, corresponding to the word senses of the homographic ideogram. When the difference between the two activation values is smaller than a certain threshold, it

continues to add new nodes of the words following the homographic ideogram to the network. The system can choose the best meaning from the additional calculation. Figure 4 shows an example of reconstructed CSN depending on the next input word. The top in the figure 4 shows a case of assigning appropriate word sense. The bottom in the figure 4 shows a case of reconstructing the CSN without assigning it.

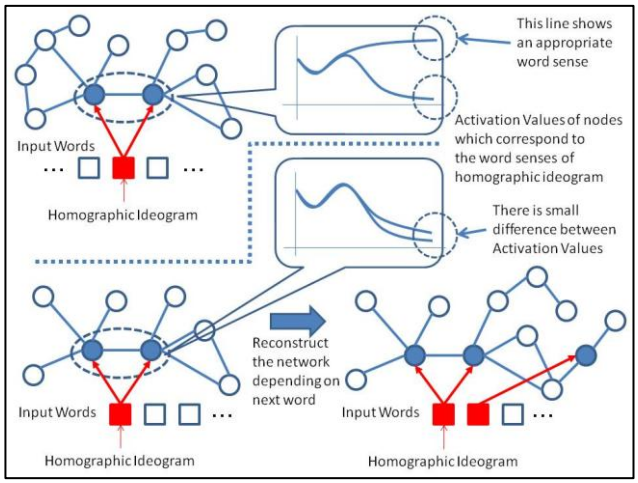

Figure 4 Example of Reconstructed CSN.

4. Experiments for WSD by the Proposed Method

4.1 Simulation for WSD by the Proposed Method

By using the proposed CDNM, we can assign appropriate word senses to the given ambiguous word by comparing some activation value of homographic ideograms. Let an input sentence be "The picture frame of Picasso's picture dropped from the wall, it struck my head, and the forehead bled." In this sentence, "picture frame" and "forehead" are homographic ideographs in Japanese. The word order in the system follows the Japanese language one. At first, we construct CSN based on the ACD for the first input word "wall". Next, the activation value of "wall" is calculated 20 times cycles. "Picasso" is next input word to construct the CSN.

Figure 5 shows activation values of the homographic ideographs in the simulation where each input word spans 20 times cycles. The horizontal axis represents time, and the vertical axis represents activation values. The words in the rectangles are major words selected from the input sentence. The first input word is "wall" which is connected to "picture-frame" in the CSN. The figure shows that activation values of "picture-frame" and "wall" increase slightly at the same time. For the homographic ideogram, "picture-frame" has bigger values than those of "forehead" (see broken-line circle in the figure 5). We can assign its appropriate pronunciation and meaning to "picture-frame" as those of Japanese ideograph "額". The homographic ideogram "forehead" has bigger values than those of "picture frame" (see thick-line circle in the figure 5). We can assign "forehead" to the Japanese ideograph "額".

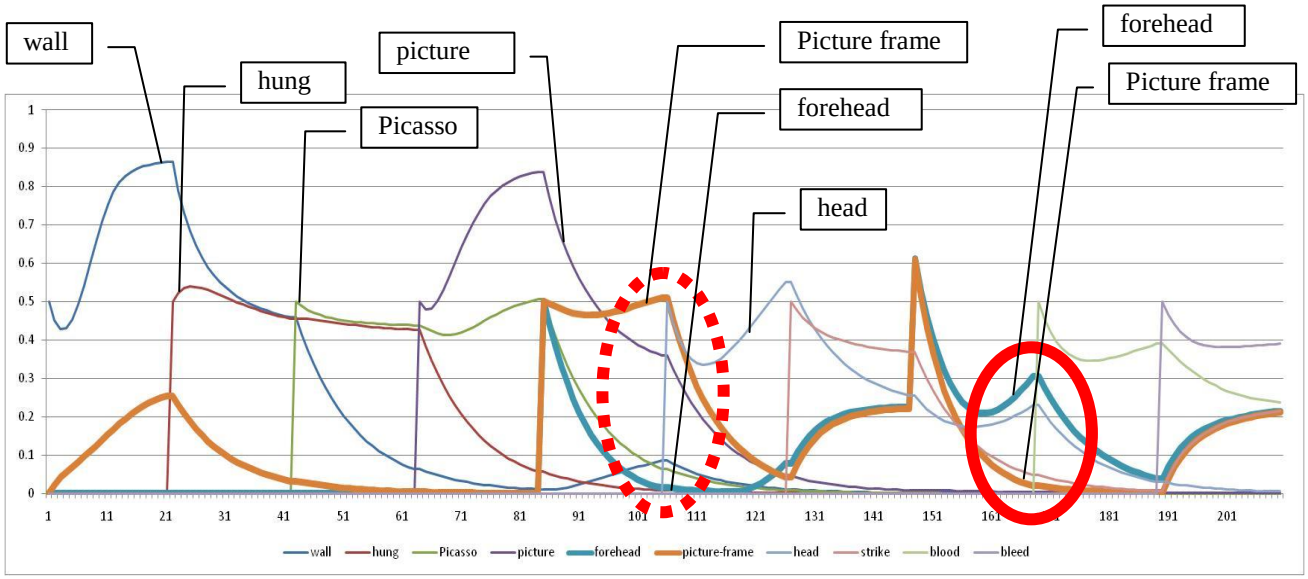

Figure 5 Activation Values for Selected Nodes in Sequential Input Word
The input sentence here is "The picture frame of Picasso's picture dropped from the wall, it struck my head and the forehead bled."

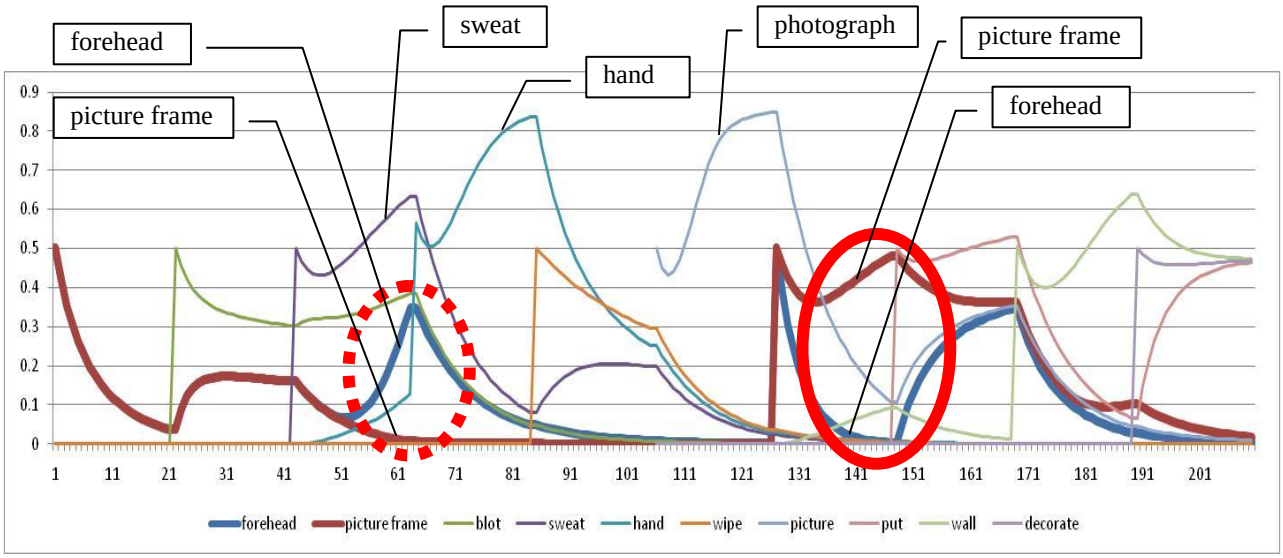

Figure 6 Activation Values for Selected Nodes in Sequential Input Word
The input sentences here are "The sweat that blots to the forehead is wiped. The photograph is put in the picture frame and the wall is decorated with it."

The modified CDNM can improve the accuracy rate of assigning appropriate word senses by using contextual information located before the homographic ideogram, when we sequentially input words in the sentence including a homographic ideogram after several sentences. We provide two sentences that include an ambiguous word whose meanings correspond to those of the ideogram. The system can assign the two appropriate word senses respectively to the input two sentences. When the beginning of the input words is a homographic ideogram and the difference between the two activation values which correspond to the homographic ideogram is smaller than a certain threshold, the CSN is reconstructed depending on the following input words to disambiguate the ideogram.

Let an input sentence be "The sweat that blots to the forehead is wiped. The photograph is put in the picture frame and the wall is decorated with it." The word order in the system follows the Japanese language one.

At first, the system constructs the CSN based on the ACD for the first input word "forehead". The activation value of "forehead" is calculated 20 times cycles. Next, it continues to add new nodes of the word "blot" following the homographic ideogram "forehead" because the difference between the two activation values is smaller than a certain threshold.

Figure 6 shows activation values of the homographic ideographs in the simulation where each input word spans 20 times cycles. The input words are based on two sentences including some homographic ideograms.

For the homographic ideogram in former sentence, "forehead" has bigger values than those of "picture frame" (see broken-line circle in the figure 6). We can assign its appropriate pronunciation and meaning to "forehead" as those of Japanese ideograph "額". For the homographic ideogram in latter sentence, "picture frame" has bigger values than those of "forehead" (see thick-line circle in the figure 6). We can assign "picture frame" to the Japanese ideograph "額".

4.2 Evaluation for WSD by the Previous and Proposed Method

The test data sets, which include homographic ideograms "額", obtained from web site documents are used to evaluate the CNDM. These test sets have 28 sentences which include many stimulus words in ACD as much as possible. We use the CDNM to assign appropriate word senses to the given ambiguous homographic ideogram "額". We can assign the meaning by comparing two activation value of homographic ideogram node (額 /gaku/) and one of its node (額 /hitai/).

	Accuracy rate	Number of sentences which are reconstructed CSN for WSD
Previous CDNM	67.9%	
New CDNM	89.3%	9 sentences

Table 1 The accuracy rate of homographic ideogram "額" in test set

Table 1 shows the accuracy rate of homographic ideogram's correct pronunciation in all the test data. In both CDNM, the activation values of nodes are calculated to each input word in the sentence sequentially. The new CDNM shows higher accuracy than that of the previous method. The new CDNM is reconstructed with CSN for WSD to assign appropriate word senses among 9 sentences.

It seems difficult to assign appropriate ones for this mode when input sentence has comparatively long phrase

modifiers near the homographic ideogram.

5. Comparison of Proposed Method and Conventional Method

5.1 Naive Bayes Method for Word Sense Disambiguation

We use the framework of multinomial Naive Bayes text classification for word sense disambiguation. Let $s_i, (i = 1,2,\cdots,m)$, denote a word sense of a homographic ideogram. S is a set of the word senses. Let $w_j, (j = 1,2,\cdots,n)$, denote words appear in the paragraph. We obtain the optimum word sense s which maximizes $P(s_i \mid w_1,\cdots,w_n)$ by the following functions.

$$\begin{aligned} s &= \arg\max_{s_i \in S} P(s_i \mid w_1,\cdots,w_n) \\ &= \arg\max_{s_i \in S} P(w_1,\cdots,w_n \mid s_i)P(s_i) \end{aligned},$$

where $P(s_i)$ is a number of paragraphs including s_i divided by total number of the paragraphs. We have a Naive Bayes assumption that words surrounding w_j is independent each other:

$$P(w_1,\cdots,w_n \mid s_i) = \prod_{j=1}^{n} P(w_j \mid s_i) .$$

We can determine a probability that it belongs to class s_i by the Bayes' rule:

$$s = \arg\max_{s_i \in S} P(s_i) \prod_{j=1}^{n} P(w_j \mid s_t) .$$

Next, we apply the Jeffreys-Perks law to solve zero frequency problems. Naive Bayes is a simple and effective method in statistical machine learning techniques. Despite of its simplicity, this method is often applied to word sense disambiguation.

5.2 Evaluation Experiments and Results

A pair of training and test data from corpora is used to evaluate the proposed CDNM. The accuracy rate of homographic ideogram's correct pronunciation in all the test data are compared among our model and the Naive Bayes method.

The test data sets, which include homographic ideograms "額" or "札", obtained from web site documents are used to evaluate the CNDM. These data sets include ten words on the left and ten words on the right of the ambiguous words. The number of data sets including an ideograph "額" is 2017 sentences, and the number of data sets including an ideograph "札" is 2018 sentences. The ideograph "札" has two pronunciations /fuda/ and /satsu/ just like the ideograms "額", the former means card or tag and the latter means bank note.

Next, we labelled correct pronunciations for the homographic ideograms in all the training and test data. Pairs of training data (about 98% of data sets) and test data (about 2% of data sets) are used to evaluate the performance of proposed method. The evaluation for a word sense disambiguation is designed to check the effectiveness of CSN which is high density network expanded from the homographic ideogram. Therefore, the test data includes many stimulus words and the associated word from stimulus words which correspond to the homographic ideogram as much as possible.

	CDNM	Naive Bayes
An ideograph "額" pronunciation as /hitai/ and /gaku/	89.3%	86.6%
An ideograph "札" pronunciation as /fuda/ and /satsu/	90.6%	90.6%

Table 2 The comparison of test results among two method

Table 2 shows accuracy rates of disambiguation for "額" or "札" in test data. Our method shows the best score for the two homographs as the correct pronunciations' ratio in all test data of homographic ideogram. The pairs of training and test data are relatively small. Moreover there is a possibility being included for the same sentences as both data because of extracting the sentences including a lot of stimulus words by priority as much as possible. Therefore it seems that the correct answer accuracy of the statistical method is high.

6. Conclusion

In this research, we proposed a method for disambiguation of pronunciations of homographic ideographs as well as the meanings by using the CDNM. When the system can not assign appropriate word sense to given ambiguous homographic ideogram, the CSN is reconstructed depending on the next input words to disambiguate word sense. The evaluation of the proposed method showed that the accuracy rates are high when CSN has high density whose node are extended using around the homographic ideogram. Human-begin can disambiguate word sense based on two words on the left and two words on the right of the ambiguous word (Choueka & Lusignan, 1985). To improve correct answer accuracy, it is necessary to reassign appropriate word sense to the given ambiguous homographic ideogram by using not only contextual information located before the ideogram but also that which follows it if a supposed word sense is found inapposite by using the words which followed the ambiguous word.

7. Future Work

The ideograph "額" has two major senses when its pronunciation is /gaku/. One is an amount of money and the other is a picture frame. It is necessary to carry out more experiments to disambiguate the two meanings. In the association experiments, stimulus words have to be presented unambiguously so that homonyms are presented with their meanings to avoid ambiguities. In the near future, we will conduct such association experiments to assign appropriate word senses with CDNM.

The ACD is a relatively small dictionary. We will extend

it to a large-scale dictionary by extracting concepts from corpora automatically. This extension will be useful for higher level contextual understanding system such as WSD system, document summarization system or analyzer for free-answer questions of the questionnaires.

8. Acknowledgements

We wish to express our gratitude to the students at SFC, Keio University who were very helpful for the association experiments, and also to the members of Ishizaki Laboratory who helped us to construct and modify the Associative Concept Dictionary.

9. References

Choueka, Y. and Lusifnan, S. (1985). Disambiguation by short contexts, *Computers and the humanities*, Vol. 19, pp 147--157.

Kudo,T., Matsumoto, Y., (2002). Japanese Dependency Analyisis using Cascaded Chunking, *CONLL 2002*, pp63-69.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H. and Asahara, M. (1999). Japanese Morphological Analysis System ChaSen Manual version 2.0 Manual 2nd edition (in Japanese)., NAIST Technical Report, *NAIST-IS-TR99009* Nara, Institute of Science and Technology.

McCelland J.L. and Rumelhart D.E.(1981). An Interactive Activation Model of Context Effects in Letter Perception: Part 1. An Account of Basic Findings, *Psychological Rev*., Vol.88, No.5, pp375--407.

Murata, M., Utiyama, M., Uchimoto, K., Ma, Q. and Isahara, H. (2003). CRL at Japanese dictionary based task of SENSAVAL-2 -- Comparison of various types of machine learning methods and features in Japanese word sense disambiguation -- (in Japanese)., *Journal of NLP* , Vol.10 No.3, pp115--133.

Okamoto, J. and Ishizaki, S. (2001). Construction of Associative Concept Dictionary with Distance Information, and Comparison with Electronic Concept Dictionary (in Japanese)., *Journal of NLP* , Vol.8 No.4, pp37--54.

Okamoto, J. and Ishizaki, S. (2007). Word Sense Disambiguation on Contextual Dynamic Network Using Associative Concept Dictionary, *PACLING 2007*, pp.93--100.

Okamoto, J. Uchiyama, K. and Ishizaki, S. (2008). A Contextual Dynamic Network Model for WSD Using Associative Concept Dictionary, *LREC 2008*, .

Veronis, J. and Ide, N. M. (1990). Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries, *Coling '90*., pp389--394.

Waltz, D. L. and Pollack, J. B. (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation, *Cognitive Science*, Vol.9, pp.51--74.