A Hybrid and Connectionist Architecture for a SCANning Understanding

Stefan Wermter*
Universität Hamburg, FB Informatik,
Bodenstedtstr. 16, W-2000 Hamburg 50, Germany
Email: wermter@nats4.informatik.uni-hamburg.de

Abstract. This paper describes a general architecture SCAN for hybrid symbolic connectionist processing of natural language phrases. SCAN's architecture shows how learned connectionist domain-dependent semantic representations can be combined with encoded symbolic syntactic representations. Within this general architecture we focus on a connectionist model for semantic classification based on a scanning understanding of phrases. We specify strategies at the topmost theory level and we show how these strategies are realized in a recurrent connectionist plausibility network at the underlying representation level. In particular, this model demonstrates that a recurrent connectionist network can *learn* a semantic memory model for phrase classification based on a scanning understanding.

1 Introduction

In the past the debate about the use of symbolic versus connectionist representations has shown strong arguments for both perspectives. From a strictly symbolic perspective, connectionist representations just take the role of implementing symbolic processes at a lower level, and connectionist implementations do not lead to new fundamental results for cognitive science and artificial intelligence (e.g., [8]). On the other hand, from a strictly connectionist perspective, connectionist representations are most appropriate for cognitive science and artificial intelligence, and symbolic interpretations only emerge from connectionist representations at a higher level (e.g., [7]).

However there is recent evidence that it is advantageous to combine symbolic and connectionist processing [2]. So far different hybrid models have been proposed for sentence analysis and for inferencing (e.g., [4] [5] [12]). In this paper we will describe a hybrid architecture for a scanning understanding of phrases which is based on Marr's general framework for artificial intelligence models [6]. In particular, we will focus on a model for learning a semantic classification of phrases as they occur in real-world book titles. Typical phrases are for instance: "Learning to use the spss batch system", "Investigation of copper and nickel after high-energy implantation of helium atoms".

We will describe plausibility judgements, reading strategies, and different constraints at the computational theory level and their realization in a recurrent plausibility network at the representation level. We will examine the learning and generalization in plausibility networks and we will analyze the learned internal representation of classified phrases with respect to uncertain class assignment, lexical ambiguity, and context learning.

2 Overview of SCAN - A Hybrid Architecture for a Scanning Understanding of Phrases

In contrast to an in-depth understanding, a scanning understanding focuses on most important syntactic and semantic properties of words and phrases. SCAN is a hybrid architecture for a scanning understanding of phrases [15] and figure 1 shows the general architecture organized at the levels of computational theory and representation based on Marr's general framework for artificial intelligence models [6]. At the theory level of SCAN's architecture, the goals and concepts of a task are specified based on plausibility judgements, reading strategies, and different constraints. At the representation level various symbolic and connectionist representations can be combined. While symbolic representations in SCAN are used for *encoding* primarily domain-independent knowledge, connectionist representations are used for *learning* domain-dependent semantic knowledge.

*This research was supported in part by the Federal Secretary for Research and Technology under contract #01IV101A/O.

ECAI 92. *10th European Conference on Artificial Intelligence* Edited by B. Neumann
Published in 1992 by John Wiley & Sons, Ltd

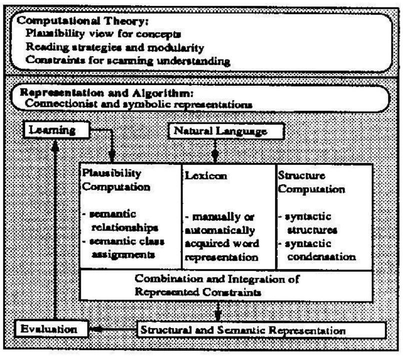

Figure 1: Overview of SCAN

In previous work we examined this hybrid architecture for structural disambiguation tasks [13] [14]. In this paper we focus on the task of learning a semantic classification of unrestricted phrases. In particular, we concentrate on the *plausibility computation* for semantic class assignment and the *learning* of a semantic representation of phrases during semantic classification.

3 Strategies for Semantic Classification at the Theory Level

There has been evidence that *plausibility judgements* are important for semantic memory models. For instance, it has been argued that uncertain fuzzy plausibility judgements are more appropriate for semantic memory models than propositional matches of purely symbolic processes [9]. In general, a concept representation based on plausibility is preferred to a definition of a set of necessary and sufficient discrete symbolic properties.

Furthermore, specific tasks influence *reading strategies*. For instance, in a detailed study it has been shown that reading for immediate recall requires relatively more structural processing while reading for comprehension requires more semantic processing [1]. This provides support for different structure-oriented and semantics-oriented modules for an architecture for a scanning understanding of phrases.

Last, various syntactic, semantic, and contextual *constraints* influence semantic classification. For instance, syntactic constraints in its most simple form can emphasize the significance of nouns versus other more domain-independent syntactic categories (e.g., prepositions, determiners) which provide less significant knowledge for semantic classification.

Other important constraints for semantic classification are supplied by the sequential context in a phrase. Only the context in a phrase allows a *lexical disambiguation* of different contextual senses of the same word. Therefore, the preceding incremental context should be used to constrain the interpretation of a current word. Furthermore, the sequential context is also important for making incremental predictions for class assignments which may be modified later when more context has been seen.

4 Plausibility Networks for Semantic Classification at the Representation Level

In order to learn and represent classifications of title phrases we used recurrent connectionist plausibility networks. The general design of a recurrent plausibility network is shown in figure 2. A feedforward network with $n+1$ layers L_0 to L_n is extended with recurrent layers at each hidden layer L_1 to L_{n-1}. Each arrow in figure 2 describes a fully connected $x:y$ relationship between x units of one layer and y units of the connected layer.

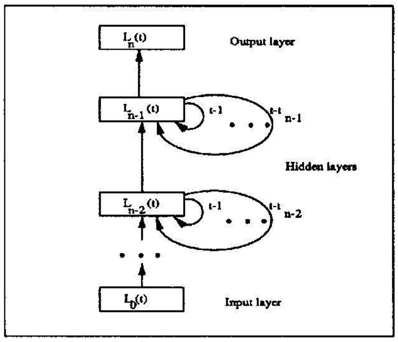

Figure 2: General structure of a recurrent plausibility network

The recurrent layers can exist for different preceding time steps $t-1$ to $t-t_{max}$ where max is the maximum preceding time step at that layer. These recurrent layers realize distributed delays of previous states of the network at different time steps and enable the network to represent preceding context. This general recurrent plausibility network extends the concept of simple recurrent networks [3] and time delay networks [11] by introducing unrestricted distributed delays at different layers. By representing previous internal states in additional recurrent layers it is possible to use the backpropagation learning rule for training [10].

4.1 Experiments for Semantic Classification

Our semantics-oriented strategies for classifying unrestricted book titles are based on a plausibility view of concept representation. One main point of the plausibility view is that a concept is not described with necessary and sufficient properties but based on properties which are *plausible*. According to this plausibility view single words, relationships between words, class membership of phrases can be learned and represented based on continuous values rather than discrete symbols. For the task of learning a semantic classification we used unrestricted classified title phrases from a University library and we represented each word in a title with a significance vector $(c_1 c_2 \cdots c_n)$ where c_i represents a certain class dimension. A *significance value* $v(w, c_i)$ is computed for each class dimension as the frequency of occurrences of word w in class c_i (the class frequency) divided by the frequency of occurrences of word w in the corpus (the corpus frequency).

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

These significance vectors were computed based on a corpus of 30206 words from 10 classes of a real-world library classification. The occurring classes were: theology/religion (TR), history/politics (HP), law (LA), mathematics (MA), chemistry (CH), computer science (CS), electrical engineering (EE), materials/geology (MG), art/architecture (AA), and music (MU). Each title was represented with its sequence of *semantic and plausible significance vectors* for its words.

For our classification exeriments we designed a plausibility network with three layers and one distributed delay at the second layer. This particular plausibility network is similar to a simple recurrent network [3] but uses plausible significance vectors for a classification task rather than artificially generated vector representations for a prediction task. In our experiments for learning and generalizing a semantic classification we used 2000 unrestricted titles from 10 classes, 1000 title phrases for training and 1000 different titles with different representations for testing the generalization behavior. The significance vectors of a title were presented sequentially one by one to the input layer of the plausibility network. The input layer had 10 units, one for each class dimension of a significance vector. The output layer had 10 units for specifying the desired class of a title phrase. A network with 10 hidden units showed the best performance on training and test set. This network was trained for 200 epochs with the complete training set with a learning rate of 0.000001. This small training rate prevented the network from changing the weights too fast for this relatively big number of phrases. Then we increased the learning rate to 0.00001 in order to speed up learning and trained the network for another 200 epochs. The summarized performance of the network is illustrated in table 1.

Evaluation after each	Error rate for recurrent plausibility network	Error rate for average significance vectors
training word	16.4%	49.4%
test word	18.3%	51.5%
training title	2.4%	18.0%
test title	5.5%	22.2%

Table 1: Semantic classification of unrestricted phrases

The first column shows the results for the recurrent plausibility network. As a comparison the second column shows the results for a classification based on the average of the significance vectors of a title. A class assignment is considered correct if the single output unit of the desired class is activated. We can see that for training and test set, the recurrent plausibility network performs better than the average significance vector representation because the plausibility network contains knowledge about the actual sequence and context in a title. Furthermore, the error rates for the plausibility network are lower at the end of a complete phrase: only 2.4% of the training titles and 5.5% of the unknown test titles could not be classified correctly, in other words 97.6% of the 1000 training titles and 94.5% of the 1000 unknown test titles were classified correctly.

In a second set of experiments we tested the influence of a simple syntactic heuristic for a reduction to most significant content words. We eliminated words that belonged to the syntactic categories of prepositions (e.g., "of"), conjunctions (e.g., "and"), personal pronous (e.g., "his"), or determiners (e.g., "the") if the word occurred four times or more. Then, we performed the same training and testing with the reduced titles using the same parameters and architecture as for the complete titles.

Evaluation after each	Error rate for recurrent plausibility network	Error rate for average significance vectors
training word	9.5%	25.1%
test word	13.6%	30.4%
reduced training title	1.9%	18.0%
reduced test title	5.7%	25.7%

Table 2: Semantic classification of unrestricted phrases without insignificant words

Similar as for the complete titles, table 2 shows that the plausibility network performed better than the classification based on the average significance vector since the learned preceding context improved classification performance. The elimination of insignificant words led to a slight improvement on the training set (error rate 1.9% versus 2.4%) but a slight deterioration on the test set (5.7% versus 5.5%). The reduction to significant content words makes training easier because there are less words with ambiguous class assignments. On the other hand, complete titles show a slightly better generalization performance due to more ambiguous insignificant words. In general, the percentages of the classification performance with and without insignificant words are rather close (both about 98% correct for training and 94% correct for generalizing to completely new phrases).

4.2 Analysis of Constraints in Learned Internal Representation

After training we examined the dynamics of the hidden layer as the learned internal representation of the plausibility network. Figure 3 shows titles with their internal representation of the hidden units and their incremental class assignment. The internal representation which contains the dynamics of incremental class assignment for phrases is distributed over collections of hidden units, and single units can contribute to different class assignments. For instance, in cooperation with other units the first hidden unit can participate in the class assignment to the AA class (examples 1 and 3), to the CH class (4) and to the MG class (6).

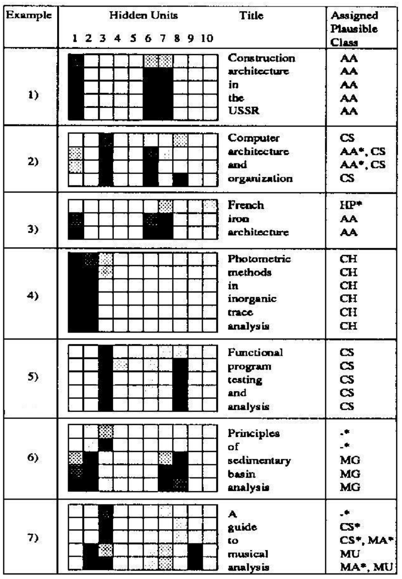

Figure 3: Internal representation of unrestricted phrases

Focusing on the individual titles, the first example has the correct class assignment for AA right from the beginning since "construction" is significant for the AA class. The second title is initially assigned to the CS class but after "computer architecture" two classes are assigned, CS correctly but AA incorrectly (marked with the sign "*"). Only the word "organization" provides more evidence for the final correct class assignment CS. The third title is another example for initial incorrect *expectations for class assignment* which have to be modified later when more specific knowledge is available. Here, "French" provides weak support for the HP class which is later modified to the correct AA class.

Uncertain class assignments (marked with "-") occur if no output unit is active. The last two examples in figure 3 illustrate that the plausibility network may initially be uncertain about a class assignment. This can be seen in the low values of the hidden layer as well as in the low values for the output units, for instance for "Principles of...". This start of a title is not yet significant enough for a class assignment. Our last example in figure 3 shows one of the very few titles with an incorrect final class assignment due to the insignificant preceding context ("a guide to") and two competing significant subsequent words ("musical analysis").

Finally, the first three examples show the incremental internal representation and class assignment for three titles which contain the word "architecture". Similarly, the fourth to seventh title all contain the word "analysis". Nevertheless different classes are assigned even at the same word, for instance in the first three examples for the word "architecture". The preceding context allows for a *lexical disambiguation* between different forms of "architecture": construction architecture, computer architecture, iron architecture. This demonstrates that the *plausibility networks learn to represent the preceding context*. Without a different internal representation of the preceding context, different classes could not be assigned for the same word.

5 Conclusion

We have outlined a general hybrid architecture SCAN for a scanning understanding of phrases which combines symbolic and connectionist processing. Within the framework of this hybrid architecture we have analyzed a model for semantic classification of phrases. From a perspective of connectionist models, recurrent plausibility networks extend concepts from simple recurrent networks and time delay networks and they can be embedded into a bigger hybrid architecture. From a natural language perspective, difficult problems for classification, like uncertain class assignment, correction of initial misclassification, context representation, and lexical disambiguation can be learned and generalized from a real-world corpus. We conclude that connectionist plausibility networks can learn and represent a semantic memory model for semantic classification within a hybrid symbolic connectionist architecture for a scanning understanding of phrases.

References

[1] D. Aaronson and S. Ferres. Reading strategies for children and adults: a quantitative model. *Psychological Review*, 93(1):89–112, 1986.

[2] M. G. Dyer. Symbolic neuroengineering for natural language processing: a multilevel research approach. In J. A. Barnden and J. B. Pollack, editors, *Advances in Connectionist and Neural Computation Theory, Vol.1: High Level Connectionist Models*. Ablex Publishing Corporation, Norwood, NJ, 1991.

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

[4] J. A. Hendler. Marker passing over microfeatures: towards a hybrid symbolic/connectionist model. *Cognitive Science*, 13:79–106, 1989.

[5] W. G. Lehnert. Symbolic/subsymbolic sentence analysis: exploiting the best of two worlds. In J. A. Barnden and J. B. Pollack, editors, *Advances in Connectionist and Neural Computation Theory, Vol.1: High Level Connectionist Models*. Ablex Publishing Corporation, Norwood, NJ, 1991.

[6] D. Marr. *Vision*. Freeman, San Francisco, 1982.

[7] J. L. McClelland, D. E. Rumelhart, and G. E. Hinton. The appeal of parallel distributed processing. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume Vol. 1, pages 3–44. MIT Press, Cambridge, MA, 1986.

[8] S. Pinker and A. Prince. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. In S. Pinker and J. Mehler, editors, *Connections and Symbols*, pages 73–193. MIT Press, Cambridge, MA, 1988.

[9] L. M. Reder. Plausibility judgements versus fact retrieval: alternative strategies for sentence verification. *Psychological Review*, 89(3):250–280, 1982.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume Vol. 1, pages 318–362. MIT Press, Cambridge, MA, 1986.

[11] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37 (3):328–339, 1989.

[12] D. L. Waltz and J. B. Pollack. Massively parallel parsing: a strongly interactive model of natural language interpretation. *Cognitive Science*, 9:51–74, 1985.

[13] S. Wermter. Integration of semantic and syntactic constraints for structural noun phrase disambiguation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1486–1491, Detroit, MI, 1989.

[14] S. Wermter. Combining symbolic and connectionist techniques for coordination in natural language. In Marburger, editor, *Proceedings of the 14th German Workshop on Artificial Intelligence*, pages 186–195, Schloß Eringerfeld, FRG, 1990.

[15] S. Wermter. Scanning understanding: A symbolic connectionist approach for natural language phrases. Technical Report (forthcoming), University of Hamburg, Hamburg, FRG, 1992.