Analysis of Scene Identification Ability of Associative Memory with Pictorial Dictionary

Tatsuhiko TSUNODA *, Hidehiko TANAKA

Tanaka Hidehiko Laboratory, Department of Electrical Engineering
Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan
{tsunoda,tanaka}@mtl.t.u-tokyo.ac.jp

Abstract

Semantic disambiguation depends on a process of defining the appropriate knowledge context. Recent research directions suggest a connectionist approach which use dictionaries, but there remain problems of scale, analysis, and interpretation. Here we focus on word disambiguation as scene selection, based on the Oxford Pictorial English Dictionary. We present a results of a spatial-scene identification ability using our original associative memory, We show both theoretical and experimental analysis, based on a several different measures including information entropy.

1 Introduction

The difficulty of semantic disambiguation in natural language processing originates with the complexity of defining disambiguating knowledge contexts (Barwise J. and Perry J., 1983). These knowledge contexts must provide unique interpretations for co-dependent words, and help resolve "semantic garden path" sequences. For example, in "John shot some bucks,"a unique reading requires semantic agreement on "shot" and "bucks," suggesting either a hunting or gambling context. The semantic garden path can be illustrated by prefixing the above sentence with "John travelled to the woods," which might suggest the hunting context, but then appending "The illegal casino was hidden far from town," to dramatically change the interpretation suggested by the first two sentences.

The core of the problem is the disciplined and dynamic construction of a disambiguating knowledge context. While it might be possible to write static rules which provide disambiguating information in the context of complete knowledge, such rule-based models are both time and space inefficient.

Recognizing these problems, Waltz D.L. and Pollack J.B.(1985) and Cottrell G.W.(1989) proposed a fascinating connectionist approach, which uses early ideas from semantic networks to resolve semantic ambiguity by dynamic spreading activation. This spreading activation construction of disambiguating context is based on a high density associative cognitive model, but still has problems: (1) no automated learning method to adaptively construct the model, (2) non-scalable, and (3) no method of confirming hypothesized disambiguation. Shastri L.(1988) proposes a similar structure, which uses a statistical semantic network. Sharkey N.E.(1989) has proposed a system for processing script-based narratives based on combining local representation and relaxation techniques with parallel distributed learning and mapping mechanisms. Miikkulainen's system DISCERN(Miikkulainen R., 1993) is also suggestive of adaptive processing, and uses self-organizing representation of words and memory depending on semantics. However, all of these models share the problems enumerated above.

Research directions for improvements suggest the use of existing collections of machine-readable dictionaries. Recently, Nishikimi M. et al. (1992) has proposed a new relationship between language acquistion and learning based on scene analaysis. Furthermore, Bookman L.A.(1993) has proposed a scalable architecture for integrating associative and semantic memory using a thesaurus. Based on this idea of using existing sources of word meanings, Veronis and Ide (Veronis J. and Ide N.M., 1990; Ide N.M. and Veronis J., 1993) use several dictionaries and to improve the ratio of words disambiguated to ambiguous words.

In addition to ideas for the source of disambiguating knowledge, many researchers have incorporated some kind of preference heuristics for improving the efficiency of determining disambiguating constraints. Although these methods are essential for semantic processing they lack any coherent method for (1) evaluating performance, and (2) acquiring new disambiguating knowledge from real-world sensors.

Of course all of these problems result from the complexity of defining appropriate disambiguating knowledge contexts. To help control and reduce this complexity, Kohonen T.(1984) has suggested the classification of disambiguating information into four types: (1) spatial contact, (2) temporal contact, (3) similarity, (4) contrast. Kohonen also emphasizes the existence

*Supported by the Fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists

of a contextual background in which primary perceptions occur, but we claim that this kind of information can be expressed in the existing four types.

The previous approaches noted above can all be interpreted as using a complex mixture of the information types proposed by Kohonen. This complexity makes it very difficult to identify or create a stable model of learning the appropriate disambiguating knowledge from the real world.

Our original contribution here is to propose a basic method of word disambiguation based on spatial scene identification, and to provide a detailed analysis of its performance. The disambiguating knowledge is represented in the form of a stochastic associative memory, constructed from the Oxford Pictorial English Dictionary (OPED). This pictorial dictionary claims to provide word sense meanings for most ordinary life scenes. The process of disambiguation is modelled as determining a unique mapping from ambiguous input words to a particular pictorial dictionary scene as modelled in the associative memory. The simple representation of pictorial knowledge based on the OPED makes analysis simpler, and provides a potentially smooth connection to visual sensory data.

2 Scene Identification

In order to identify spatial scenes based on input sentences, some kind of information of defining each scene must exist. As explained in the OPED, "The dictionary is edited regarding the depiction of everyday objects and situations, in order to allow greater scope for the treatment of these objects and situations in the context of English-speaking countries"[from *Forward* in OPED]. Each scene or pictorial entry in the OPED accompanied by a word list of entries from the scene (see next section). This bundle of information is the basis for organizing our associate memory model.

2.1 Constraints

Here we assume some constraints on the method of representing and using the OPED scenes:

- Only ordinal living scenes (384 scenes including thousands of subscenes) are handled. All scenes are hypothesized to be constructable by combinations of these scenes.

- Most of the words in OPED are noun terms accompanied by adjective terms. In this system, spatial-scenes are identified by using only these words. No syntactical information is used.

- Compound words are decomposed into primitive words.

- The associative memory has the ability to incrementally learn, but our analysis here uses a fixed set of scenes and words.

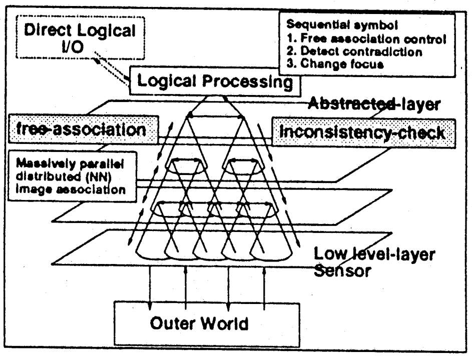

Figure 1: PDAI&CD architecture

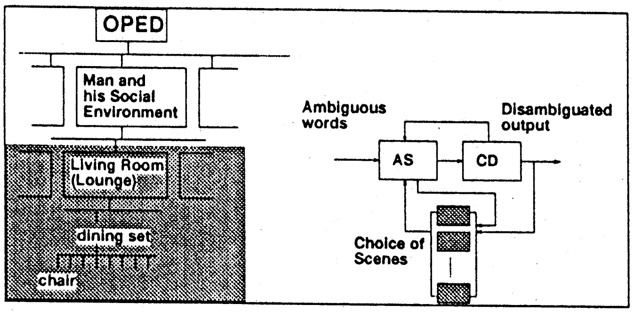

Figure 2: Structure of OPED and diagram of PDAI&CD

- Morphological analysis is done by using the electronic dictionary of Japan Electronic Dictionary Research Institute (EDR).

2.2 PDAI&CD and WAVE

The spatial scene identification system analyzed in this paper is one module of a general inference architecture called Parallel Distributed Associative Inference and Contradiction Detection (PDAI&CD)(Tsunoda T. and Tanaka H., 1993), which uses an associative memory WAVE(Tsunoda T. and Tanaka H.) based on neural networks and a logical verification system. We have previously presented an application of that architecture to semantic disambiguation (Tsunoda T. and Tanaka H., 1993). It features a cognitive model of fast disambiguation depending on context with bottom-up associative memory together with a more precise top-down feedback process (Fig.1). After one scene is selected by previously input words, the system can disambiguate meaning of following words (as in the right side of Fig.2). In the future, we plan to combine natural language processing with visual image from sensory data. Our representation of the spatial data from the OPED is considered to be a simplest approximation of such visual sensory images.

Table 1: Examples of semantic disambiguation

Ex.	Ambiguous word	Sentence # (Context)	Classified scene	Meaning of word
1	ball	(a)	Billiards	globe
		(b)	Carnival	dance
2	lead	(a)	Kitchen	cord
		(b)	Atom I	metal

2.3 Semantic Disambiguation

Words in OPED have different meanings corresponding to their use in different scenes. When a set of ambiguous words uniquely determines a scene, we conclude that the words have been successfully disambiguated. We acknowledge that many other processes may be involved in general word sense disambiguation, but use this scene-selection sense of word sense disambiguation from here on.

We illustrate typical two examples below. The system with OPED and our associative memory can recognize these sentences and classify into each scene in the dictionary. Once a scene is identified, it assigns each ambiguous words uniquely. We call it semantical disambiguation of words here. The correspondances of the sentences and each meaning of word is summarized in Table.1.

1. **ball** :

   (a) Tom shot a white cue ball with a cue. The ball hit a red object ball and he thought it's lucky if it will ...

   (b) Judy found that she was in a strange world. Devils,dominos,pierrots,exotic girls, pirates,... where am I? 'Oh!', she said to herself, as she found she wandered into a ball.

2. **lead** :

   (a) It's not sufficient to shield only by the 1m-thick concrete. The fission experiment requires additional 10cm-thick blocks of lead. Fission fragments released by the chain reaction of...

   (b) He said to his son, "Please pull out the plug of the coffee grinder from the wall socket. Be careful not to pull by the lead. Huum...here I found the kettle."...

Our system is able to disambiguate each meaning in these examples actually.

3 Representation and Processing Theory

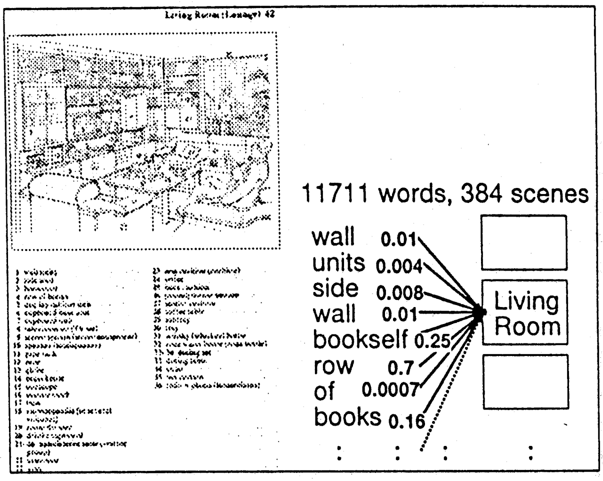

Figure 3: Living room scene and link example on the associative memoryWAVE

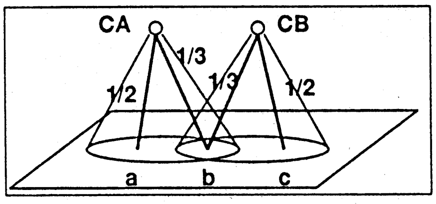

Figure 4: Weight of links and category selection

3.1 Representation of OPED

The Oxford Pictorial English Dictionary(OPED) has very simple form of text and picture (Fig.3). In this example, the upper part is a picture of a living room scene, and the lower part consists of words of corresponding parts as follows:

```
1 wall units
2 side wall
3 bookself
...
```

OPED has originally a hierachical structure of categorization (as in the left side of Fig.2), but we use the middle level of it (shaded part in the figure), which is most easily interpretable.

To provide the associative memory model for processing words and selecting scenes, we encode the OPED entries in the WAVE model as depicted in Fig.3. The weights between scene elements are automatically learned during the construction of the associative memory.

3.2 Simplified Model of Associative Memory WAVE

The aim of using associative memory for identification is to select the most likely scene based on incomplete word data from sentences. $\mathbf{I}_i$ and C_i are set to be elements of input space S_I, scene space S_C, respectively. In an ideal state, the appropriate scene C_i is uniquely indexed by association from a complete input vector: $\mathbf{I}_i \xrightarrow{A} C_i$.

In the typical situation, however, the complete index is not provided and we require a way of ranking competing scenes by defining a weighted activation value which depends on the partial input, or set of ambiguous words, as follows:

$$\begin{aligned} C_i &= f(\sum_j W_{ij} I_j) && (1) \\ f(x) &= \frac{1}{1+e^{-x}} && (2) \\ & && (3) \end{aligned}$$

where the weight of each component is given by the conditional probability value

$$W_{ij} = P(C_i \mid I_j) \qquad (4)$$

A maximum-likelihood scene is selected by a winner-take-all network:

$$C_{i*} = \max_i [C_i] \qquad (5)$$

This type of associative memory has following features:

- Unlike correlative models (Amari S. and Maginu K., 1988), neither distortion of pattern nor pseudo local minimum solutions arise from memorizing other patterns.
- Memory capacity is $O(mn)$ compared to $O(n^2)$ of correlative model, where m is average number of words per scene, and n is the total number of possible words.
- Unlike back-propagation learning algorithms, incremental learning is possible at any time in WAVE.

3.3 Recalling probability and estimation of required quantity of information

The measure of scene selectivity is reduced to the condition whether given words are unique to the scene. If all input words are common to plural scenes, they can not determine the original scene uniquely. For example, the system can not determine whether to choose category CA or CB only by seeing element 'b' in Fig.4. If 'a' or the set {a, b} is given, it is able to select CA. Here we estimate the selectivity by the ratio of successful cases to all of possible cases as follows(n is the number of total elements, k is the number of elements related to each scene, and m is the total number of scenes; incomplete information is defined as a partial vector of elements number s ($0 < s < k$)).

The probability that s elements are shared simultaneously by two patterns is

$$V(n,k,s) = \frac{{}_kC_{s-1}\ {}_{n-k}C_{k-s-1}}{{}_nC_k} \qquad (6)$$

To extend this probability to generalized cases of m patterns, we use the number s of elements of the (partial) input vector. It can be estimated by counting the negative case where more than one pattern shares elements.

$$\begin{aligned} & P(n,k,s,m) && (7) \\ &= (\sum_{r=1}^{s} V(n,k,r))^{m-1} - P(n,k,s-1,m) && (8) \\ &= (p_1 - p_2)(\sum_{q=0}^{m-2} p_1^q p_2^{m-2-q}) && (9) \\ &= V(n,k,s)(\sum_{q=0}^{m-2} p_1^q p_2^{m-2-q}) && (10) \end{aligned}$$

$$\left(p_1 = \sum_{r=1}^{s} V(n,k,r), \quad p_2 = \sum_{r=1}^{s-1} V(n,k,r) \right)$$

The results using this formula are shown in the next section.

3.4 Information Entropy

As an alternative method of evaluation of spatial-scene information of OPED, we consider here self-information entropy and mutual-information entropy along with the information theory of Shannon C.E.(1948).

- **Self-information entropy:**
  Fig.5 illustrates a talking scene. Although sentences involving many ambiguous words are handed from the speaker to the listener, the listener can disambiguate them with some kind of knowkedge common to these people. Conversely, the listner can determine scene by the handed sentences. The entropy of scene selection ambiguity is reduced by the interaction. We can define a concept of self-information (SI) of the spatial-scene idetification module as the entropy of ambiguous words or scenes. Assuming equal probability to the scene selection with no handed word, the entropy of the spatial-scene identification can be calcualted.

  $$SI_0 = -\sum_j P(C_j)\log_2 P(C_j) = \log_2 384 = 8.59 bits$$

  After the identification, the meaning of each word can be selected according to each a selection distribution function updated by the Bayesian rule.

  $$\begin{aligned} SI_1 &= CE(C \mid X) && (11) \\ &= < -\sum_{ji} P_{ji} \log P_{ji} > && (12) \\ P_{ji} &= P(C_j \mid x_i) = P(x_i \mid C_j) && (13) \end{aligned}$$

  Each P_{ij} is equal to W_{ij} as in Eq.(2). <> represents ensemble average over each x_i.

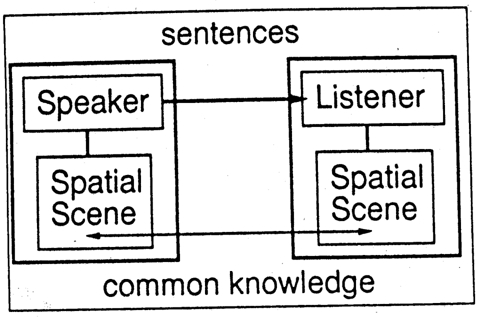

Figure 5: Common knowledge between speaker and listener to disambiguate semantics of handed sentences.

Table 2: Mutual-information of OPED

	Scene entropy	Mutual-inform.
Without input	8.59 bits	—
1 word input	0.80 bits	7.79 bits
2 words input	0.32 bits	0.48 bits

- **Mutual-information entropy:**
  Mutual-information entropy (MIE) can be defined as the contribution of additional words to identify a scene, and consequently, the selectiveness of the target word or scene. In order to select a word meaning or scene from the possible space Y, the space C of all other words are considered in the calculation of conditional entropy (CE). Mutual-information entropy per word is calculated by following formula:

  $$MIE(\theta;\theta') = CE(C \mid \theta) - CE(C \mid \theta')$$

  Here, θ is a set of previous state parameters, and θ' is that of next one. Mutual-inforamtion can be interpreted as the reduction from a previous conditional entropy to corresponding updated conditional entropy with additional words. We provide a theoretical estimation of self-information of spatial-scenes with the dictionary in Table 2. The result suggests that it has the spatial-scene identification ability with a few words preservation. It also supports the consequence of a logical-summation algorithm shown in next section.

4 Analyses of identification module

Here we propose analyses of OPED and results of theoretical simulations. As formula (9) is expensive(11711! times), we use a Monte-Carlo simulation to abstract its characteristics. Iteration time in each case is 1,000.

- Fig.6 (a) shows a distribution of number of elements involved in each scene in OPED. It approximated a Gaussian distribution and has a average value of 184.2. This value is used in the theoretical simulations.

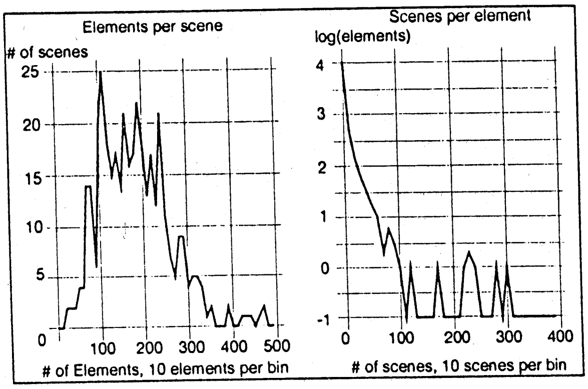

Figure 6: (a) Distribution of number of elements per scene and (b) Distribution of number of scenes per elements

- Fig.6 (b) shows a distribution of number of scenes which are related to one element. The region where more than 100 scenes are related to one word are those for trivial words like 'a', 'the', 'of', 'that', 'to', 'in', 'and', 'for', 'with', 's'. Although we could ignore these words for an actual application, we use them for fairness.

- Selection probability in the case that partial words of scenes are input to the associative memory is illustrated in Fig.7. The recall rate increases as the input vector (set of words) becomes more similar to complete vector (set of words) pattern. Only about five words are enough to identify each scene at recognition rate of 90 percent. Compared to the average number of 184 words in each scene, this required number is sufficiently small. It proves good performance of the associative memory used in this module. Theoretical results of a random distribution model is also shown in Fig.7. The cause of the discrepancy between the experiment and theoryis described later. The dotted line 'EXACT' in the figure is a result using logical-summation. The crossing point of the 'OPED' line and the 'EXACT' line is remarkable. The former has the advantage of expecting with relatively high-probability (likelihood) using input words of small number. Though with more additional words, the algorithm is defeated by the simple logical-summation. As our architecture PDAI&CD uses dual-phase of expectation and evaluation, we can get a solution with maximum-likelihood satisfying constraints automatically.

- Fig.8 shows the distribution of number of elements contributing to identify each scene uniquely.

- In order to clarify the discrepancy of the experimental and theoretical results, the number of elements overlapped in any two scenes are counted.

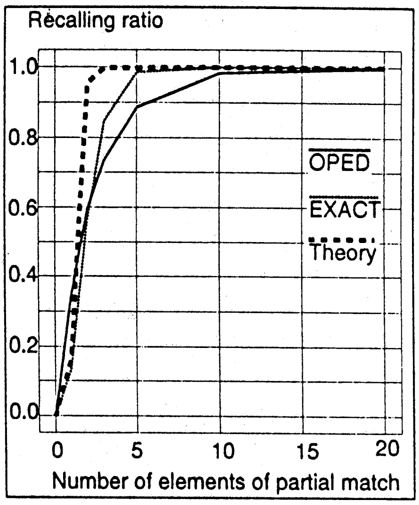

Figure 7: Recalling probability to number of partial input elements

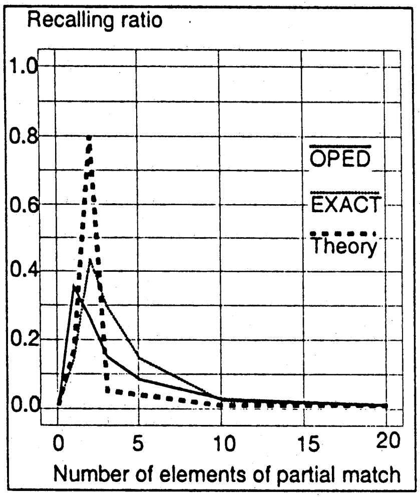

Figure 8: Distribution of number of partial input elements to identify scenes

As in Fig.9, the number of overlapping elements in the theoretical calculation is very small compared to the experiments with OPED. OPED-2 in the figure illustrates the same value without using trivial words like 'a', 'the', 'of', 'that', 'to', 'in', 'and', 'for', 'with', 's'. But the existence of these words can not explain the whole discrepancy. This will be described in the next section in more detail.

- As further investigation in order to explain the discrepancy of 'EXACT'(logical-summation) and 'OPED'(with our associative memory), distribution of weight values is shown in Fig.10. Logical-summation method is achieved by a special algorithm similar to the associative memory. Only the difference is that it uses equal weight value without any variance. But in practical, the experimental result of 'OPED' as in Fig.10 shows an existence of enormous variance in the distribution of weight value. Though the variance helps the selectivity with a few words, it disturbs the expectivity with more than three words conversely. Here we summarize the interpretation of the gaps among the theoretical expectation, the result of logical-summation('EXACT'), and the system('OPED'):

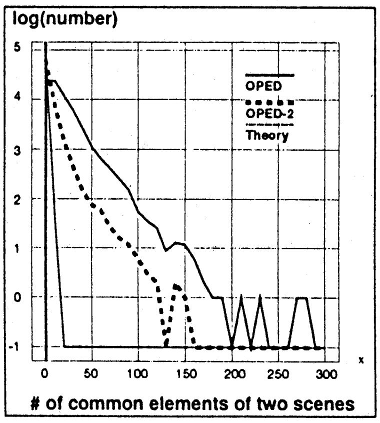

Figure 9: Distribution of number of elements common to two scenes

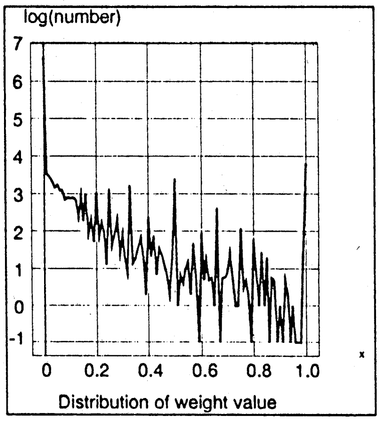

Figure 10: Distribution of weight value

1. Exsistence of trivial words in most of the scenes.
2. Variance of weight distribution.
3. Difference of characteristics between algorithms.

- Abstracted results are summarized in Table.3. In this table, the number of registered words in dictionary itself is different from the number of the total words analyzed by our system. The discrepancy arises mainly from the fact that we analyzed compound words into simple words (e.g. 'research laboratory' to 'research' and 'laboratory').

Table 3: Summarized results

Total # of scenes	384 scenes
Registered # of words	27,500 words
Total # of words	11,711 words
Average # of words / scene	184.2 words
Max # of words in one scene	478 words
Required # of words to identify scenes at 90% ratio	5 words
Required # of words to identify scenes at 90% ratio by exact match algorithm	4 words
Theoretical estimation of required # of words to identify scenes at 90% ratio	2 words

5 Summary

We analyzed the selectivity of our 384 living scenes with many sets of words which are part of 11,711 words used in the dictionary OPED. The average number of words in one scene is about 184. The probability of recalling correct scenes with input partial words is different from the theoretical simulation of random assignment constructed with values of these parameters. Unlike random generation of arbitrary symbols, semantics of natural language consists of highly-correlated meanings of words. Although the theoretical simulation of the simplified model suggests a rough estimation of disambiguation requirements we should analyze the dictionary itself as in this paper.

Another suggestive analysis is using Shannon's information or entropy, which gives us more accurate information depending on probability of each phenomenon. It shows how to estimate the amount of semantic ambiguity.

Spatial-scene identification is one of the simplest kind of context necessary to disambiguate meaning of words and offer a new method for future integration of natural language processing and visual pattern recognition.

6 Acknowledgements

The authors acknowledge Randy Goebel, Nancy Ide, Jean Veronis, Hiroaki Kitano, Koiichi Hashida, Katashi Nagao and Lawrence A. Bookman for helpful discussions and suggestions. Also the authors thank Kazuhiro Nakadai and Satoshi Murakami for transformation of the pictorial dictionary into machine-readable one. This research is supported by Fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists and Grant-in-Aid for Scientific Research on Priority Areas by the Ministry of Educations, Science and Culture, Japan.

References

[1] Amari S. and Maginu K. (1988). Statistical Neurodynamics of Associative Memory. *Neural Networks, Vol. 1-1*, pp.63-73.

[2] Barwise J. and Perry J. (1983). *Situation and Attitudes*, MIT-Press.

[3] Bookman L.A. (1993). A Scalable Architecture for Integrating Associative and Semantic Memory. *Connection Science, Vol. 5.*

[4] Cottrell G.W. (1989). *A Connectionist Approach to Word Sense Disambiguation*, Pitman, Morgan Kaufmann Pub.

[5] Ide N.M. and Veronis J. (1993). Extracting Knowledge Bases from Machine-Readable Dictionaries: Have We Wasted Our Time? In *KB & KS 93*, pp.257-266.

[6] Kohonen T. (1984). *Self-Organization and Associative Memory*, Springer-Verlag.

[7] Miikkulainen R. (1993). *Subsymbolic Natural Language Processing : An Integrated Model of Scripts, Lexicon, and Memory.*, MIT-Press.

[8] Nishikimi M., Nakashima H. and Matsubara H. (1992). Language Acquisition as Learning. In *Proceedings of COLING-92*, pp.707-713.

[9] Shannon C.E. (1948). A Mathematical Theory of Communication. *Bell System Tech. J., Vol.27*, pp.373-423, 623-656.

[10] Sharkey N.E. (1989). A PDP Learning Approach to Naural Language Understanding. In Alexander I. Ed., *Neural Computing Architectures : The Design of Brain-like Machines*, MIT-Press, pp.92-116.

[11] Shastri L. (1988). *Semantic Networks: An Evidential Formalization and its Connectionist Realization*, Morgan Kaufmann.

[12] Tsunoda T. and Tanaka H. (1992). Semantic Ambiguity Resolution by Parallel Distributed Associative Inference and Contradiction Detection. In *Proceedings of IJCNN-Nagoya93, Vol.1*, pp.163-166.

[13] Tsunoda T. and Tanaka H. (1993). Winner Associative Voting Engine (WAVE). In *Proceedings of IJCNN-Beijing92, Vol.3*, pp.589-594.

[14] Veronis J. and Ide N.M. (1990). Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. In *Proceedings of COLING-90*, pp.389-394.

[15] 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.