Artificial Intelligence Review (1989) **3**, 23–31

Connectionist systems for natural language understanding

B. Selman
Department of Computer Science, University of Toronto, Toronto, Canada M5S 1A4

Abstract We will discuss various connectionist schemes for natural language understanding (NLU). In principle, massively parallel processing schemes, such as connectionist networks, are well-suited for modelling highly integrated forms of processing. The connectionist approach towards natural language processing is motivated by the belief that a NLU system should process knowledge from many different sources, e.g. semantic, syntactic, and pragmatic, in just this sort of integrated manner. The successful use of spreading activation for various disambiguation tasks in natural language processing models lead to the first connectionist NLU systems. In addition to describing in detail a connectionist disambiguation system, we will also discuss proposed connectionist approaches towards parsing and case role assignment. This paper is intended to introduce the reader to some of the basic ideas behind the connectionist approach to NLU. We will also suggest some directions for future research.

Introduction

There has been much discussion on the extent to which different forms of processing should be done in a parallel, integrated manner in natural language understanding (NLU) systems. Most conventional NLU systems follow a model where syntactic processing functions as the front end to the system; thereby ensuring that syntactic and semantic processing are strictly separated. However, this separation is frequently counterproductive and a more integrated form of processing is needed, especially for disambiguation tasks (Hirst, 1983).

One way of achieving a more integrated form of processing is by adding a marker passing component[1] which runs in parallel with the syntactic and semantic components (Charniak, 1983). The connectionist approach towards NLU takes this paradigm of integration a step further in allowing many sources of knowledge, such as syntax, semantics and pragmatics, to be handled in a highly integrated manner. We will describe the basic idea behind this approach as illustrated by an example of a connectionist network for word-sense disambiguation, and then proceed to describe some of the connectionist work on other aspects of NLU, such as parsing and case role assignment. Finally, we will conclude with a short discussion on open research issues.

Connectionist word-sense disambiguation systems

Before we discuss a connectionist disambiguation system, we will give a short description of connectionist models in general.

A typical connectionist model or network (e.g. Feldman & Ballard, 1982) consists of a large number of simple computing *units*. Each unit has a number of *inputs* and one *output*, which is in turn connected to zero or more other units. Each connection has a *weight* associated with it, and each unit is assigned a numerical value called its *activation level* (which is often restricted to be one of a fixed set of values). Without loss of generality, we can assume that the output of a unit is equal to its activation level. A unit updates its activation level with each time step; its new activation level is a function of its previous activation and the sum of the weighted outputs from other units connected to it. All units update their activation level in parallel. When a connection has a positive weight, it is called an *excitatory* connection, otherwise it is called an *inhibitory* connection. If the output of unit A is linked to unit B by an excitatory link, an increase in the activation level of unit A will tend to increase the activation level of unit B. If the units are linked by an inhibitory link, an increase in activation level of unit A will tend to decrease that of unit B.

An important aspect of any natural language processing system is its ability to disambiguate. Consider, for example, lexical ambiguity; Gentner (1982) found that the 20 most frequent nouns have an average of 7.3 senses each, and the 20 most frequent verbs have an average of 12.4 senses each. Yet, people perform word-sense disambiguation quite effortlessly.

The first connectionist models for NLU dealt with word-sense disambiguation. These models were motivated by two factors: firstly the belief that the disambiguation requires integrated, parallel processing of knowledge from various sources, and secondly the fact that spreading activation in connectionist networks can be used to model *semantic priming*. Semantic priming occurs when activation of concept structures in the brain reduces the reaction time for subsequent judgements involving associated concepts (Collings & Loftus, 1975). Psycholinguistic research shows that semantic priming speeds up word-sense disambiguation.[2]

Cottrell and Small (1983) use a representation scheme similar to that of Schank (1975) in his work on conceptual dependency theory. They distinguish three levels[3] in their model: a lexical level, a word-sense level, and a case logic level. At the lexical level, the input to the system, incoming words, activates the associated word units. This provides input to the word-sense level, where units representing the different senses of the words at the lexical level become activated. Finally, at the case logic level, units expressing the possible relationships between the predicates and objects at the second level are activated. An example network will clarify their approach. Consider the following input sentence.

A man threw up a ball.

Figure 1(a) shows the state of the network after receiving '*a man threw*' as input. The number of plus signs represents the level of activation; a unit is said to be

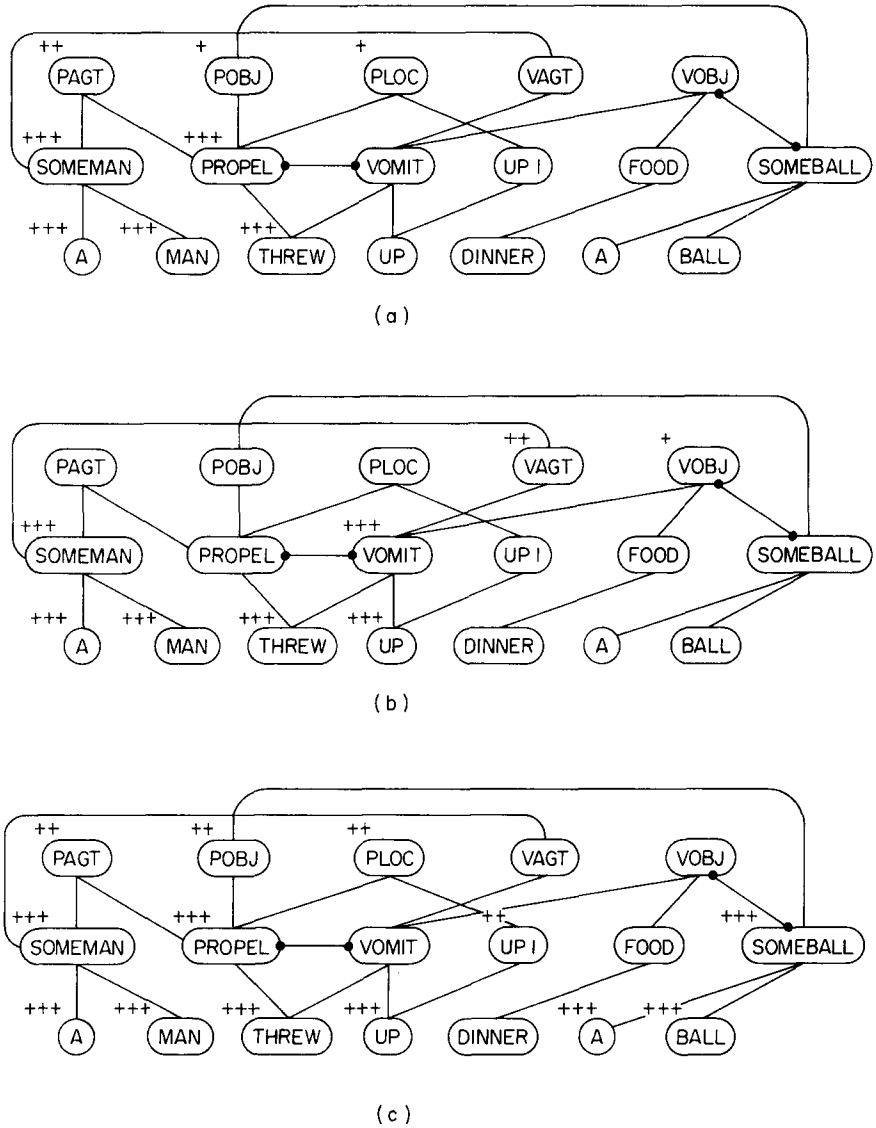

Fig. 1 The states of a network by Cottrell & Small (1983) for different inputs. Links denote excitatory connections, except for those ending in (•—•) which denote inhibitacy connections. (Figure adapted from Small *et al.*, 1982.)

active if it has one or more plus signs.[4] Note that the links are symmetric, so activation can flow in either direction between linked units.

The first phrase in the input, '*a man*', activates the units A and MAN. These units activate the unit SOMEMAN. The article '*a*' will also excite other units, for example, SOMEWOMAN (not shown in Fig. 1), but an inhibitory link between SOMEMAN and SOMEWOMAN prevents them from being simultaneously active (because no person can be both).

The unit THREW excites PROPEL and VOMIT, but the VOMIT unit needs additional input to become active. Thus only PROPEL is activated. Although not explicitly shown in Fig. 1, the weight on the link between THREW and VOMIT is smaller than that between THREW and PROPEL; this represents the fact that '*a man threw*' is more commonly associated with throwing an object than vomiting. Moreover, the inhibitory link between PROPEL and VOMIT will prevent them from both being active. Finally, the units SOMEMAN and PROPEL will activate PAGT, indicating that the unit SOMEMAN represents the agent of propel.

The pattern of activity will change drastically when the word '*up*' is presented as the next input; Fig. 1(b) shows this new pattern. The units THREW and UP together activate VOMIT. Now, the unit VOMIT will become more active than PROPEL, and their inhibitory link will cause the activation of PROPEL to decrease. The system will now settle into a new stable state, representing the fact that the phrase '*threw up*' usually refers to vomiting.

Finally, the system will receive the input '*a ball*', which reinforces the unit POBJ and inhibits the analogous VOBJ. The system will now settle into the stable pattern shown in Fig. 1(c). This pattern represents the preferred interpretation of the complete sentence.

The networks by Cottrell & Small (1983) do not explicitly represent the syntactic structure of the sentence, as can be seen in Fig. 1. However, in word-sense disambiguation tasks syntax often plays an important role, for example, in resolving noun-verb ambiguities. Waltz & Pollack (1985) propose a connectionist parser that incorporates an explicit representation of syntactic structure. Since the basic idea behind their approach is quite similar to that of Cottrell and Small, we will only give a high-level description of their disambiguation networks.

Waltz & Pollack distinguish four levels in their networks: an input level, a syntactic level, a lexical level, and a contextual level. A network is custom built for each input sentence. The sentence is run through a conventional chart-parser to construct the syntactic layer of the network representing all possible parses of the input. (Aside from handling lexical ambiguities, Waltz and Pollack's system also resolves some syntactic ambiguities.) The lexical layer is a representation of all possible senses of the words. The contextual layer represents the context in which the sentence should be interpreted.

In order to construct the context layer, Waltz & Pollack (1985) propose a connectionist scheme for contextual priming. The network contains a set of units representing concepts and a set of units representing *microfeatures* (Hinton, 1981) of these concepts. Each concept is connected to a representative set of microfeatures; for example WEEKEND will be connected to DAY and WEEK, but also to RESTAURANT, BAR, and so forth. Each concept will have at least one microfeature in common with another concept, and most concepts share many microfeatures with other concepts. Therefore, in general, all units will be indirectly connected to one another.

In such a scheme the activation of a specific concept spreads throughout the network thereby activating related concepts. The level of activation of a particular unit represents the strength of its relation to the priming concept. Thus, the current con-

text is given by the total pattern of activity in the network at that moment. When new sentences are introduced this pattern of activity may change, representing a change in context.

Other aspects of natural language understanding

The connectionist networks discussed in the previous section, although quite successful at disambiguation tasks, can each handle only a small number of input sentences. To build larger networks, we require a more principled way of designing connectionist networks for NLU. For example, we need a systematic method for setting the weights on the connections.

In an effort to obtain general, provably correct networks, research thus far had to be limited to specific subtasks of natural language processing, such as parsing and case role assignment. Eventually, networks for the various tasks must be integrated in order to obtain a general connectionist natural-language-processing system.

Parsing

Fanty (1985) and Selman & Hirst (1985; Selman, 1985) have proposed general connectionist parsing networks. These networks have several features in common: both deal with context-free parsing, result in provably correct networks, and only deal with sentences up to some maximum length.[5] The schemes differ in their underlying architecture: Fanty's parser employs a deterministic weight update rule (Feldman & Ballard, 1982), while Selman and Hirst use a stochastic update scheme similar to that of the Boltzmann machine (Fahlman, *et al.*, 1983) and apply simulated annealing (Kirkpatrick, *et al.*, 1983). To illustrate the general idea behind connectionist parsing, we will consider in some detail the scheme proposed by Selman & Hirst (1985).

Selman and Hirst (1985) propose a connectionist network in which the grammar rules are captured using a localist representation. That is, each syntactic category is represented by a *single* computing unit, called a *main unit*. Each context-free grammar rule is represented by a group of main units called a *connectionist primitive*. Figure 2 shows two examples of such primitives. The output of each unit is either $+1$ (unit is *active*) or -1 (unit is *inactive*). The activation of all units of a connectionist primitive in a particular state of the parsing network corresponds to the use of the associated grammar rule in the parse.

The primitives are linked together by *binder units*.[6] The units are linked such that all possible parse trees for all sentences up to a fixed length can be represented in the network. Since parse trees can have common substructures, such sub-

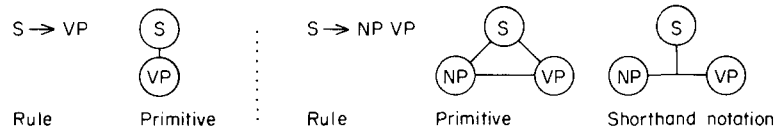

Fig. 2 Two examples of connectionist primitives and their associated grammar rules. In this case all links denote excitatory connections. (From Selman & Hirst, 1985).

structures are shared in the network, greatly reducing the total number of units needed. Finally, the set of main units representing the terminal symbols of the grammar forms the input layer of the network. An incoming sentence will activate a subset of those units.

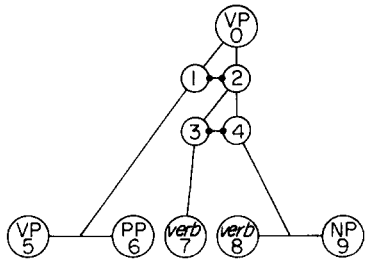

Fig. 3 Various connectionist primitives linked together: excitatory link (—), inhibitory link (•—•), main unit (◯), binder unit (○). (From Selman & Hirst, 1985.)

Figure 3 shows an example of part of a parsing network. The following rules are represented:

VP	→	*VP PP*	(a)
VP	→	*verb*	(b)
VP	→	*verb NP*	(c)

These grammar rules allow the verb phrase in the input sentence, represented by unit No. 0, to be parsed in three possible ways. The binder units (No. 1, No. 2, No. 3 and No. 4) are linked by inhibitory and excitatory connections, such that when the network reaches a stable state, the active binders (those with output +1) tell us which one of the three grammar rules was used in the parse. So, if binder No. 1 stays active, rule (a) is used in the parse; if No. 2 and No. 3 stay active, rule (b) is used; and if No. 2 and No. 4 stay active rule (c) is used.

All units in the network employ the same stochastic updating rule, similar to the one used in the Boltzmann machine. After a sentence is presented to the network, thereby activating various units in the input layer, the network follows a predetermined annealing schedule and gradually settles into a state in which the active units represent a correct parse of the input, provided the input is a sentence of the language generated by the grammar. During annealing the value of the 'temperature' parameter is gradually decreased. This influences the probability that the output of a unit is changed. The simulation is started off at a high temperature: units will change state with a probability of 0.5. As the temperature is lowered the network becomes more 'rigid'; finally, at a temperature of 0 the update function becomes deterministic. Simulated annealing has been proposed as a good search technique for finding minima of multivariable functions such as the energy function discussed below.[7]

So far, we have not discussed how the weights are set. As noted above, for a general connectionist approach to an NLU task we do not want to choose weights by trial and error. One of the main advantages of the Boltzmann machine archi-

tecture is that its computation can be characterized as a search for a global minimum in the *energy function* of the network.[8] The energy function maps each possible state of the network into a real number, called the energy of the state; a different weight setting will lead to a different energy function. So, to obtain a parsing network, the weights must be set such that states corresponding to correct parses have minimum energy. This can be achieved based on the local environment of each unit (a slightly modified update rule that facilitates the design of a provably correct network is introduced in Selman, 1985).

In Selman and Hirst's system, parsing is treated as a constraint satisfaction problem in which the constraints are given by the grammar rules. The search for a minimum energy state corresponds to a search for a parse tree which is the best possible match given the constraints and the input. When presented with an ungrammatical sentence, the network will settle into a minimum energy state which corresponds to the best possible partial parse of the sentence. When presented with a syntactically ambiguous sentence, the network will settle into a state representing one of the possible parses. This is consistent with the human tendency to settle for one possible consistent interpretation when faced with ambiguous input.

Case role assignment

McClelland & Kawamoto (1986) propose a connectionist system for case role assignment that takes into account both word order and semantic constraints. Given a sentence the network will establish 'who did what to whom'. The structure of their model is too complicated to be described in detail here. Their system learns from previous experience and generalizes to unseen sentences.

They employ a distributed representation in which each word is represented by a set of semantic microfeatures, each corresponding to a single unit in the network. Similarly, various roles, such as Agent and Patient, are each represented by a group of units. This representation facilitates the learning of semantic associations between words and allows for generalization during learning.

Interestingly, the weights on the connections in the network are not preset. Instead, the network is presented with a set of examples, each consisting of a sentence (of a rather retricted syntactic form) and its case role assignment. After each example the weights are adjusted according to the perceptron learning procedure (Rosenblatt, 1962). Following this training period, when presented with a sentence the network will settle into a state representing the correct case role assignment. Moreover, the network correctly handles some sentence which it has never seen before, i.e. the learning mechanism generalizes from the examples. Thus, we have another principled way of setting the weights of the network, in this case by learning from examples.

Conclusions

We have discussed various connectionist systems for NLU. Disambiguation networks process syntactic and semantic knowledge in a highly integrated manner.

However, these networks only handle a few input sentences each. A much better understanding of the design of such networks is needed in order to develop more general, larger scale systems.

The work on parsing and case role assignment is a step in the direction of more general networks. Although, these approaches are quite interesting in the themselves, they will need to be integrated into a complete, connectionist natural language processing system.

Waltz & Pollack (1985) and McClelland & Kawamoto (1986) have quite successfully used distributed representations in their schemes. However, more research is needed in the area of the distributed representation of complex, articulate structures where the meaning of the whole is determined by the meaning of its parts, and where each part itself has a complex structure (Hinton, 1988).

Finally, we should also mention in passing that various connectionist NLU systems, e.g. Cottrell (1985), conform quite well with psycholinguistic constraints. These models show that the connectionist approach also offers promise with respect to psychological models for linguistics.

Notes

[1] A marker passing mechanism sends markers along links in a network-based knowledge representation scheme (Fahlman, 1979).

[2] Marker passing, the discrete form of spreading activation, has been used to implement semantic priming in conventional NLU systems (Hirst, 1983).

[3] In connectionist models it is common to group units into one or more *levels*.

[4] The level of activation is a continuous value, so, for example, one plus sign refers to a value between k and 2k, where k is a positive number.

[5] The sentence length restriction is in some sense an unavoidable aspect of the connectionist approach (Charniak & Santos, 1987); see Pollack (1987) for a proposed means of dynamically changing the network.

[6] Fanty (1985) uses units with a similar function in his parser.

[7] For a quite different approach towards parsing using simulated annealing see Sampson (1986).

[8] Feldman (1985) evaluates various connectionist models based on energy minimization.

Acknowledgments

I would like to thank Sue Becker, Jim des Rivières, Maya Guha, Graeme Hirst, and Niels Lobo for useful comments and discussions.

References

Charniak, E. (1983) Passing markers: A theory of contextual influence in language comprehension. *Cognitive Science*, **7**, July–September 1983, 171–190.

Charniak, E. & Santos, E. (1987). A connectionist context-free parser which is not context-free, but then it is not really connectionist either. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, Seattle, WA, July 1987, 70–77.

Collins, A.M. & Loftus, E.F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, November 1975, **82**, 407–429.

Cottrell, G.W. (1985). *A connectionist approach to word sense disambiguation*. Doctoral dissertation, Computer Science Department, University of Rochester, Rochester, NY 14627, April 1985.

Cottrell G.W. & Small, S.L. (1983). A connectionist scheme for modelling word sense disambiguation. *Cognition and Brain Theory*, **6**(1), 1983, 89–120.

Fahlman, S.E. (1979). *NETL: A system for representing and using real-world knowledge*. MIT Press. Cambridge, Massachusetts, USA.

Fahlman, S.E., Hinton, G.E. & Sejnowski, T.J. (1983). Massively parallel architectures for AI: NETL, Thistle and Boltzmann machines. *Proceedings of the National Conference on Artificial Intelligence*, Washington, August 1983, 109–113.

Fanty, M. (1985). *Context-Free Parsing in Connectionist Networks*. Technical Report 174, Computer Science Department, University of Rochester, Rochester NY, November 1985.

Feldman, J.A. (1985). *Energy and the Behavior of Connectionist Models*. Technical Report 155, Computer Science Department, University of Rochester, Rochester NY, November 1985.

Feldman, J.A. & Ballard, D.H. (1982). Connectionist models and their properties. *Cognitive Science*, **6**, 205–254.

Gentner, D. (1982). Some interesting differences between nouns and verbs. *Cognition and Brain Theory*, **4**, 155–184.

Hinton, G.E. (1981). Implementing semantic networks in parallel hardware. In *Parallel Models of Associative Memory*, (eds.) G.E. Hinton and J.A. Anderson, Erlbaum, Hillsdale, NJ, U.S.A.

Hinton, G.E. (1988). Representing part-whole hierarchies in connectionist networks. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, Montreal, Canada, 48–54.

Hirst, G. (1983), Semantic interpretation against ambiguity. PhD thesis, Department of Computer Science of the Brown University, December 1983. Appeared as *Semantic interpretation and the resolution of ambiguity* (Studies in natural language processing), Cambridge University Press, 1987.

Kirkpatrick, S., Gelatt, C.D. Jr & Vecchi, M.P. (1983). Optimization by simulated annealing, *Science*, **220**, 4598, 671–680.

McClelland, J.L. & Kawamoto, A.H. (1986). Mechanisms of sentence processing: assigning roles to constituents of sentences. In *Parallel Distributed Processing* by D.E. Rumelhart, J.L. McClelland and the PDP Research group, vol. 2, Bradford/MIT Press, Cambridge, USA, 273–325.

Pollack, J.B. (1987). Cascade back-propagation on dynamic connectionist networks. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society, Seattle*, WA, July 1987, 391–404.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan, New York.

Sampson, G. (1986). A stochastic approach to parsing. *Proceedings of Coling '86*, 1986, 151–155.

Schank, R.C. (1975). *Conceptual Information Processing*. North-Holland publishing company, Amsterdam.

Selman, B. (1985). *Rule-based Processing in a Connectionist System for Natural Language Understanding*. Technical Report CSRI-168, Computer Systems Research Group, University of Toronto, April 1985.

Selman, B. & Hirst, G. (1985). A Rule-Based Connectionist Parsing System, Proceedings of the Seventh Annual Conference of the Cognitive Science Society, Irvine, CA, August 1985, 212–219. An extended version entitled 'Parsing as an Energy Minimization Problem' appeared in *Genetic Algorithms and Simulated Annealing* (ed.) Lawrence Davis, Pitman, London. 155–168.

Small, S.L., Cottrell, G. & Shastri, L. (1982). Towards Connectionist Parsing. *Proceedings of the National Conference on Artificial Intelligence*, Pittsburgh, PA, August 1982, 247–250.

Waltz, D.L. & Pollack, J.B. (1985). Massively parallel parsing. *Cognitive Science*, **9**, 1985, 51–74.