Modular Preference Moore Machines in News Mining Agents

Stefan Wermter and Garen Arevian
University of Sunderland
The Informatics Centre, SCET
St. Peter's Campus, St Peter's Way
Sunderland SR6 0DD, United Kingdom
www.his.sunderland.ac.uk

Abstract

This paper focuses on Hybrid Symbolic Neural Architectures that support the task of classifying textual information in learning agents. We give an outline of these symbolic and neural preference Moore machines. Furthermore, we demonstrate how they can be used in the context of information mining and news classification. Using the Reuters newswire text data, we demonstrate how hybrid symbolic and neural machines can provide an effective foundation for learning news agents.

1 Introduction

With the expansion of the Internet, a need has arisen to design more sophisticated learning agents which are capable of processing relevant information. Much initial work in the field of internet agents has used manual encoding techniques or simple techniques from information retrieval [24]. However, it becomes increasingly apparent that automatic adaptation, learning, dealing with incompleteness and robustness are necessary requirements [33]. Recently, there has been a new focus on machine learning techniques and language processing, for instance for newswires and World Wide Web documents [20; 21; 8].

Agents [3; 19] can be designed to perform various tasks, whether they be classification [12; 23], information retrieval and extraction [10; 8], routing of information [32; 30] or automated web browsing [2; 22; 7]. In general, robust learning architectures have been identified as important current areas for natural language processing [4; 9].

Statistical techniques have been shown to perform successfully in the classification of language [5]. When documents are organized in a large number of topic categories, the categories are often arranged in a hierarchy. For instance, a naive Bayes classifier is significantly improved by taking advantage of a hierarchy of classes [18]. However, these statistical methods require assumptions about the distribution.

Furthermore, self-organizing maps (SOMs) [14] have been used. A SOM forms a non-linear projection from a high-dimensional space onto low-dimensional space and has been used in the WEBSOM project [13; 15]. The SOM algorithm computes an optimal collection of models that approximates the data by applying a specified error criterion; this allows the ordering of the reduced dimensionality onto a map. The SOM is acting as a similarity graph of the data and is useful for structure visualization, data mining, knowledge discovery and retrieval [1; 11].

Another approach has been proposed whereby artificial life algorithms are applied in web-mining [19]. Adaptive and distributive algorithms seem to have the ability to capture the complexities of such a dynamic and complex environment as the World Wide Web, which can be regarded as a very large database of heterogeneous documents.

In summary, most internet agents, for instance classifiers, search engines, extractors, etc., still use *ad hoc* heuristic coding rather than adaptive machine learning techniques. However, adaptive learning web agents using neural network paradigms such as [27; 15; 30] hold a lot of promise as they support robustness and learning, are relatively autonomous in their learning behavior and offer the potential of on-line adaptivity. Our experiments with neural agents have been tested on noisy, real-world data and benchmarked on corpora like the Reuters corpus [30; 31]. In this paper we extend these results to hybrid symbolic neural preference Moore machines and demonstrate their results.

2 Preference Moore Machines as Modular Agents

One main motivation for modular agent systems is that they have a greater generalization ability, and classification tasks and target functions may be reached more readily [26]. Failure of one component does not necessarily mean an overall failure of the task, and indeed benefits arise from the communication between the various modular agents.

Another benefit is that such agent systems, for instance for modular classification, could form their own representations for a specific subtask. For example, mixture of experts approaches show that performance can be improved. For example, though recurrent networks are able to encode sequentiality, finite state machines might be more robust in encoding rules and relationships, and interaction between them could give better generalization.

Recurrent neural systems not only are able to embody some sort of contextual information, but they have the inherent ability to simulate any finite-state machine [16], essentially allowing an abstraction of the information within a recurrent

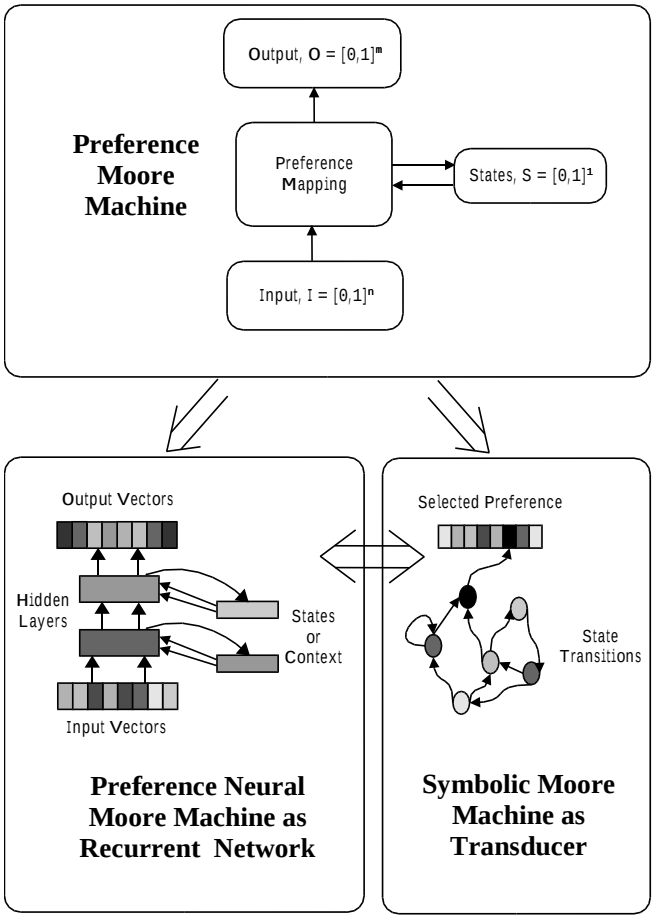

Figure 1: Relationship between Preference Moore Machines and Recurrent Networks

neural network [25; 6] into discrete representations such as grammatical rules and relations. This connection between recurrent networks and finite state machines can be exploited.

How is it possible to link the contextual information embodied in a recurrent neural network with a finite state system? There has been work on introducing Preference Moore Machines [29]. Here we want to extend this framework and bridge the gap between neural networks and such symbolic machines.

A Preference Moore Machine is a synchronous sequential machine that codes a sequential preference mapping, using current state S and the input I preferences, to assign an output preference O and a new state S. A Moore Machine is able to transduce knowledge from an input to output while maintaining context.

Preference Moore Machines can be seen as neural networks (Neural Preference Moore Machine) or as symbolic transducers (Symbolic Preference Moore Machine) as shown in Figure 1. For a Neural Preference Moore Machine, the internal state of the system and the context are represented as a n-dimensional vector. Using the Euclidean distance metric, different assignments can be made between this vector representation and a symbolic interpretation.

As non-neural techniques, symbolic transducers are considered as symbolic Preference Moore Machines because symbolic regularities are sometimes known. Rather than extracting them from training material, it is possible to encode the relationships and generate a transducer from regular expressions.

Symbolically encoded and neurally learned versions of preference Moore machines potentially represent very different forms of knowledge, and can be seen as different agents which can be combined as top-down and bottom-up models. This leads to systems using different agents with different representations. For a combination of such different preference Moore machines, we described initial operations for the integration of preference Moore machines [29].

Recently, we have shown that recurrent neural networks can indeed act as robust and scalable classifying agents for sequential tasks such as the classification of a stream of textual information of arbitrary lengths [32]. This work on neural agents [32; 31; 30; 33] has shown that a single agent system is a feasible approach to the task of textual classification. In this paper, we develop a multiple agent system to explore the possibility of coupling symbolic transducer agents with the neural agents.

3 Symbolic Preference Moore Agent

3.1 Classification Material

One recent and well known news collection is the Reuters text classification test collection [17]. This corpus contains documents from the Reuters newswire. All news titles in the Reuters corpus belong to one or more of eight main categories: Money/Foreign Exchange (**money-fx**, **mf**), Shipping (**shipping**, **sh**), Interest Rates (**interest**, **in**), Economic Indicators (**economic**, **ec**), Currency (**currency**, **cr**), Corporate (**corporate**, **co**), Commodity (**commodity**, **cm**), Energy (**energy**, **en**). We use exactly all 10 733 titles of the so-called ModApte split of the Reuters corpus whose documents have a title and at least one associated topic category. 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.

3.2 Construction of Transducers from Regular Expression

The titles are symbolically tagged according to the most frequent occurrence of the tag for a particular word. This results in a sequence of tags of the form e.g. *(en cm cm co co)*, which represents a semantic tag sequence for a specific title.

Issues such as the exclusion of stop-words [32], stemming and possible loss of information due to the rounding have been considered previously; for example, in the case of the removal of stop-words (i.e. insignificant words such as 'the', 'a', 'and', etc., that may have an average distribution and are domain-independent across all categories), it was shown that there is only a little improvement in terms of classification accuracy. However, it can also be argued that in a semantic sequence, stop-words may indeed have an important influence

since they may be an indication of a unique sequence; for example, the 'of' in the phrase 'Bank of England', could bias the sequence towards 'England' if there are enough examples of the phrase itself in a set of titles.

A regular expression is a very specific way of defining a pattern, and hence the rules of finding that pattern. The regular expressions were incorporated into a finite-state transducer using [28]. For the discussion example shown in Figure 2, the regular expression would be denoted as $((0^*en^+mo^+0^*)^+)$, the '*' signifying that the symbol should appear either never or at least once in the sequence, the '$^+$' meaning that the symbol should appear at least once and the '0' simply standing for an arbitrary tag. The order of the sequence is from left to right. Therefore, the transducer would be able accept the sequence *(0 en mo 0 en mo 0)* but not *(0 mo en mo 0)* as it explicitly expects *(en)* at the start of a sequence followed by *(mo)* ultimately at the end of the sequence.

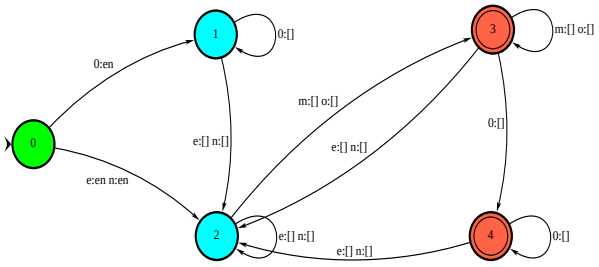

Figure 2: A transducer encoding the regular expression $((0^*en^+mo^+0^*)^+)$ for classifying a specific sequence of tags 'en' followed by 'mo' into the "energy" category. The tags must appear at least once in a stream of symbols interspersed with an arbitrary number of other tags - this transducer is more robust for sparser representations (e.g. the body of a newswire article or longer sequences from longer titles).

Figure 3 is a transducer able to handle sequences that are encoded by the regular expression $(co^*mf^*in^+in^*mf^+)^+)$ - this only accepts sequences that are ambiguous, but being explicitly in the "interest" category, one instance of the symbol *(in)*, followed by an arbitrary number of other *(in)* symbols, and one instance of the symbol *(mf)* must be present for correct transduction to the appropriate category.

3.3 Experimental Results

We give examples from 3 semantic classes to illustrate our experiments.

A Preference Moore Transducer was constructed that encoded the regular expression $(mf^*cr^*in^*en^*co^*cm^*ec^*sh^+sh^*ec^*mf^*cr^*in^*en^*co^*cm^*ec^*)$ for classifying the semantic sequence tags for the category "shipping". The symbolic input sequences were the tagged semantic titles described above. The transducer was able to achieve 84% classification of the "shipping" test sequences.

A similar Preference Moore Transducer was constructed for the category "energy"; 94% correct classification was achieved by the regular expression,

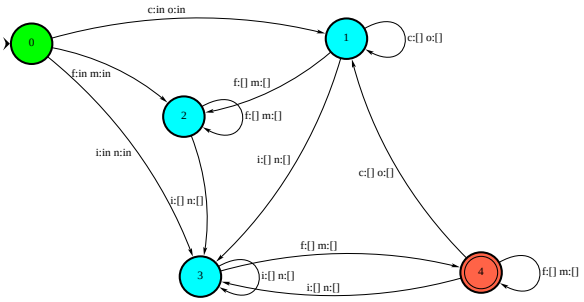

Figure 3: A transducer encoding the regular expression $((co^*mf^*in^+in^*mf^+)^+)$ for detecting a specific semantic sequence of tags 'in' followed explicitly by 'mf' that must appear at least once in a stream of tags interspersed with an arbitrary number of other symbols, in this example, "co" and "mf", to give a transduction to the "interest" category - this transducer is able to handle a denser representation of semantic sequences, with more specific rules, and also shorter sequences that have a very specific pattern or set of features

$(en^+co^*cr^*mf^*in^*ec^*co^*en^+)$. A Moore Machine for the category "commerce" achieved 100% classification performance using the regular expression $(mf^*cr^*in^*sh^*en^*cm^*co^+co^*mf^*cr^*in^*sh^*en^*cm^*)$. We were thus able to map the semantic relationships between sequences via such symbolic Moore transducers, which are able to maintain context during the mapping.

3.4 Evaluation of Results

In general, these initial simple symbolic machines perform reasonably well, given the simple representation. However, some semantic sequences from a particular category may wrongly be classified for several reasons - for example, the category allocations may depend on human-level interaction that does not take into consideration strict semantic representation but rather a more heuristic allocation to a particular category that may be arbitrary.

Also, the loss of information that may have taken place from the conversion of the numerical occurrences to their symbolic tags seemed to be minimal - the semantic tags did indeed preserve the appropriate level of information for acceptable figures of accuracy to be achieved using our symbolic Moore machines.

One basic heuristic in the construction of the regular expressions for the semantic sequences was to encode the presence of the category tag itself somewhere within the sequence - i.e. it was assumed that in general, sequences would be weighted towards having a greater number of the semantic tags belonging to that of the category itself.

The semantic sequences for the "commerce" category itself were shorter on average in the Reuters corpus, and hence it was less difficult to construct a transducer that would be able to confidently encode for the sequences.

Table 1 shows some representative results; regular expressions encoding the semantic preference rules were coded for each of the 8 main categories. Data-sets of the symbol-coded sequences were created. Using the specific Moore machine,

Category of Sequences	COmmerce Transducer	SHipping Transducer	ENergy Transducer
Commerce	**100%**	78%	64%
Shipping	10%	**84%**	10%
Energy	48%	30%	**94%**

Table 1: Performance of specific preference Moore machines with various input sequences of semantic categories; the bold figures (essentially the recall value for the transducers) show that the specific rules designed to handle the respective semantic sequences were performing well for the required categories.

we were able to achieve useful performance outputs. We cross-tested the respective data-set collections with the transducers for the other categories - and this indicated that the heuristic rules derived from the semantic sequences were indeed quite specific to the categories sequences themselves. As was expected, the "shipping" and "energy" transducers did not not work well with the "commerce" data, whereas "commerce" data with the shorter "commerce" semantic sequences achieved 100% performance. However, "shipping" and "energy" overlap with the category "commerce", hence the values were high (this is desirable as it shows that the regular expressions were indeed able to generalize). However, studying the results of the performance of the "shipping" transducer, it can be seen that "energy" and "commerce" are a weak subset of "shipping", or that the specificity of the "shipping" category is such that usually there is a very low degree of ambiguity.

Finally, we present the average recall and precision values for a symbolic preference Moore Machine in Table 2:

	Recall	Precision
SPMM	92.66%	58.91%

Table 2: Average recall and precision values for the 3 example transducers discussed.

With a simple symbolic preference Moore Machine, we can reach good recall values while the precision is still unsatisfactory.

4 Neural Preference Moore Agent

While a symbolic Preference Moore Machine is encoding top-down knowledge, a neural Preference Moore Machine is learning bottom-up. We briefly describe the various forms of neural Preference Moore machine used: a neural preference Moore machine with one context layer and a neural Preference Moore Machine with two hidden layers (see Figure 4) were trained using semantic vector representations at the input layer.

Input representations are obtained to represent the *plausibility* of a specific word occurring in a particular semantic category. The main advantage is that they are independent of the number of examples present in each category:

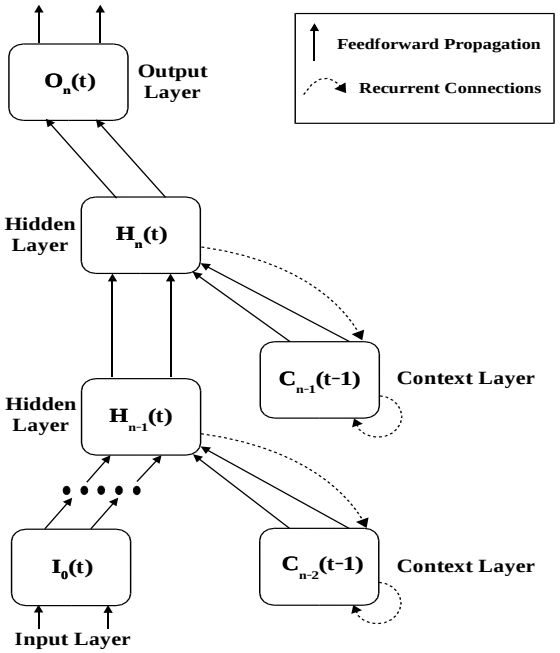

Figure 4: A Neural Preference Moore Machine with 2 hidden layers.

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

where:

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

The *normalized* frequency of the number of times a word w appears in a semantic category x_i (i.e. *the normalized category frequency*) was computed as a value $v(w, x_i)$ for each element of the semantic vector, divided by normalizing the frequency of the number of times a word w appears in the corpus (i.e. *the normalized corpus frequency*).

The performance of the best trained neural Preference Moore machines is shown in Table 3 with the recall and precision rates.

Evaluation	Recall	Precision
PMM 1 layer training	85.15	86.99
PMM 1 layer test	91.23	90.73
PMM 2 layers training	89.05	90.24
PMM 2 layers test	93.05	92.29

Table 3: Recall and precision for classifying newswire titles using neural preference Moore machines.

5 Discussion and Conclusion

We have described two different types of Preference Moore Machine agents - firstly, symbolic preference Moore machines based on finite-state automata theory which make use of transducers, and secondly, neural preference Moore machines based on the distributed learning of neural networks. We demonstrate that both approaches, though radically different in their computational paradigm, can indeed produce two related agent models that operate from a heuristically coded top-down supervisory mode, and from a bottom-up unsupervised mode.

Using the formalism that introduced Preference Moore Machine integration [29], we demonstrate the potential for integrating the different computational approaches on a standard real-world benchmarking corpus for the task of textual classification and information-mining.

The agent system is able to interact via the Preference Moore formalism and this allows different representations to be combined complementarily; the important gain is in terms of the robustness of the classification system as the transducer agent is able to code in a more abstract manner certain rules that the distributed recurrent agent may have missed; or conversely, the neural network agent may be able to correct the more difficult rules that are not easily encoded via regular expression syntax. The symbolic agent is better able to handle exceptions as manually coded expressions while neural classification agents would be able to handle difficult semantic sequences.

References

[1] N. M. Allinson and H. Yin. Interactive and semantic data visualisation using self-organizing maps. In *Proceedings of the IEE Colloquium on Neural Networks in Interactive Multimedia Systems*, 1998.

[2] M. Balabanovic and Y. Shoham. Learning information retrieval agents: Experiments with automated web browsing. In *Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford, CA*, 1995.

[3] M. Balabanovic, Y. Shoham, and Y. Yun. An adaptive agent for automated web browsing. Technical Report CS-TN-97-52, Stanford University, 1997.

[4] T. Briscoe. Co-evolution of language and of the language acquisition device. In *Proceedings of the Meeting of the Association for Computational Linguistics*, 1997.

[5] E. Charniak. *Statistical Language Learning*. MIT Press, Cambridge, MA, 1993.

[6] A. Cleeremans, D. Servan-Schreiber, and J. McClelland. Finite-state automata and simple recurrent networks. *Neural Computation*, 1:372–381, 1989.

[7] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In *International Conference on Tools for Artificial Intelligence*, Newport Beach, CA, November 1997.

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

[9] H. Cunningham, Y. Wilks, and R. Gaizauskas. New methods, current trends and software infrastructure for NLP. In *Proceedings of the NEMLAP-2*, Ankara, 1996.

[10] D. Freitag. Information extraction from html: Application of a general machine learning approach. In *National Conference on Artificial Intelligence*, pages 517–523, Madison, Wisconsin, 1998.

[11] T. Honkela. Self-organizing maps in symbol processing. In S.Wermter and R.Sun, editors, *Hybrid Neural Systems*. Springer, Heidelberg, Germany, 2000.

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

[13] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. WEBSOM - self-organizing maps of document collections. *Neurocomputing*, 21:101–117, 1998.

[14] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 1995.

[15] T. Kohonen. Self-organisation of very large document collections: State of the art. In *Proceedings of the International Conference on Aritricial Neural Networks*, pages 65–74, Skovde, Sweden, 1998.

[16] Stefan C. Kremer. On the computational power of Elman-style recurrent networks. *IEEE Transactions on Neural Networks*, 6(4):1000–1004, July 1995.

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

[18] A. McCallum, R. Rosenfeld, T. Mitchell, and A. Y. Ng. Improving text classification by shrinkage in a hierarchy of classes. In *Proceeedings of the 15th International Conference on Machine Learning*, pages 359–367, San Francisco, CA, 1998.

[19] F. Menczer, R. Belew, and W. Willuhn. Artificial life applied to adaptive information agents. In *Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995.

[20] Kazuhisa Niki. Self-organizing information retrieval system on the web: SirWeb. In Nikola Kasabov, Robert Kozma, Kitty Ko, Robert O'Shea, George Coghill, and Tom Gedeon, editors, *Progress in Connectionsist-Based Information Systems. Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems*, volume 2, pages 881–884. Springer, Singapore, 1997.

[21] R. Papka, J. P. Callan, and A. G. Barto. Text-based information retrieval using exponentiated gradient descent. In M. C. Mozer, M. I. Jordan, and T. Petsche, ed-

itors, *Advances in Neural Information Processing Systems*, volume 9. The MIT Press, 1997.

[22] M. Perkowitz and O. Etzioni. Adaptive web sites: an AI challenge. In *International Joint Conference on Artificial Intelligence*, Nagoya, Japan, 1997.

[23] M. Sahami, M. Hearst, and E. Saund. Applying the multiple cause mixture model to text categorization. Technical report, AAAI Spring Symposium on Machine Learning in Information Access, 1996.

[24] H. Schuetze, D. A. Hull, and J. O. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Proceedings of the Special Interest Group on Information Retrieval*, 1995.

[25] D. Servan-Schreiber, A. Cleeremans, and J. L. McClelland. Encoding sequential structure in simple recurrent networks. Technical Report Technical Report CMU-CS-88-183, Carnegie Mellon University, Pittsburgh, PA, 1988.

[26] N. Sharkey and A. Sharkey. Separating learning and representation. In S. Wermter, E. Riloff, and G. Scheler, editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 17–32. Springer, 1996.

[27] R. Sun and T. Peterson. Multi-agent reinforcement learning: Weighting and partitioning. *Neural Networks*, 1999.

[28] G. van Noord. FSA utilities: A toolbox to manipulate finite-state automata. In Darrell Raymond, Derick Wood, and Sheng Yu, editors, *Automata Implementation*, pages 87–108. Lecture Notes in Computer Science 1260, Springer Verlag, 1997.

[29] S. Wermter. Preference Moore machines for neural fuzzy integration. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 840–845, Stockholm, 1999.

[30] S. Wermter, G. Arevian, and C. Panchev. Recurrent neural network learning for text routing. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 898–903, Edinburgh, UK, 1999.

[31] S. Wermter, G. Arevian, and C. Panchev. Network analysis in a neural learning internet agent. In *Proceedings of the International Conference on Computational Intelligence and Neurosciences*, pages 880–884, Atlantic City, PA, USA, 2000.

[32] S. Wermter, C. Panchev, and G. Arevian. Hybrid neural plausibility networks for news agents. In *Proceedings of the National Conference on Artificial Intelligence*, pages 93–98, Orlando, USA, 1999.

[33] S. Wermter and R. Sun. *Hybrid Neural Systems*. Springer, Heidelberg, 2000.