Meaning and the Mental Lexicon

Will Lowe*
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
Scotland
wil@cogsci.ed.ac.uk

Abstract

This paper presents a network model of the mental lexicon and its formation. Models of word meaning typically postulate a network of nodes with connection strengths, or distances, that reflect semantic similarity, but seldom explain how the network is formed or how it could be represented in the brain. The model presented here is an attempt to address these questions. The network organizes semantically similar words into clusters when exposed to sequentially presented text. Lexical co-occurrence information is calculated and used to create a hierarchical semantic representation. The output is similar to semantic networks first described by [Collins and Loftus, 1975], but is created automatically.

1 Introduction

The mental lexicon refers to the representations that allow word recognition on the basis of auditory and visual stimuli. The lexicon is understood as two linked levels of representation: The first level consists of form-based representations that reflect a word's phonological or graphemie properties. The second *level* contains semantic representations that reflect its meaning relations with other words [Marslen-Wilson, 1989],

Priming studies are an important source of evidence for the semantic organization of the lexicon. When subjects are presented briefly with a letter string, followed by another, and asked to decide whether the latter is a real word, the response time when both strings are related is reliably faster than when they are unrelated. Priming effects can be found using stimuli that are graphemically, morphologically, or semantically related ([Taft, 1991] for a review).

"The author is supported by a Medical Research Council studentship.

Substantial progress has been made modelling the form-based lexical representations in the light of graphemie or phonological similarity [Plaut *et* a/., 1994], but there is currently no principled measure of semantic similarity. Word meaning is much more difficult to quantify.

The network described here is an attempt to address this problem. It is inspired by two relatively independent approaches to semantic representation from cognitive psychology and computational linguistics. After considering each approach I describe the network's implementation and present results. The next section describes the structure and development of semantic representations, and how the model relates to previous work. Finally 1 consider the model's psychological relevance with reference to developing categorizations and semantic priming.

2 Lexical-semantic Networks

A highly influential theory of lexical-semantic representation from cognitive psychology is based on the semantic network. A semantic network consists of a set of nodes and connections of varying strengths, or lengths, between them [Collins and Loftus, 1975]. Each concept is assigned a node, and connection strengths reflect the amount of conceptual relevance each node has to its partner. The stronger, or shorter, connections represent a high level of similarity. Weaker, or longer, connections hold between less related nodes. In a lexical-semantic network (LSN), each node represents a word and the distance between nodes reflects the amount of semantic similarity between each word. The Logogen model [Morton, 1979], Interactive-activation model and spreading activation accounts, are all types of LSN [Neely, 1991].

LSN accounts explain semantic priming effects in the following way: Each node has an activation level. When a stimulus is presented it activates all nodes in the network to some degree. If one node is activated strongly enough its activation will pass a threshold and fire. The stimulus will be recognized as that word. Each*time a

word is presented, activation spreads from the most activated node to nearby nodes, decaying over time. For example, if'doctor' is presented shortly before 'nurse', the node associated with 'nurse' will reach threshold faster and fire sooner. Its resting activation level is raised by activation spreading from 'doctor' during the interstimulus interval.

3 Data-intensive Semantics

Recent work in computational linguistics suggests that large amounts of semantic information can be extracted automatically from large text corpora on the basis of lexical co-occurrence information [Lund *et* al., 1995; Schiitze, 1993]. This approach is particularly well suited to neural network implementation [Finch, 1993] because co-occurrence statistics track conditional probabilities, and neural networks have straightforward interpretations as statistical models [Bishop, 1995].

The data-intensive approach to semantics is consistent with, and inspired by theories of meaning that emphasize the importance of use [Wittgenstein, 1958] (see also [Church and Mercer, 1993]). Lexical co-occurrence information reflects a word's distributional profile, which is a reflection of its use.

The success of the data-intensive semantics research shows that, with a large enough sample, there is sufficient information in a strictly linguistic environment to recover much semantic structure. It seems plausible, therefore, to investigate the possibility that the brain makes use of such information. The recent discovery that semantic and associative priming effects in the lexical decision task are significantly correlated with co-occurrence statistics [Lund *et al,* 1995; Spence and Owens, 1990] support this possibility. Lund et al. constructed a high-dimensional space on the basis of lexical co-occurrence counts. Words that were close together in the space gave larger priming effects than those further away.

4 Modelling the Lexicon

LSN theories provide an intuitive way to understand word meaning and its relation to priming. However, there is no theory of how the nodes of a network are formed, or how the distance (or strength) relations between them become organized.

The data-intensive approach to semantics is an effective predictor of semantic priming, and reflects an influential approach to understanding word meaning. However, the approach requires an extremely high-dimensional co-occurrence space for lexical-semantic representation. It is not obvious how such a space could be represented in the brain.

The model presented below is a first attempt at explaining how the semantic level of the lexicon could be organized, consistent with the LSN and data-intensive semantics approaches, in a way that is computationally tractable and biologically reasonable.

4.1 Overview of the Model

The model consists of an input layer that picks out words from a text stream, a dynamic proto-lexicon which records co-occurrences between the present target word and words either side of it, and a self-organizing map. The proto-lexicon is initially empty and the self-organizing map weights are set to random values.

4.2 Implementation

Proto-lexicon

The proto-lexicon represents each word in terms of the number of times it has been seen to co-occur directly before and after each other word in the vocabulary. Specifically, in an n—word vocabulary each word $(1 \leq i \leq n)$ is associated with the vector $x^i = [\xi_1, \ldots, \xi_{2n}]$ normalized to unit length, where $\xi_j\ (1 \leq j \leq n)$ denotes the frequency with which W_j has preceded w_i, before t, and $\xi_k\ (n+1 \leq k \leq 2n)$ denotes the frequency with which W_k has succeeded w_i,. Thus at each time step, x^i represents the model's best guess for the conditional probabilities

$$p(w_j(t-1) \mid w_i(t)) \tag{1}$$

and

$$p(w_k(t+1) \mid w_i(t)) \tag{2}$$

for all words j, *k* and time *t.* Each successive x^i is an improved estimation of the true distributional profile of each word.

In large-scale applications it is usual to distinguish a fixed subset of high-frequency words to serve as context (see [Church and Mercer, 1993] for a review). Co-occurrence vectors calculated using high-frequency words are less sparse and provide better samples. This technique complicates the relation between the co-occurrence vectors and quantities (1) and (2), though the results are robust to approximation. The model is presented without approximation.

Self-organizing Map

The self-organizing map is presented with the current proto-lexical representation for each word as it is encountered in the text stream. The winning unit is the unit *i** with weight vector w_i*such that

$$\| x^w - w_{i*} \| < \| x^w - w_i \| \tag{3}$$

for all $i \neq i*$. Output unit weights are updated after each word presentation using a variation of Kohonen's self-organizing map algorithm [Kohonen, 1982]:

$$\Delta w_{ij} = \eta\, \Phi(i, i^*)(\xi_j - w_{ij}) \tag{4}$$

where η is the learning rate and $\Phi(i, i^*)$ a neighbourhood function. i represents the unit to be updated, and i^* the winning unit. $\Phi(i, i^*)$ takes the value 1 when $i = i^*$ and decreases as a function of the distance between i and i^* according to:

$$\Phi(i, i^*) = e^{-\frac{\|i - i^*\|^2}{2\sigma^2}} \quad (5)$$

During the course of training σ and η are slowly reduced [1].

Input

Input consists of words taken sequentially from a 20,000 word corpus generated by a stochastic context-free grammar described in [Elman, 1990]. A sample section of input is shown below.

man like boy lion eat mouse

Each time a word is recognized in the text stream, its representation in the proto-lexicon is updated and presented to the self-organizing net as an training item.

4.3 Results

After moving through the corpus once, clusters of semantically related words emerge. Each word is presented to the network the winning output unit is recorded and labelled. Figure 1 shows the winning unit for each word after 20,000 word presentations.

Consistent with the LSN approach, the network clusters each word with other words that are used in similar contexts. Similar words tend to be nearer to one another than to dissimilar words. Verbs have been represented together on the right side: Psychological verbs 'see' and 'smell' are represented together, as are destructive verbs 'smash' and 'break' within the main verb group. On the left side, categorial similarity among the nouns is equally well preserved - human and animal nouns group separately, adjacent to one another.

However, figure 1 does not reflect the full extent of the net's categorization. 'Man' is equidistant from 'boy' and from 'book', but is related much more closely to one than the other. This fact is represented by the network, not in the pattern of winning units, but by the pattern of activation across all output units. Figures 2,3 and 4 show activation plots for 'man', and for 'boy' and 'book' with the unit specialized for 'man' marked. 'Boy' is associated much more strongly with 'man', than with 'book' because 'boy' is the highest unit on a plateau containing all the human nouns, whereas 'book' is in a separate region shared by inanimate nouns.

Figures 5 and 6 also show how the network creates a hierarchical semantic representation: 'See' and 'smell'

[1] $\sigma(t) = \sigma_i(\frac{\sigma_f}{\sigma_i})^{\frac{t}{t_{max}}}$ where σ_i represents the initial setting and σ_f the final value. $\eta(t) = \eta_i(\frac{\eta_f}{\eta_i})^{\frac{t}{t_{max}}}$ where η_i and eta_f represent the initial and final learning rates, respectively. For these simulations $\sigma_i = 20, \sigma_f = 0.5, \eta_i = 1$ and $\eta_f = 0.01$.

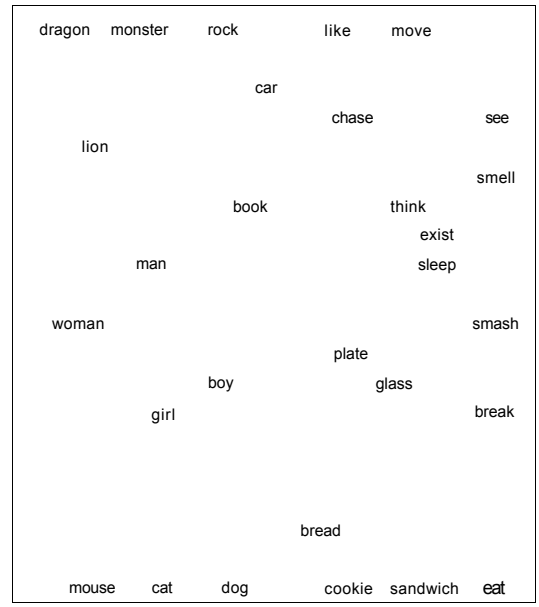

Figure 1: Output map after 20,000 word presentations.

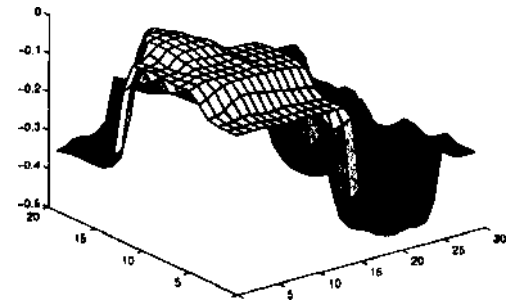

Figure 2: Activation plot for the word 'man'.

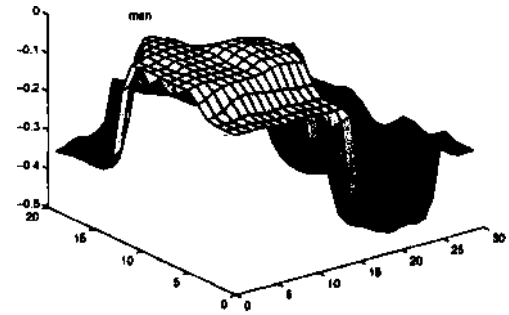

Figure 3: Activation plot for the word 'boy' with 'man' marked.

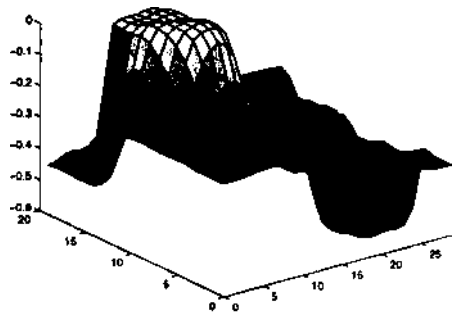

Figure 4: Activation plot for the word 'book' with 'man' marked.

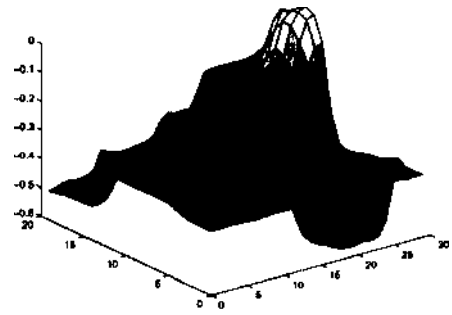

Figure 5: Activation plot for the psychological verb 'smell'.

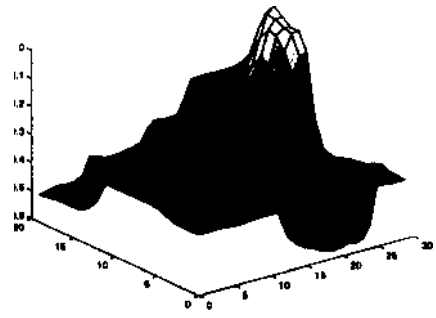

Figure 6: Activation plot for the psychological verb 'see'.

share an activated region but this region is a section of the larger region that covers all verbs.

5 Structure and Development of Semantic Representation

Structure

Although figure 1 resembles a classical semantic net (without connections), unlike many other LSN models, it uses a distributed coding scheme; the activation levels of all units are used to express semantic relationships among words. Distributed coding is both more reliable and more biologically realistic than classical localist coding. It is more reliable because representations need not be compromised by the loss of single units. In topographic maps, if a unit specialized to some input pattern is lost, one of its neighbours will become the winner for that pattern on subsequent tests, but the shape of the output map will remain largely intact due to graded transitions in feature-specificity across neighbouring units. Graded transitions between cell response profiles due to topographic organization have been reported in many brain regions [Knudsen e*t* a/., 1988].

In the model graded transitions between winning units also allow uncommitted units to capture new words with distributions similar to more than one word in the initial vocabulary.

The data-intensive approach to semantics explains semantic similarity in terms of points in a high-dimensional space. By the end of training, the proto-lexical representations define such a space. Here 29 words create a 58-dimensional space. In order to form an output representation of the type required by the LSN approach, and to explain how such a space could be represented in the more limited dimensions of the brain, this space must be reduced to a more manageable size. Two properties of the self-organizing map algorithm make it especially well-suited to this task:

1. The self-organizing map algorithm creates a nonlinear projection from a collection of data points in a high dimensional space defined by each input vector to a one or two-dimensional grid of output units. The projection attempts to preserve the topology of the input space in the lower dimensional output space. Thus the algorithm performs precisely the data-reduction necessary.
2. The algorithm has a straightforward physiological interpretation: it models the development of feature selectivity due to lateral inhibition among cortical nerve cells [Sirosh and Miikkulainen, 1993; Kohonen, 1993]

It is possible to pinpoint cell groups relevant to naming and semantic memory tasks using electrode mapping techniques during neurosurgery [Ojemann, 1983;

1991]. In a review of language localization studies, [Bradshaw and Mattingley, 1995] conclude that lexical representation probably depends on many areas distributed across the cerebral cortex. This is consistent with the structure of the model.

Development

During the course of training, output units become increasingly sensitive to a particular input pattern. The weight update equation ensures that the weight vector of the winning unit and its neighbours are moved toward the input pattern. However the input patterns are also moving. Representations in the proto-lexicon alter each time a new word is read. This means that different words get different amounts of training devoted to them. During development, the region over which updates occur is shrinking. This entails that input patterns that are far from one another in the input space are separated first on the output map, and patterns that are very close to each other are separated later [Kohonen, 1982]. If the algorithm is interrupted before completion, the final output representation provides a usable partial categorization of the input data (See also section 7).

6 Related Work

The self-organizing map algorithm has been used successfully to model semantic clustering [Miikkulainen, 1993], but Miikkulainen's networks require a static set of input patterns to be constructed by a separate extended-backpropagation system with theta-role assignments, before clustering can begin. In contrast, this system operates with raw text, without staged processing, and without supervised components. [Ritter and Kohonen, 1989] have also used self-organizing maps, but used heavily pre-processed input data, and a much smaller corpus.

Elman's Simple Recurrent Networks [Elman, 1990] also make use of of co-occurrence statistics, represented in the hidden node activations. However, in the Simple Recurrent Network, hidden unit response profiles must be manually extracted and submitted to cluster analysis before semantic similarity information becomes directly available: In the model presented here co-occurrence information is explicit.

7 Psychological Relevance

The model presented here illuminates two psychological phenomena: learning to distinguish between semantically similar words, and semantic priming effects.

Discrimination

The shape of the output map reflects the network's increasing ability to distinguish words that are semantically distinct. The semantic difference between 'mouse' and 'sandwich' is greater than between 'mouse' and 'cat'. The developmental schedule of the network reflects this, by learning to distinguish 'mouse' from 'sandwich' *before* distinguishing it from 'cat'. While 'cat' remains undistinguished the two words are essentially the same to the system. The discriminative capacities of the system also depend upon exposure; high frequency words are distinguished sooner in general. These observations are consistent with research into child language acquisition, suggesting that broad semantic distinctions between frequent items are discovered first [Harris, 1992]. [Finch and Chater, 1994] have shown how partial categorizations may be used to bootstrap more complex representational structure during language development.

Semantic Priming

The network described above is a type of LSN. Consequently it is possible to formulate an account of semantic priming within this framework. With the current architecture each word that is encountered activates the output layer independently, via its proto-lexical representation. To allow an explanation of priming it is necessary to relax this restriction and assume that word recognition advances all unit activations to the levels particular to the recognized word in small amounts over a brief time period, and that unit activations then decay over time until the activation surface is flat.

The explanation of semantic priming is then straightforward: If a prime word is presented before activation levels have fully decayed, then residual activation will still be present in some units. For example, let 'man' be the target word, and let the related word 'boy' and the unrelated 'book' be primes. 'Man' (fig. 2) will take longer to be recognized as a word when preceded by 'book' (fig. 4) than when it is preceded by 'boy' (fig. 3) because the activation surface for 'man' is almost identical to that of 'boy', but quite different from the activation surface for 'book'; fewer increments are necessary to convert the activation surface for 'boy' into the surface for 'man' than to convert the surface for 'book' into the surface for 'man'.

This account of priming is similar to other LSN models. Priming effects depend on the distance between the prime and target words because the the map is organized such that activation tends to drop off with distance from the winning node.

8 Conclusion

The network presented here models the formation and arrangement of the semantic level of the mental lexicon. It is consistent with the lexical-semantic network approaches to lexical arrangement and semantic priming, and with the data-intensive approach to semantics. The network represents semantically similar words together using lexical co-occurrence information that is

calculated as the network moves through a text corpus. The network uses a topographic mapping technique that is widespread in the brain, and provides a biologically reasonable account of mental lexicon.

Acknowledgments

I would like to thank my supervisors Richard Shillcock and Mark Ellison, and Joanna Bryson for encouragement and much useful discussion.

References

[Bishop, 1995] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[Bradshaw and Mattingley, 1995] J. L. Bradshaw and J. B. Mattingiey. *Clinical Neuropsychology: Behavioral and Brain Science.* Academic Press, 1995.

[Church and Mercer, 1993] K. W. Church and R. L. Mercer. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics,* 19:1-24, 1993.

[Collins and Loftus, 1975] A. M. Collins and E. F. Loftus. A spreading activation theory of semantic processing. *Psychological Review,* 82:407-428, 1975.

[Elman, 1990] J. L. Elman. Finding structure in time. *Cognitive Science,* 14:179-211, 1990.

[Finch and Chater, 1994] S. Finch and N. Chater. Distributional bootstrapping: From word class to protosentence. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society,* 1994.

[Finch, 1993] S. Finch. *Finding Structure in Language.* PhD thesis, Centre for Cognitive Science, Edinburgh University, 1993.

[Harris, 1992] M. Harris. *Language Experience and Early Language Development.* Lawrence Erlbaum Associates, 1992.

[Knudsen *et al,* 1988] E. 1. Knudsen, S. du Lac, and S. D. Esterly. Computational maps in the brain. *Annual Review of Neuroscience,* 10:41-65, 1988.

[Kohonen, 1982] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics,* 43:59-69, 1982.

[Kohonen, 1993] T. Kohonen. Physiological interpretation of the self-organizing map algorithm. *Neural Computation,* 6:895-905, 1993.

[Lund *et ai,* 1995] K. Lund, C. Burgess, and R. A. Atchley. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society,* pages 660-665, 1995.

[Marslen-Wilson, 1989] W. Marslen-Wilson. Access and integration: Projecting sound onto meaning. In W. Marslen-Wilson, editor, *Lexical Processing and Representation,* chapter 1, pages 3-24. MIT Press, 1989.

[Miikkulainen, 1993] R. Miikkulainen. *Subsymbolic Natural Language Processing.* MIT Press, 1993.

[Morton, 1979] J. Morton. Word recognition. In J. Morton and J. C. Marshall, editors, *Psycholinguistics Series 2: Structures and Processes.* Elek, 1979.

[Neely, 1991] J. H. Neely. Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner and G. W. Humphreys, editors, *Basic Processes in Reading: Visual Word Recognition,* chapter 9. LAwrence Erlbaum Associates, 1991.

[Ojemann, 1983] G. A. Ojemann. Brain organization for language from the perspective of electrical stimulation mapping. *Behavioral and Brain Science,* 6:189-230, 1983.

[Ojemann, 199l] G. A. Ojemann. Cortical organization of language. *Journal of Neuroscience,* 11(8):2281—2287, 1991.

[Plaut *et ai,* 1994] D. C. Plaut, J. L. McClelland, M. S. Seidenberg, and K. E. Patterson. Understanding normal and impaired word-reading: Computational principles in quasi-regular domains. Technical report, Carnegie Mellon University, 1994.

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

[Schiitze, 1993] H. Schiitze. Word space. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems,* volume 5, pages 895-902. Morgan Kaufmann, 1993.

[Sirosh and Miikkulainen, 1993] J. Sirosh and R. Miikkulainen. How lateral interaction develops a self-organizing feature map. In *Proceedings of the IEEE International Conference on Neural Networks,* 1993.

[Spence and Owens, 1990] D. P. Spence and K. C. Owens. Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research,* 19:317-330, 1990.

[Taft, 1991] M. Taft. *Reading and the Mental Lexicon.* Lawrence Erlbaum Associates, 1991.

[Wittgenstein, 1958] L. Wittgenstein. *Philosophical Investigations.* Blackwell, 1958.

Extracting Propositions from Trained Neural Networks

Hiroshi Tsukimoto
Research & Development Center, Toshiba Corporation
70, Yanagi-cho, Saiwai-ku, Kawasaki, 210
Japan

Abstract

This paper presents an algorithm for extracting propositions from trained neural networks. The algorithm is a decompositional approach which can be applied to any neural network whose output function is monotone such as sigmoid function. Therefore, the algorithm can be applied to multi-layer neural networks, recurrent neural networks and so on. The algorithm does not depend on training methods. The algorithm is polynomial in computational complexity. The basic idea is that the units of neural networks are approximated by Boolean functions. But the computational complexity of the approximation is exponential, so a polynomial algorithm is presented. The authors have applied the algorithm to several problems to extract understandable and accurate propositions. This paper shows the results for *votes* data and *mushroom* data. The algorithm is extended to the continuous domain, where extracted propositions are continuous Boolean functions. Roughly speaking, the representation by continuous Boolean functions means the representation using conjunction, disjunction, direct proportion and reverse proportion. This paper shows the results for *iris* data.

1 Introduction

Extracting rules or propositions from trained neural networks is important[1], [6], Although several algorithms have been proposed by Shavlik, Ishikawa and others [2],[3], every algorithm is subject to problems in that it is applicable only to certain types of networks or to certain training methods.

This paper presents an algorithm for extracting propositions from trained neural networks. The algorithm is a decompositional approach which can be applied to any neural network whose output function is monotone such as sigmoid function. Therefore, the algorithm can be applied to multi-layer neural networks, recurrent neural networks and so on. The algorithm does not depend on training methods, although some other methods[2], [3] do. The algorithm does not modify the training results, although some other methods [2] do. Extracted propositions are Boolean functions. The algorithm is polynomial in computational complexity.

The basic idea is that the units of neural networks are approximated by Boolean functions. But the computational complexity of the approximation is exponential, so a polynomial algorithm is presented. The basic idea of reducing the computational complexity to a polynomial is that only low order terms are generated, that is, high order terms are neglected. Because high order terms are not informative, the approximation by low order terms is accurate[4].

In order to obtain accurate propositions, when the hidden units of neural networks are approximated to Boolean functions, the distances between the units and the functions are not measured in the whole domain, but in the domain of learning data. In order to obtain simple propositions, only the weight parameters whose absolute values are big are used.

The authors have applied the algorithm to several problems to extract understandable and accurate propositions. This paper shows the results for *votes* data and *mushroom* data.

The algorithm is extended to the continuous domain, where extracted propositions are continuous Boolean functions. Roughly speaking, the representation by continuous Boolean functions means the representation using conjunction, disjunction, direct proportion and reverse proportion. This paper shows the results for *iris* data.

Section 2 explains the basic method. Section 3 presents a polynomial algorithm. Section 4 describes the experiments. Section 5 extends the algorithm to continuous domains and applies it to *iris* data.

The following notations are used. *x,y,..* stand for variables. f,g,.. stand for functions.

2 The basic method

There are two kinds of domains, that is, discrete domains and continuous domains. The discrete domains can be reduced to {0,1} domains by dummy variables. So only {0,1} domains have to be discussed. Here, the domain is {0,1}. Continuous domains will be discussed later.