A Microfeature Based Approach Towards Metaphor Interpretation

Ron Sun
Department of Computer Science
The University of Alabama
Tuscaloosa, AL 35487, USA

Abstract

This paper advocates a microfeature based approach towards developing computational models for metaphor interpretation. It is argued that the existing models based on semantic networks and mappings of complex symbolic structures are insufficient and inappropriate for modeling metaphors. A connectionist model of metaphor interpretation based on microfeatures is presented, which tries to take into account some important issues, such as accurate capturing of similarity, automatic formation of features, contextual effects, elimination of long paths in conceptual hierarchies, salience imbalance, and feature enhancement. Some of these issues have broad implications in cognitive modeling.

1 Introduction

Metaphor is an important cognitive phenomenon, and it is of great interest to AI, philosophy, psychology, linguistics, and literary studies. In fact there has been a surge of interests in the past few decades in the philosophy and linguistics communities in the nature and the process of metaphor interpretation; more recently, there are also increasingly more interests in the AI community in non-literal language and analogical reasoning, in both of which metaphor occupies a prominent place. Among the voluminous studies of metaphor, computational models are relatively few. Most existing computational models of metaphor are based on semantic networks and mappings of complex symbolic structures, or in other words, based entirely on the traditional symbolic AI methods. I would like to argue that such symbolic models are insufficient and inappropriate for modeling metaphor in general, on the basis of a number of important considerations. I will instead propose a microfeature based (connectionist) approach towards developing computational models for metaphor interpretation.

2 Semantic Network Based Approaches

Most existing models of metaphor interpretation take semantic network based approaches: they represent the requisite linguistic (and world) knowledge in some kinds of semantic networks with hierarchically structured concepts, and metaphor interpretation is accomplished through traversing and mapping the hierarchical conceptual structures in some ways. Similarities of words and other linguistic entities are computed from structural relations among entities within a hierarchy. Such models (especially, for example, Martin 1988 and Fass 1991) achieved certain successes within limited domains and/or strictly controlled environments. However, there are many valid questions and objections to these models.

Let us look into some representative existing models. Martin (1988) presents a system for dealing with *conventional* metaphors, i.e., metaphors that reflect a core set of correspondences that can manifest in various ways (Lakoff and Johnson 1980). For example, "How can I *enter* Emacs?" or "I am *in* Lisp", where a computer program is uniformly viewed as an enclosure. His system consists of large semantic networks with conceptual hierarchies and a set of procedures that operate on them. When a verb (e.g. *enter*) is found to be incompatible with (violating some constraints of) the object (e.g. *Emacs*), a procedure is called to find a coherent mapping between the domain of the verb and the domain of the object, from a core set of conventional metaphors stored in the system; once such a mapping is found, the verb in the source domain is mapped into a corresponding concept in the target domain (i.e., a concept suitable for the object, e.g. *invoke* Emacs). New variations of known conventional metaphors can also be handled, through detecting their similarity to existing ones by finding a path through the conceptual hierarchy between the new instance and the known instances. For example, once the system knows "I am in Lisp", it can also make sense of "I am in Emacs" by finding the relation between Lisp and Emacs in the hierarchy (the sibling relation in this case, since they are both computer programs).

Veale and Keane (1992) deal also with conventional metaphors. They divide the interpretation process into two steps: first a scaffold of core (semantic network) structures is constructed in accordance with some conventional metaphors which is already known and stored in the system; then the structure is fleshed out with details, from general world knowledge and/or the context. What attributes to transfer from the source domain (e.g. physical enclosures) to the target domain (e.g. computer programs) and their respective pre-conditions for valid

transfers are specified a priori.

Fass (1991) views metaphor as the existence of constraint violation as well as a set of correspondences between a source domain and a target domain. The representation is based on a kind of semantic network, where graph search algorithms find different types of paths between different concepts represented in "semantic vectors", which include constraints and preferences that are used to keep mappings coherent.

The shortcomings of these models are as follows:

- In semantic networks, conceptual hierarchies require hand-coding, In any domain of realistic sizes, this poses a serious problem for practical reasons — the difficulty of knowledge acquisition in system building.
- A more serious problem is that such semantic networks are often made up of ad hoc concepts and hierarchies, that is, hierarchies specifically designed for one particular kind of circumstance or only for the purpose of getting one particular result (in the current context, for generating plausible interpretations of test examples of metaphors).
- A related problem is the fixedness of such hierarchies. Conceptual hierarchies are explicitly and manually constructed a priori. However, in cognition, at least some concepts (if not all) can be more flexible; that is, they can have one or the other superordinate concept, depending on (1) the current context (e.g., contextual priming; cf. Barsalou 1989), (2) the current goals (cf. Schank 1977), and (3) even personal, idiosyncratic connections (i.e., individual differences; cf. Barsalou 1989). Thus a more flexible representation is required.

  It may be argued that this problem can be remedied by introducing some new components to a semantic network that can modify the connections on the fly in accordance with various factors. Although this is theoretical possible, semantic networks are notoriously complex and difficult to modify, even statically, let alone dynamically. Therefore, it seems that some fundamentally different approaches are called for.
- With semantic networks, it may also be difficult to find certain information, and complex search algorithms are required. Such algorithms typically trace links in various ways. One ramification of this is that time required to process the relations between two concepts is proportional to the length(s) of the path(s) between the two concepts, which does not seem plausible cognitively. This is especially troublsome, considering the ad hoc-ness of the conceptual hierarchies as mentioned before.

There are some more specific issues that are directly related to metaphor interpretation:

- Similarity in these models, which is necessary for determining the relation between two concepts, is determined mainly from the sibling relation between the concepts involved. This may not be flexible enough. For example, whales are closer in a conceptual hierarchy to bears (both can be represented as siblings, with mammals as their parent), but in a metaphor whales may be closer to sharks instead, based on some of their attributes prominent in the context. In other words, similarities in metaphor may not reflect, or be restricted to, conceptual hierarchical structures, especially when such structures are artificially and inflexibly constructed.
- Another way of capturing similarities is to have explicit links in semantic networks that directly indicate pair-wise similarities (cf. Eskridge 1993). The problem with such an approach is the complexity of representation: everything is somehow similar to everything else in some ways, and therefore there may need to be pair-wise connections between (almost) every pair of concepts. The number of similarity links is thus $2 * C_n^2 = n * (n-1)$, where n is the total number of concepts. This is much too complex representationally, and is even more so dynamically (for search algorithms to go through).
- Most importantly, metaphor cannot be simplistically construed as *constraint violation, structural mapping,* or *selective inference* — there is more to it. In interpreting metaphors, certain subtle senses are brought up from metaphoric interaction (Black 1955). Such senses are not apparent from mapping structures across domains; consequently they are not dealt with in many existing models. An example will help to illustrate the issue: in Martin's model, the metaphoric use of the word *enter* as in "How can I enter Emacs?" is mapped to *invoke* based on the structural correspondence of [agent of entering = agent of invoking] and [the place to enter = the program to invoke]. However, what is missing is the sense inherent in the metaphor of getting into an enclosure of some sort, and therefore the model also misses the sense that an agent is now able to access everything inside but is insulated somewhat from everything outside. These senses are strong in the word *enter*, but not in the word *invoke*. Through metaphor, a new sense is added (transferred) to the target domain, from the interaction with the source domain.

  An even more illustrative example is from Ortony (1989): "Billboards are warts." In this metaphor, one of the prominent features in the source domain, the ugliness of warts, is transferred to the target domain to achieve the effect of "billboards are as ugly as warts". And in this process, the ugliness of warts is becoming even more prominent in the source domain.

3 A Microfeature Based Approach

Now I want to argue that a radically different approach, a microfeature based approach, can better deal with these problems, and can thus produce better computational models of metaphor interpretation.

Let us look into microfeatures. Many connectionist models emphasize "distributed" representation that employs fine-grained meaning elements, microfeatures, for capturing the meaning/semantics of concepts, in addi-

tion to, or instead of, explicit conceptual links. Such microfeatures can either be extracted manually from (macro-level) domain theories, such as in Sun (1995), or acquired through applying learning algorithms that automatically develop a fine-grained internal (uninterpretable) representation.

Representations with microfeatures have some interesting properties that are especially relevant in the current context. Firstly, with such representations, we do not have some of the ad hoc-ness of hand-coded hierarchies, because we no longer make arbitrary design decisions. Secondly, because the representation is "flattened" (i.e. there is no path to trace), the problem of the complexity of search algorithms and the time needed to traverse links is no longer existent. Thirdly, microfeature representations tend to be more context sensitive, and are thus more flexible. Fourthly, similarities are easy to compute with microfeature representations; various similarity measures can be implemented with slightly different node activation functions and link weights (see Sun 1995 for details), without either explicit similarity links or explicit conceptual hierarchies. Lastly, although there is no *explicit* hierarchies, hierarchies can nevertheless be embodied in microfeatures and used in inference, as demonstrated by Sun (1993).

The question now is exactly how we use microfeature based representations for metaphor interpretation in a computational architecture. One possible candidate architecture is CONSYDERR (Sun 1995), a connectionist architecture that utilizes both localist and distributed representation and embodies all of the aforementioned properties of microfeature based representation. It is meant to be a comprehensive model of robust commonsense reasoning, and as such, it should be able to handle metaphor and non-literal language as well.

The most important idea behind CONSYDERR is the two-level dual-representation scheme, that is, there are two parts (levels) in the model: one is localist, representing concepts as individual nodes, and the other is distributed, representing concepts as microfeatures (hence the dual representation). There are two-way connections between each pair of corresponding representations across levels. The working of the model is divided into three phases. In the top-down phase, the computation is as follows:

$$A_{x_i} = \max_j A_{a_j}$$

where a_j is any node in the top level (the concept level, or CL) and x_i is its microfeature (in the microfeature level, or CD). In the settling phase, the computation is as follows:

$$A_{a_j} = \sum_k W_k * I_k$$

and

$$A_{x_i} = \sum_k w_k * i_k$$

where W_k's and w_k's are intra-level links weights, and I_k's and i_k's are the activations of related nodes. [1] In the bottom-up phase, the computation is as follows:

$$A_{a_j} = \max(A_{a_j}, \sum_{A_{x_i} \in F_{a_j}} \frac{A_{x_i}}{|F_{a_j}|})$$

where a_j is any node in the concept level, F_{a_j} is its microfeature set, and $|F_{a_j}|$ is the size of the microfeature set. Through the interaction of the two levels, many difficult patterns in commonsense reasoning can emerge naturally (Sun 1995).

Now the question is as follows: How can we utilize this model to implement a microfeature-based computational model of metaphor? There are some existing theories that we may draw upon regarding detailed metaphoric interaction. [2] Ortony (1979) posited a theory of metaphor based on salience imbalance and attribute enhancement: as touched upon before, a metaphor enhances, or highlights, some attributes in the target domain that are highly salient in the source domain but are far less salient in the target domain; therefore something new will be produced in the target domain as a result of the metaphor (so in some sense, we can say that some highly salient attributes get transferred from the source domain to the target domain).

There are, however, many unanswered questions and unspecified details that need to be looked into. For one thing, the idea of "attributes" needs to be better characterized: What constitutes an attribute? How can we ascertain if a concept has a certain attribute or not? What is the proper granularity of such attributes? and so on. Another related question is how attributes of different concepts are compared: How corresponding attributes are found (Tourangeau & Sternberg 1982), and how similarities between these attributed are determined (as touched upon in Ortony 1979 under the rubric of attribute inequality), since slightly different attributes can be matched, or one can be converted into another as appropriate when transfers occur. If we start to allow certain structures among attributes in order to handle such matches, then we open the door back to the full-fledged semantic networks, which has already been shown as inappropriate.

Another particularly important issue in relation to this theory is what attributes should or should not get transferred from the source domain to the target domain, since surely not all the highly salient attributes in the source domain can be transferred. So the question is how the selection can be done. One question that I raised earlier in relation to semantic network based approaches is also relevant here and is not dealt with in Ortony's theory: How does the context affect the salience of attributes of a concept and thus affect the outcome of metaphoric interaction? Certainly the claim that the salience of each attribute will stay the same always, regardless of circumstances, is an indefensible position.

The solution below, albeit a little simplistic, captures the basic ideas discussed above. In CONSYDERR, a

[1] This phase will not be used in the present work.

[2] The interaction of the two semantic domains of the two terms involved is hypothesized by Black (1955) to be the main factor underlying metaphor interpretation. This work is based on the belief that the process can be captured in micorfeatures.

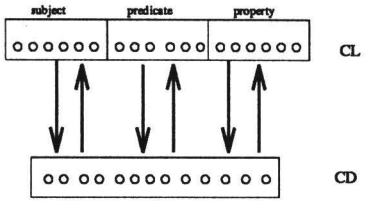

Figure 1: A Connectionist Model for Metaphor Interpretation

Note that there are three separate segments of nodes for the subject and the predicate of an input sentence and for the properties.

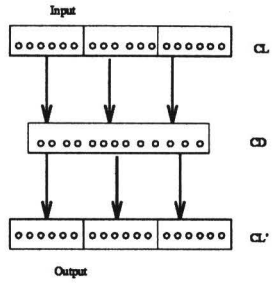

Figure 2: An Equivalent Model Used In Training

1. billboard	2. pole	3. fence	4. gate	
5. wart	6. mouth	7. nose	8. ear	9. hair

Figure 3: A List of Domain Objects

layer of nodes (*CL*) is established for representing concepts in a domain, including physical objects and their properties (e.g., *big, billboard*, etc. as in "the billboards are big"), and another layer (*CD*) is established for representing microfeatures that may be needed in processing the meaning of metaphoric (and non-metaphoric) statements. The weight on a link from a concept node to a microfeature node represents the salience of that microfeature in relation to the concept. What we want to achieve is the following: When a predicate metaphor, such as "Billboards are warts", is presented to the concept layer by the activation of the nodes for "billboard" and "wart", these activated nodes in turn activate the related microfeature nodes in *CD*, and then, through nonlinear interaction in the microfeature layer, a proper interpretation emerges in the form of proper activations (that is, those and only those microfeatures involved in the final interpretation remain activated); these activated microfeatures will then go back up to activate the proper nodes in the concept layer that represent the proper interpretation of the metaphor ("ugly" in this case). See Figure 1 for a sketch of the model.

I will start *tabula rasa*, without hand-coding any a priori knowledge of microfeatures and concepts. Thus the knowledge must be acquired through learning. I first fully connect the two layers with small random initial weights, in order for them to develop proper connectivity patterns through data presentation using connectionist learning procedures. Although there is apparently no readily applicable learning algorithm for such a two-layer three-phase model, we can "unfold" the model to come up with an equivalent three-layer model by duplicating the concept layer and having the duplicated layer carry the bottom-up weights (while the original concept layer retain its top-down weights). The input activations in the training data will be applied to the original concept layer and the output activation representing the right interpretation of the input will be applied to the duplicated concept layer. See Figure 2. Since the metaphoric interaction is rather complex (Ortony 1979, Black 1955) a linear combination operation may be insufficient. To improve the processing power, I adopt a sigmoidal activation function instead of the operations specified earlier. Now the backpropagation algorithm can be applied directly to update the two sets of weights, until the correct interpretation is obtained for each training case. Proper microfeatures are developed in the meantime. The weights acquired in the end represent the salience of the respective microfeatures formed in the *CD* layer.

An example test domain is as follows. There are two sets of concepts: one set includes five facial objects and the other includes four land objects (see Figure 3). Each object has a number of properties, which should be activated properly upon the presentation of a corresponding objects. All the properties together form another set of concepts. Each of the above three sets is represented by a corresponding collection of concept nodes in the CL layer of the model. The interpretation of the meaning of each object concept is based on the properties listed in Figure 4. Metaphoric statements are constructed systematically from applying a concept from one set of objects to another concept from the other set, such as

Billboards are warts
Warts are billboards
Noses are gates
Gates are noses

including all the possible combinations (there are $2*4*5 = 80$ of them). A subset of these constructed statements is used for training, and the rest for testing. With the resulting network, given the sentence "Billboards are warts", the node in the subject segment of the concept layer representing "billboard" is activated and the node in the predicate segment representing "wart" is also activated; with these input activations, during the top-down phase, some of the microfeatures in the bottom layer are activated in response to the top-down activation flow; then in the bottom-up phase, those activated microfeatures in turn activate the proper nodes in the property segment of the top layer, which in this case include, most prominently, "ugly". Therefore the resulting interpretation is "billboards are ugly", among some other less prominent properties. See Figure 5 for more examples of metaphor interpretation.

There is another issue that is worth mentioning. Context can force changes in feature salience; that is, a microfeature that is very prominent in one context can be negligible in another. I need a way to modulate microfeature salience (i.e., weights), which can be accomplished by including an extra set of nodes (a context module) in

big	medium	small	
man-made	natural		
ill	healthy		
ugly	beautiful		
facial	on-ground		
breathing	smelling	hearing	
flat	protrusive		
standing-up	surrounding	covering	entering
protective			
obvious	advertising		
non-sense	making-sense		

Figure 4: A List of Properties Associated with Domain Objects

Billboards are warts	ugly
Warts are billboards	obvious
Ears are gates	entrance
Gates are ears	hearing

Figure 5: Some Examples of Metaphor Interpretation

the concept layer, which receives inputs and sends out signals that change the weights on the link connecting concepts and their corresponding microfeatures. In other words, these signals are used to manipulate or modulate microfeatures on the fly (Sun 1995). This change (or modulation) is done by higher-order links (three-way connections) of concepts and microfeatures; that is, the link weight between a concept and a microfeature is:

$$w'_{ij} = w_{ij} * w_c.$$

where w_c is a modulation signal from the context module. See Figure 6. The same idea in developing training procedures is again applied, by unfolding the architecture into multiple layers and by handling the updating of high-order weights (details omitted).

4 Further Steps

This model produces the desired outcomes, through handling the metaphoric interaction by two sets of non-linear (sigmoidal) operations successively applied, which have been shown to be able to approximate any measurable function and are thus sufficient for the purpose. However, we may want a more linear and more precise characterization.

In the interest of specifying the detailed operations of metaphoric interaction, I propose to use a process that involves *enhancement*. That is, we selectively enhance the links between the target concept and some of its

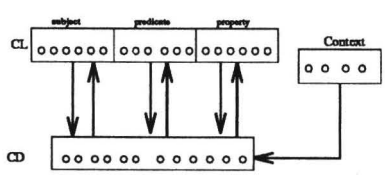

Figure 6: The Overall Architecture for Metaphor Interpretation

microfeatures (this is similar in a way to "slippage" as proposed by Hofstadter). In accordance with the theory of salience imbalance (Ortony 1979), these microfeatures selected to be enhanced are among those that have strong links with the source concept (i.e., highly salient for the source concept) but weak links with the target concept. After the enhancement, there will be some highly salient microfeatures shared by both the source concept and the target concept (so these microfeatures have been transferred), which is not the case before the enhancement (provided that the statement to be interpreted is metaphoric; Ortony 1979). [3]

To perform enhancement, let us go back to the original CONSYDERR architecture: without the sigmoidal operations, the concepts/microfeatures relations are once again transparent. Enhancement can then be accomplished by setting up links with high weights between the target concept and the microfeatures that have strong connections with the source concept, with each weight proportional to (a function of) the corresponding weight between the microfeature and the source concept and the original weight between the microfeature and the target concept. That is,

$$\Delta w^t_{ji} = \text{sign}(w^t_{ji}) \max(w^s_{jk}, \alpha * \frac{1 - e^{-\beta_1 |w^s_{jk}|}}{1 + e^{-\beta_1 |w^s_{jk}|}} * \frac{1 - e^{-\beta_2 |w^t_{ji}|}}{n + e^{-\beta_2 |w^t_{ji}|}})$$

when w^s_{jk} and w^t_{ji} are of the same sign, where n, α, β_1, and β_2 are parameters, i is the currently activated target concept, k is the currently activated source concept, and j is the microfeature in question. These parameters can be chosen in such a way that (1) the first fraction will have a steep slope (when β_1 is large), because only those microfeatures of the source concept that are highly salient can get transferred, and (2) the second fraction will have a gentle and low slope (when β_2 is small and n is used to compress its range), because we want the amount of transfer to correspond, to a small degree, to the original weight between the target concept and the microfeature in question since that weight represents a propensity of the target concept to obtain that microfeature. (Therefore, we have $\beta_1 >> \beta_2$.)

With this update rule, if the source weight (w^s_{jk}) and the target weight (w^t_{ji}) are both high, then the target weight will basically remain the same, as it should be in case the statement is a literal comparison; if both are low, there will be very little enhancement as expected; if the target weight is high and the source weight is low, there will be little enhancement; if the source weight is high and the target weight is low, there will be significant enhancement; in this last case, however, the higher the target weight the more enhancement there will be (up to a certain limit).

With the enhanced top-down links from the target concept (and without the source concept), the three

[3] According to Ortony (1979), if the source concept and the target concept originally share some highly salient features, then the statement will not be metaphoric, but a simple literal comparison. Metaphors occur when none of those highly salient features of the source domain are highly salient features of the target domain.

phases of CONSYDERR are reenacted. From this three-phase cycle, some properties representing possible interpretations will be activated, in the end, in the property segment of the CL layer. We then use the most strongly activated properties as the interpretation of the metaphor.

5 Discussions

Going back to the issues raised earlier, we see that the model fare well in all of these aspects:

- First of all, with the microfeature based approach there is no need for complex search algorithms and the time needed to reach a conclusion is not dependent on the path length between them. Simple microfeature transformation generates the equivalent of complex search behavior without actually performing the search (see Sun 1993 for more details).
- There is no hand-coding that creates arbitrary conceptual hierarchies; rather, conceptual structures are embedded in microfeature representation and, more importantly, are extracted from data.
- The microfeatures themselves are not arbitrarily chosen either; they are formed automatically through learning correct mappings.
- Because such automatic formation of microfeatures, they naturally capture the similarity of concepts, as demonstrated by the resulting correct interpretations in the experiments.
- Because of automatic formation of microfeatures, attribute inequality is avoided, since there will be automatic decomposition of attributes to microfeatures that capture similarities among attributes through distributed representation.
- The models are capable of highlighting certain aspects of the target domain by *transferring* highly salient features from the source domain in metaphor interpretation. Thus, metaphors are not simply constraint violation, mapping between domains, or selective inference (although these are all integral part of metaphor interpretation and are embodied in the model).
- The effect of context is taken into consideration in the model; with microfeature, this effect can be realized by context modules and high order links.

There exist other related connectionist models (see, e.g., Eskridge 1993, Lange and Wharton 1993). These models are mostly localist, and as such, they are more akin to semantic network based approaches than to the microfeature based approach proposed in the present work. These models, however, do have some noteworthy features; for example, they allow very complex structures in their representations and they compute and utilize complex structural correspondences (e.g., Eskridge 1993). They are also capable of explicitly expressing goals and plans. I should note that these representations are possible in CONSYDERR, since it has a localist level (*CL*) that can readily implement the aforementioned structures (as in Sun 1992).

One advantage of the models proposed is that they employ the same mechanism for both metaphoric or literal statements. Metaphors are not viewed as an extraneous process serving only ornamental purposes. Since the model presented here is a variation of CONSYDERR, which is a unified model of various kinds of commonsense reasoning, ranging from inheritance reasoning (Sun 1993) to evidential causal reasoning (Sun 1995) and to similarity-based induction (Sun 1995), the model is integrative and not ad hoc. Some of the issues addressed in this work are also applicable to other areas.

Moreover, because of the possibility of incorporating structural correspondences and structural similarities into CONSYDERR, it is possible to extend the proposed models of metaphor interpretation to deal with more sophisticated kind of analogy and analogical reasoning (involving complex relations). Some work is currently being carried out in this direction.

Acknowledgements

I thank John Barnden and Larry Bookman.

References

L. Barsalou, (1989). Intraconcept similarity and its implications. in *Similarity and Analogical Reasoning*, Cambridge U. Press.

M. Black, (1955). Metaphor. *Proc.of the Aristotelian Society.*

T. Eskridge, (1992). A hybrid model of continuous analogical reasoning. In *Advances in Connectionist and Neural Computation Theory II.* Ablex

D. Fass, (1991). Met: A method for metonymy and metaphor. *Computational Linguistics*, 17(1)

G. Lakoff and M. Johnson, (1980). The metaphoric structure of human conceptual system. *Cognitive Science*, 4, 193-208.

T. Lange and C. Wharton, (1992). REMIND: retrieval from episodic memory. *Advances in Connectionist and Neural Computation Theory II.* Ablex

E. R. MacCormac, (1988). *A Cognitive Theory of Metaphor.* MIT Press.

J. Martin, (1988). A Computational Theory of Metaphor. Ph.D Thesis, University of California, Berkeley.

A. Ortony, (1979). Beyond literal similarity. *Psychological Review.* 86(1), 161-180.

R. Schank, (1977). *Scripts, Plans and Goals.* LEA.

R. Sun, (1992). On variable binding. *Connection Science.* 2, 93-124.

R. Sun, (1993). An efficient feature-based connectionist inheritance scheme, *IEEE Transaction on SMC*, 23(2).

R. Sun, (1995). Robust reasoning. *Artificial Intelligence.* June.

R. Tourangeau and R. Sternberg, (1982). Understanding and appreciating metaphors. *Cognition.* 11, 203-244.

T. Veale and M. Keane, (1992). Conceptual scaffolding. *Computational Intelligence*, 8(3) 494-519.