Hybrid Neural Plausibility Networks for News Agents

Stefan Wermter, **Christo Panchev** and **Garen Arevian**
The Informatics Center
School of Computing, Engineering & Technology
University of Sunderland
St. Peter's Way, Sunderland SR6 0DD, United Kingdom
Email: stefan.wermter@sunderland.ac.uk

Abstract

This paper describes a learning news agent HyNeT which uses hybrid neural network techniques for classifying news titles as they appear on an internet newswire. Recurrent plausibility networks with local memory are developed and examined for learning robust text routing. HyNeT is described for the first time in this paper. We show that a careful hybrid integration of techniques from neural network architectures, learning and information retrieval can reach consistent recall and precision rates of more than 92% on an 82 000 word corpus; this is demonstrated for 10 000 unknown news titles from the Reuters newswire. This new synthesis of neural networks, learning and information retrieval techniques allows us to scale up to a real-world task and demonstrates a lot of potential for hybrid plausibility networks for semantic text routing agents on the internet.

Introduction

In the last decade, a lot of work on neural networks in artificial intelligence has focused on fundamental issues of connectionist representations (Hendler 1991; Miikkulainen 1993; Giles & Omlin 1993; Sun 1994; Honavar 1995). Recently, there has been a new focus on neural network learning techniques and text processing-such as for newswires and world wide web documents (Papka, Callan, & Barto 1997; Lawrence & Giles 1998; Craven *et al.* 1998; Joachims 1998).

However, it has been an open research question (Wermter & Sun 1998) as to whether hybrid neural architectures will be able to learn large-scale real-world tasks, such as learning the classification of noisy newswire titles. Neural networks with their properties of robustness, learning and adaptiveness are good candidates for weighted rankings and weighted routing of ambiguous or corrupted messages, in addition to well-established techniques from information retrieval, symbolic processing, and statistics (Lewis 1994).

In this paper, we develop and examine new **Hy**brid **Ne**ural/symbolic agents for **T**ext routing on the internet (HyNeT). We integrate different preprocessing strategies based on information retrieval with recurrent neural network architectures, including variable-length short-term memories. In particular, we explore simple recurrent networks and new more sophisticated recurrent plausibility networks. Recurrent plausibility networks are extended with a dynamic short-term memory which allows the processing of sequences in a robust manner. As a real-world testbed, we describe extensive experiments with learning news agents.

Recurrent Plausibility Networks

Recurrent neural networks introduce previous states and extend feedforward networks with short-term incremental memories. In fully recurrent networks, all the information is processed and fed back into one single layer. Partially recurrent networks, such as simple recurrent networks, have recurrent connections between the hidden layer and context layer (Elman 1990) or Jordan networks have connections between the output and context layer (Jordan 1986).

In other research (Wermter 1995), different decay memories were introduced by using distributed recurrent delays over the separate context layers representing the contexts at different time steps. At a given time step, the network with n hidden layers processes the current input as well as the incremental contexts from the $n-1$ previous time steps.

Figure 1 shows the general structure of our recurrent plausibility network. It combines the features of recurrent networks with distributed context layers and self-recurrent connections of the context layers. The input to a hidden layer L_n is constrained by the underlying layer L_{n-1} as well as the incremental context layer C_n. The activation of a unit $L_{ni}(t)$ at time t is computed on the basis of the weighted activation of the units in the previous layer $L_{(n-1)i}(t)$ and the units in the current context of this layer $C_{ni}(t)$ limited by the logistic function f.

$$L_{ni}(t) = f(\sum_k w_{ki} L_{(n-1)i}(t) + \sum_l w_{li} C_{ni}(t))$$

The units in the context layers perform a time-averaging of the information using the equation

$$C_{ni}(t) = (1-\varphi_n) L_{ni}(t-1) + \varphi_n C_{ni}(t-1)$$

where $C_{ni}(t)$ is the activation of a unit in the context layer at time t. We represent the self-recurrent connections using the hysteresis value φ_n. The hysteresis value of the context layer C_{n-1} is lower than the hysteresis value of the next context layer C_n. This ensures that the context layers closer to the input layer will perform as memory that represents a more dynamic context. Having higher hysteresis values, the context layers closer to the output layer will incrementally build more stable sequential memory. Therefore the larger context is built on the more recent, dynamic one.

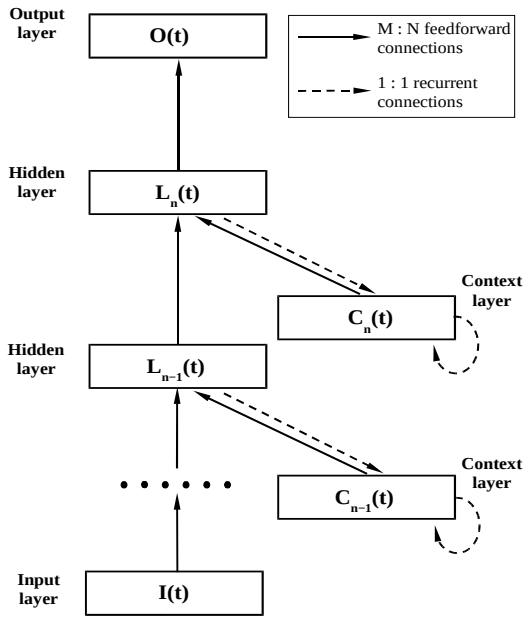

Figure 1: Recurrent plausibility network.

The Reuters News Corpus

For learning the subtask of text routing, we used the Reuters text categorization test collection (Lewis 1997). This corpus contains documents which appeared on the Reuters newswire. All news titles in the Reuters corpus belong to one or more of eight main categories[1]: Money/Foreign Exchange (**money-fx**, **MF**), Shipping (**ship**, **SH**), Interest Rates (**interest**, **IN**), Economic Indicators (**economic**, **EC**), Currency (**currency**, **CR**), Corporate (**corporate**, **CO**), Commodity (**commodity**, **CM**), Energy (**energy**, **EN**).

Examples of typical titles from these categories are shown in Table 1. As we can see, there are abbreviated phrases or sentences, specific characters, and terms which would make it difficult to manually encode a traditional semantic grammar. We can find incompleteness or ambiguity in the category assignments. For example, the title "U.K. money market offered early assistance" occurs six times in the corpus belonging to different semantic categories. Three times it is classified into the "money-fx" category, and three times into both "money-fx" and "interest" categories. Therefore there is an inherent ambiguity in the corpus and such examples pose challenges to learning algorithms and necessarily increase the classification error, since there is no perfect unambiguous category assignment in such cases. However, since we wanted to study hybrid neural agents under hard, real-world conditions, we did not perform any cleaning up of the underlying corpus data.

We use exactly all 10 733 titles of the so-called Mod-Apte split whose documents have a title and at least one topic. The total number of words is 82 339 and the number of different words in the titles is 11 104. For our training set, we use 1 040 news titles, the first 130 of each of the 8 categories. All the other 9 693 news titles are used for testing the generalization to new and unseen examples. The description of this corpus is shown in Table 2. Since categories may overlap, i.e. one news title can be in exactly one or several semantic categories, the actual distribution of the titles over the training set is not even.

Category	Training titles		Test titles	
	Number	Average length	Number	Average length
money-fx	286	8.26	427	8.13
ship	139	7.04	139	7.32
interest	188	8.79	288	8.50
economic	198	8.43	1 099	8.50
currency	203	8.54	92	8.42
corporate	144	7.10	6 163	7.43
commodity	321	7.35	2 815	7.54
energy	176	7.80	645	8.08
All titles	1 040	7.96	9 693	7.64

Table 2: The distribution of the titles from the Reuters news corpus over the semantic categories. Note that since one title can be classified in more than one semantic category, 1 040 titles represent the total number of 1 655 category occurrences for the training set and 9 693 titles represent the total number of 11 668 category occurrences in the test set.

Hybrid Neural News Routing

In this section, we will describe different architectures for learning news title classification. Each news title is presented to the network as a sequence of word input representations and category output representations, one such pair for each word. At the beginning of a news title, the context layers are initialized with 0 values. Each unit in the output layer corresponds to a particular semantic category. Those output units (one

[1]There is also a second level categorization into 135 specific categories, but since we wanted to study the learning of difficult and possibly ambiguous classifications, we used the 8 main categories from the first level.

Semantic Category	Title
money-fx	Bankers trust issuing STG/DLR currency warrants
ship	U.S. cargo preference squabble continues
interest	Volcker says FED policy not responsible for prime rate rise
economic	German net currency reserves rise 400 mln marks to 87.0 billion - Bundesbank
currency	Stoltenberg not surprised by dollar reaction
corporate	First Granite Bancorp Inc agrees to be acquired by Magna Group Inc for stock
commodity	Indonesian coffee production may fall this year
energy	U.S. oil dependency seen rising to record level
ship&energy	Kuwait may re-register gulf tankers - Newspaper
money-fx&interest&currency	J.P. Morgan <JPM> says DLR may prevent FED easing

Table 1: Example titles from the corpus.

or more) which represent the desired semantic categories are set to 1. All other output units are set to 0. We define a news title as *classified* to a particular semantic category if at the end of the sequence the value of the output unit for the desired category is higher than 0.5. Using this output classification, we compute the recall and precision values for each title. These values are used to compute the average recall and precision rates for each semantic category, as well as for the overall training and test sets, which all determine the network performance.

Supervised learning techniques based on plausibility networks were used for the training (Rumelhart *et al.* 1995). The training regime forces the recurrent plausibility network to assign the desired category, starting from the beginning of the news title as early as possible. Supervised training is continued until the error over the training set stops decreasing. Typically, between 700 and 900 epochs through the training set were necessary. In one epoch, we present all titles from the training set. Weights were adjusted at the end of each title.

Complete Titles and Significance Vectors

In our first set of experiments, we use significance vectors as the basis for representing words for semantic category routing. Significance vectors are determined based on the frequency of a word in different semantic categories. Each word w is represented with a *vector* $(c_1 c_2 \cdots c_n)$, where c_i represents a certain semantic category. A *value* $v(w, c_i)$ is computed for each dimension of the vector as the frequency of occurrences of word w in semantic category c_i (the category frequency), divided by the frequency of occurrences of word w in the corpus (the corpus frequency). That is:

$$v(w, c_i) = \frac{Frequency\ of\ w\ in\ c_i}{\sum_j Frequency\ for\ w\ in\ c_j}\ for\ j \in \{1, \cdots n\}$$

The same significance vector can represent two different words if they actually occur with the same frequency across all semantic categories in the whole corpus. However, a news title will be represented by a sequence of significance vectors so that phrases with the same sequence of significance vectors are less likely. Figure 2 shows the significance vectors of words from the lexicon. As we can see, words like "mortgage", "bureau" or "parker" have clear semantic preferences for specific semantic categories, while domain-independent words like "a", "of", "and" have more distributed preferences.

Word	MF	SH	IN	EC	CR	CO	CM	EN
a	.07	.02	.04	.17	.04	.28	.27	.10
and	.06	.02	.03	.15	.03	.25	.34	.09
bureau	.02	.00	.00	.38	.02	.02	.56	.01
mortgage	.01	.00	.30	.18	.00	.51	.00	.00
of	.07	.02	.03	.14	.03	.28	.31	.09
parker	.13	.00	.00	.00	.13	.73	.00	.00
the	.09	.03	.05	.18	.05	.20	.31	.09

Figure 2: Examples of significance vectors.

Simple recurrent networks were trained using significance vector representations as the input word representation of the input layer. The performance of the best trained network in terms of recall and precision is shown in Table 3. The table contains detailed results for each semantic category as well as for the whole training and test sets. The two categories "money-fx" and "economic" are fairly common, which explains their lower recall and precision rates, but in general, 91.23% recall and 90.73% precision are reached for the 9 693 unknown test titles.

Complete Titles and Semantic Vectors

In our second set of experiments, we use a different semantic vector representation of the words in the lexicon. These vectors mainly represent the plausibility of a particular word occurring in a semantic category and they are independent of the number of examples observed in each category. A *value* $v(w, c_i)$ is computed for each element of the semantic vector as the *normalized* frequency of occurrences of word w in semantic category c_i (the normalized category frequency), divided by the

Category	Training set		Test set	
	recall	precision	recall	precision
money-fx	84.36	84.62	84.74	69.56
ship	81.06	93.17	77.34	94.96
interest	77.93	82.71	85.42	83.45
economic	71.70	80.79	74.74	77.82
currency	85.75	91.46	85.36	87.16
corporate	88.81	92.31	94.87	95.10
commodity	86.27	94.77	86.47	88.31
energy	81.88	92.26	85.22	91.65
Total	85.15	86.99	**91.23**	**90.73**

Table 3: Results using significance vectors

Category	Training set		Test set	
	recall	precision	recall	precision
money-fx	87.78	88.60	84.07	69.59
ship	81.65	88.13	82.73	93.88
interest	85.33	86.97	88.25	88.19
economic	76.22	83.54	78.36	80.30
currency	87.76	92.59	89.64	89.86
corporate	89.16	91.96	95.90	95.98
commodity	86.27	90.52	86.20	87.22
energy	89.19	95.48	86.58	91.56
Total	88.57	88.59	**92.47**	**91.61**

Table 4: Results using semantic vectors

normalized frequency of occurrences of word w in the corpus (the normalized corpus frequency). That is:

$$v(w, c_i) = \frac{Norm.\ freq.\ of\ w\ in\ c_i}{\sum_j Norm.\ freq.\ for\ w\ in\ c_j},\ j \in \{1, \cdots n\}$$

where:

$$Norm.\ freq.\ of\ w\ in\ c_i = \frac{Freq.\ of\ w\ in\ c_i}{Number\ of\ titles\ in\ c_i}$$

Some examples of semantic vectors are shown in Figure 3. As we can see, the domain-independent words like "a", "of", "and" have fairly even distributions, while the domain-dependent words "bureau", "parker" have more specific preferences. An example of the difference between the two types of representation is the word "mortgage". Compared to the significance representation, here the preference entirely changes from the category "corporate" to the category "interest".

Word	MF	SH	IN	EC	CR	CO	CM	EN
a	.13	.12	.11	.16	.16	.06	.11	.15
and	.12	.12	.09	.16	.14	.05	.15	.15
bureau	.00	.00	.00	.50	.11	.00	.32	.01
mortgage	.02	.00	.72	.15	.02	.09	.00	.00
of	.13	.11	.11	.15	.15	.06	.14	.14
parker	.25	.00	.00	.00	.56	.17	.00	.00
the	.15	.13	.12	.15	.18	.04	.11	.12

Figure 3: Examples of semantic vectors

Simple recurrent networks were trained using the same training parameters and network architecture as in the previous experiment, but the titles were processed using the semantic vector representations. The performance of the trained network in terms of recall and precision is shown in Table 4. Using the semantic vectors rather than the significance vectors, we improved the test recall and precision rates to 92.47% and 91.61% respectively.

News Titles without Insignificant Words

So far we have examined complete phrases in a simple recurrent network. However, a symbolic preprocessing strategy from information retrieval could be useful, one that emphasizes significant domain-dependent words. To investigate such a strategy from information retrieval, we removed insignificant words (sometimes also called "stop words") from the unrestricted title phrases. We defined the set of *insignificant words* as equally frequent, domain-independent words that belong to any of the following syntactic categories: determiners, prepositions and conjunctions. Using the semantic vector representation, we extracted a set of 19 insignificant words of which the difference between the highest and lowest values in their vector was less than 0.2. For instance, examples of these words are: *a, an, and, as, at, but, by, for, from.* These words occur 7 094 times in the training and test corpus. After removing these words, the average length of a title in the training set is 7.11 and average length of a title in the test set is 7.00.

The experiment was conducted with the same learning conditions as for the complete unrestricted titles. The test results are shown in Table 5. We can see that the removal of insignificant words improves the recall and precision rates slightly over the last two experiments, but the improvement is small. However, by removing the insignificant stop words, we have also deleted some domain-independent words; hence this experiment also suggests that noisy corrupted titles can also be processed.

Category	Training set		Test set	
	recall	precision	recall	precision
money-fx	87.40	89.12	84.97	69.36
ship	79.14	88.85	78.42	93.53
interest	86.61	93.09	88.89	88.77
economic	79.12	87.80	79.28	82.64
currency	85.55	92.59	87.39	87.16
corporate	89.51	93.71	96.47	96.48
commodity	87.25	91.18	86.74	86.51
energy	85.32	94.19	83.86	90.54
Total	88.47	90.05	**92.88**	**91.92**

Table 5: Results without insignificant words

Complete Titles with Plausibility Networks

In this experiment, we use the recurrent plausibility network described in Figure 1 with two hidden and two context layers. Different networks have been trained using different combinations of the hysteresis value for the first and second context layers. The best results were achieved with the network having a hysteresis value of 0.2 for the first context layer, and 0.8 for the second. In order to be able to compare the results of this architecture with the previous ones, the rest of the training and testing environment parameters were set to be the same as in the experiment of the complete titles with the semantic vector representation of the words. Table 6 shows the recall and precision rates obtained with the recurrent plausibility network. There was an improvement in the classification, especially for the longer titles, as the two context layers and recurrent hysteresis connections support a larger and dynamic short-term memory.

Category	Training set		Test set	
	recall	precision	recall	precision
money-fx	87.34	89.47	86.03	76.70
ship	84.65	89.21	82.37	90.29
interest	85.24	87.77	88.19	86.05
economic	90.24	91.77	81.89	83.80
currency	88.89	91.36	89.64	89.86
corporate	92.31	92.66	95.55	95.43
commodity	92.81	93.14	88.84	90.29
energy	85.27	87.74	87.69	92.95
Total	89.05	90.24	**93.05**	**92.29**

Table 6: Results using recurrent plausibility network with semantic vectors

Examples of the Network Output

Figure 4 presents the output representations after processing three unknown example titles from the test corpus with the recurrent plausibility network from the last experiment. Here we show a representative behavior of the network, and in all these examples, the computed final categories are correct. The first two examples start with the same word sequence "Bank of Japan", but later, when a sequence of more domain-dependent words has been processed, the network changes its output preference according to the semantic meaning of the different input sequences: "intervenes shortly after Tokyo opens" provides a "money-fx" and "currency" category assignment, while "easy money policy" leads to an "interest" category assignment. Both these assignments are correct according to the Reuters classification.

In the third example in Figure 4, we illustrate how the network, based on the incrementally built context, can turn the strong output preference from the category "ship" to "energy". The network starts with preference for the "ship" and "energy" categories because, according to all news documents from the corpus, "Iran"

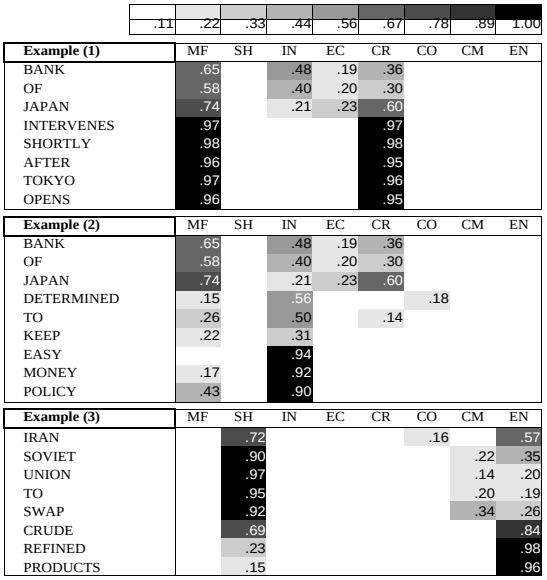

Figure 4: Examples and their category preferences

is associated with oil shipping. A recognizable number of news titles in the corpus are about the Soviet Union commodity crisis and are classified in the "shipping" and "commodity" categories. In the middle of the title however, when "crude" is processed as a significant term for the "energy" category, the output preference is changed to this category and subsequent words confirm this category assignment.

Discussion

We have described several hybrid neural architectures for classifying news headlines. In general, the recall and precision rates for simple recurrent networks and recurrent plausibility networks were fairly high given the degree of ambiguity of title/category and word/category assignments (for instance, see Table 1). The generalization performance for new and unseen news titles has been even better than the performance on the training data (for instance, see experiment 1). This is a very desirable effect and demonstrates that overfitting on the training set does not exist.

In other related work on text classification, whole documents rather than titles are often used to classify a news story. Obviously whole documents contain more information than titles, but we wanted to examine how far we can get in terms of classification based only on the title. However, since a news title contains on average around 8 words, it is unlikely that most of the words in the sequence are ambiguous and that they would lead to an incorrect category. Our good recall and precision results of at least 92% confirm that the influence of misleading word representations is limited.

Other related work on a different book title classification task using neural networks (Wermter 1995) has reached 95% recall and precision, but the library titles

are much less ambiguous and only 1 000 test titles have been used in that approach while we used 10 000 corrupted and ambiguous titles. Also, the titles in our task are about 25% longer and more difficult to classify.

Our integration of new variations of the vector space model (significance vectors, semantic vectors) with nonlinear incremental classification from neural networks was shown as a viable new way for classification and routing tasks. The removal of insignificant stop words was shown to provide better results in the simple recurrent network as well. Best results were obtained with a plausibility network with two hidden layers, two context layers and recurrent connections for the context layer. Although the performance was in general fairly high for different architectures, an additional 2% improvement in the test set would mean another 200 correctly classified titles. The architecture with plausibility network has more possibilities to encode the preceding context and reaches at least 92% minimum value for recall and precision.

There has been interesting related work on text categorization on the Reuters corpus using whole documents (Joachims 1998). For the ten most frequently occurring categories, the recall/precision breakeven point was 86% for a Support Vector Machine, 82% for k-Nearest Neighbor, 72% for Naive Bayes. However, a different set of categories and whole documents rather than titles are used and therefore the results are not directly comparable to ours. Nevertheless, they give some indication of document classification performance on this corpus. In particular, for medium text data sets or when only titles are available, our approach produces very good performance.

Conclusions

We have described and analyzed HyNeT, a novel news agent for real-world headline routing which learns to classify news titles. A neural network architecture with several context layers and recurrent hysteresis connections is proposed which particularly supports a larger and dynamic short-term memory. Different from most related approaches for language processing, we have used hybrid neural learning techniques.

HyNeT is robust, classifies noisy arbitrary real-world titles, processes titles incrementally from left to right, and shows better classification reliability towards the ends of titles. This synthesis of techniques and constraints from plausibility neural network architectures, information retrieval and learning holds a lot of potential for building robust hybrid neural architectures of semantic text routing agents for the internet in the future.

References

Craven, M.; DiPasquo, D.; Freitag, D.; McCallum, A.; Mitchell, T.; Nigam, K.; and Slattery, S. 1998. Learning to extract symbolic knowledge from the world wide web. In *Proceedings of the 15th National Conference on Artificial Intelligence*.

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

Giles, C. L., and Omlin, C. W. 1993. Extraction, insertion and refinement of symbolic rules in dynamically driven recurrent neural networks. *Connection Science* 5:307–337.

Hendler, J. 1991. Developing hybrid symbolic/connectionist models. In Barnden, J. A., and Pollack, J. B., eds., *Advances in Connectionist and Neural Computation Theory, Vol.1: High Level Connectionist Models*. Norwood, NJ: Ablex Publishing Corporation. 165–179.

Honavar, V. 1995. Symbolic artificial intelligence and numeric artificial neural networks: towards a resolution of the dichotomy. In Sun, R., and Bookman, L. A., eds., *Computational Architectures integrating Neural and Symbolic Processes*. Boston: Kluwer. 351–388.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*.

Jordan, M. I. 1986. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Conference of the Cognitive Science Society*, 531–546.

Lawrence, S., and Giles, C. L. 1998. Searching the world wide web. *Science* 280.

Lewis, D. D. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the Seventeenth Annual SIGIR Conference on Research and Development in Information Retrieval*.

Lewis, D. D. 1997. Reuters-21578 text categorization test collection, http://www.research.att.com/ lewis.

Miikkulainen, R. 1993. *Subsymbolic Natural Language Processing*. Cambridge, MA: MIT Press.

Papka, R.; Callan, J. P.; and Barto, A. G. 1997. Text-based information retrieval using exponentiated gradient descent. In Mozer, M. C.; Jordan, M. I.; and Petsche, T., eds., *Advances in Neural Information Processing Systems*, volume 9, 3. The MIT Press.

Rumelhart, D. E.; Durbin, R.; Golden, R.; and Chauvin, Y. 1995. Backpropagation: the basic theory. In Chauvin, Y., and Rumelhart, D. E., eds., *Backpropagation: theory, architectures and applications*. Hillsdale, NJ: Lawrence Erlbaum Associates. 1–34.

Sun, R. 1994. *Integrating Rules and Connectionism for Robust Commonsense Reasoning*. New York: Wiley.

Wermter, S., and Sun, R. 1998. *Nips Workshop on Hybrid Neural Symbolic Integration*. Breckenridge, CO: Nips.

Wermter, S. 1995. *Hybrid Connectionist Natural Language Processing*. London, UK: Chapman and Hall, Thomson International.