Meaning Spotting and Robustness of Recurrent Networks

Stefan Wermter, Christo Panchev and Garen Arevian
The Informatics Centre, SCET, University of Sunderland
St. Peter's Way, Sunderland SR6 0DD, United Kingdom
email: stefan.wermter@sunderland.ac.uk

Abstract

This paper describes and evaluates the behavior of preference-based recurrent networks which process text sequences. First, we train a recurrent plausibility network to learn a semantic classification of the Reuters news title corpus. Then we analyze the robustness and incremental learning behavior of these networks in more detail. We demonstrate that these recurrent networks use their recurrent connections to support incremental processing. In particular, we compare the performance of the real title models with reversed title models and even random title models. We find that the recurrent networks can, even under these severe conditions, provide good classification results. We claim that previous context in recurrent connections and a meaning spotting strategy are pursued by the network which supports this robust processing.

1 Introduction

In the last decade, there has been a lot of work demonstrating that neural networks can be used for a wide variety of problem domains and tasks [Giles and Omlin, 1993, Hendler, 1991, Honavar, 1995, Miikkulainen, 1993, Wermter, 1995]. Initial models were often criticized for working only on small problems or that they did not analyze the general conditions under which a network would perform. Recently, recurrent neural networks [Wermter, 2000, Moisl, 1992], along with self-organizing maps and reinforcement learning approaches have gained a greater interest. In this paper, we address robustness in recurrent networks, in particular for a task in semantic classification.

The task of textual classification is an important research area, for instance considering the huge numbers of webpages and text that can be classified [Craven et al., 1998, Kohonen, 1998, Niki, 1997, Balabanovic and Shoham, 1995]. Different approaches from the fields of Artificial Intelligence, Neural Networks and Information Retrieval have been explored recently [Joachims, 1998, Freitag, 1998, McCallum et al., 1998, Cooley et al., 1997, Liere and Tadepalli, 1996, Cohen, 1996]. However, there are still important deficiencies in these systems such as lack of scalability, robustness, adaptiveness, autonomousness and so on. More sophisticated neural network approaches, such a modular networks, hybrid systems and learning software agents have only been applied recently to this task.

In this paper, we describe recurrent neural networks which are analyzed in terms of their robust behavior. In previous work, we have designed a model for semantic classification of real world news titles [Wermter et al., 1999]. While this paper focused on the initial design of the recurrent plausibility networks, we now have performed new extensive experiments which shed light on the performance of such networks under different, more unconstrained conditions; this allows the testing of the robust performance of our architecture, and the testing of the behavior of the network against a variety of differing inputs.

2 Setting the Scenery: Input Representations and the Model

Recurrent plausibility networks (RPN) [Wermter, 1995], developed from simple recurrent networks (SRN) [Elman, 1990], are a class of networks which only have partial feedback; in the simplest case, where a network has an input layer, an output layer and a hidden layer, there is a feedback from the internal hidden layer to a context layer that acts as a "buffer". In our RPN, the specific architecture is one with two hidden layers whose feedback connections go to two context layers, which have connections back to the same hidden layer and also have self-recurrency. Figure 1 shows the network architecture. For a more formal definition of this class of architectures, we refer to [Wermter, 1995, Wermter et al., 1999].

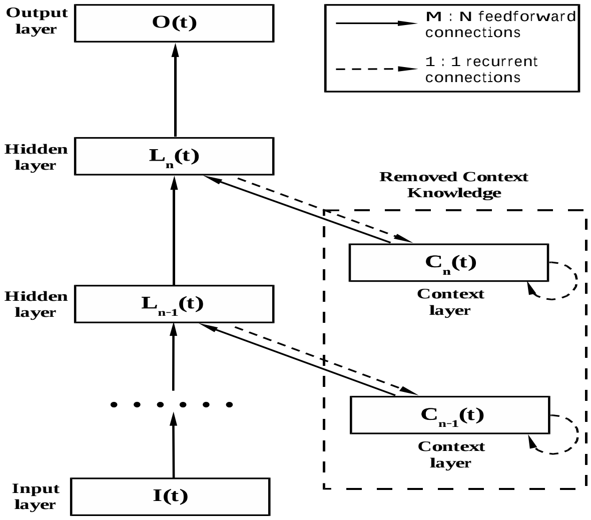

Figure 1: Recurrent plausibility network.

The RPN was trained and tested with several variations of input vectors which represent the sequence of words from real-world grammatical titles. In the experiments described, we concentrate on semantic vectors as a preprocessing strategy. We represent semantic vectors as the *plausibility* of a specific word occurring in a particular semantic category, the main advantage being that they are independent of the number of examples present in each category:

$$v(w, x_i) = \frac{Normalized\ frequency\ of\ w\ in\ x_i}{\sum\limits_j Normalized\ frequency\ of\ w\ in\ x_j},\ j \in \{1, \cdots n\}$$

where:

$$Normalized\ frequency\ of\ w\ in\ x_i = \frac{Frequency\ of\ w\ in\ x_i}{Number\ of\ titles\ in\ x_i}$$

The *normalized* frequency of occurrences of a word w 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 occurrences of a word w in the corpus (i.e. *the normalized corpus frequency*).

3 Testing for Robustness: Training with Randomized Sequences

Several sets of experiments were performed. First, the RPN was trained with titles from the corpus with the vectors representing the original word order. Then, the word-orders in the same titles were randomized and a second network trained. Furthermore, an SRN was also trained with the same randomized sequences to allow comparison of the performance between the SRN and RPN. All networks were trained with the same parameters for 1000 epochs. For the training of randomized word order in the titles, the word order was changed every 100 epochs. All networks were tested with a test set containing titles with the correct word order. The percentages for the network's recall and precision values can be seen in Table 1.

We used titles from the real-world Reuters corpus data [Lewis, 1997]. All documents in this corpus have been pre-classified into several specific categories; the news titles belong to one or more of eight main categories: “money/foreign-exchange”, “shipping”, “interest rates”, “economic indicators”, “currency”, “corporate”, “commodity” and “energy”. The corpus is composed of 10,733 titles of which one-tenth were used for the training set, the remainder to test generalization with the unseen examples.

In the first row of Table 1, we see the total recall and precision values of the test-set for titles in the original order of words. The RPN performance for the randomized title sequences is relatively higher when compared to the figures for the SRN, suggesting that the *two* context layers in the RPN act as an important element for the embodiment of contextual information, allowing robust generalization.

Category	Test set	
	recall	precision
RPN trained on original word-order of titles	**93.05**	**92.29**
RPN trained on randomized word-order of titles	**93.68**	**92.79**
SRN trained on randomized word-order of titles	**92.61**	**91.30**
RPN trained on original word-order *with context layers "removed"*	**55.98**	**55.75**

Table 1: A selection of recall/precision values for various recurrent networks trained on several word-order configurations

To further benchmark the above results, we conducted experiments where the context layers were removed (see Figure 1), essentially reducing the network to a feedforward one. The results show the degraded performance on the same input vectors. This demonstrates that the network is doing something like meaning spotting and classifies based on context but not necessarily the individual grammatical order of words. This also clearly demonstrates the importance of the context in the recurrent networks since the performance drops significantly without this context.

4 Deeper Analysis of an Example

In this section, we describe a particular example in more detail to analyze the learning behavior. The title explored is "Iran Soviet Union to swap crude, refined products", chosen as it is slightly ambiguous in terms of the classification category to which it belongs. We use this title to demonstrate the behavior and robustness of the network and the need for context layers.

The important keywords in this case that influence correct classification of the title, are "Iran" and "Soviet Union" and "crude". The word "crude" has a strong association with the correct category "energy", while "Iran" and "Soviet Union" have a slightly stronger preference for the wrong (in this case) "shipping" category. The word "product" has a relatively strong preference for the incorrect categories, namely "energy", "commodity" and "economics".

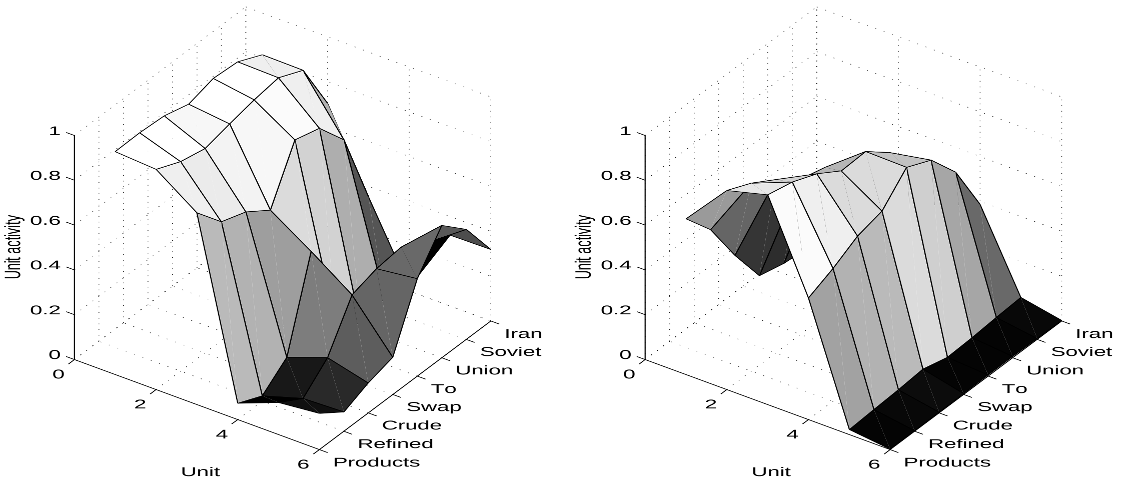

Figure 2: First and second context layers with *original/correct word-order* plotted against activation of units which have been ordered from highest to lowest activity

4.1 Input of Original Phrase in Correct Grammatical Order

Figure 2 shows the performance of both the first context and second context layers respectively using the correct and unchanged word order. The first context layer, as can be expected, shows a greater degree of fluctuation in its inner representation, but the second context layer shows a clear preference for the correct classification. It can be seen that the first context layer is more dynamic than the second. The misleading keywords are at the beginning of the sequence, so classification is relatively inaccurate. However, the word “products” at the end of the title causes correct classification. This is a demonstration of the abilities of the model to quickly react to keywords.

4.2 Input of Phrase in Reverse Word Order

To test the hypothesis that the network is robust against possible biases that may be introduced from word order, the same title was initially reversed to “Products refined crude swap to Union Soviet Iran”. Looking at Figure 3, it can be seen that the behavior of the first context layer is again fairly dynamic; however, comparison with the second layer shows that again the network has learned the correct inner representation of the context, and hence there is correct class assignment. This is a desirable result, as the keyword “crude” that allows correct classification occurs towards the beginning of the title, demonstrating a longer-term memory for the context.

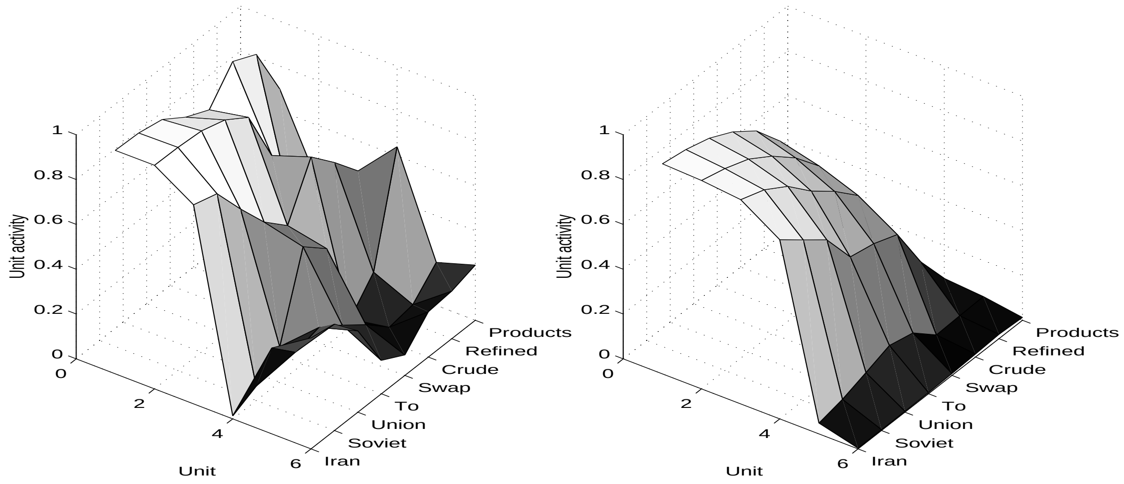

Figure 3: First and second context layers with *reversed word-order* plotted against activation of units which have been ordered from highest to lowest activity

4.3 Input of Phrase with Random Word Order

Finally, a random sequence of the input representations was chosen and tested (in this case, “Products Iran swap Union to crude refined Soviet”). Figure 4 shows the same general behavior of the activations as with the two previous sets; the second context layer is more stable in the inner representation of the context. In this example, the word “crude” is towards the end - this is interesting since there is no clear preference for any of the classes from this word. However, the classification is still correct at the end due to the incremental context.

5 Discussion and Conclusions

From the experiments above, we can derive several points which give us new insight into the classification and generalization behavior of the various recurrent networks: namely that they are robust, rely on previous

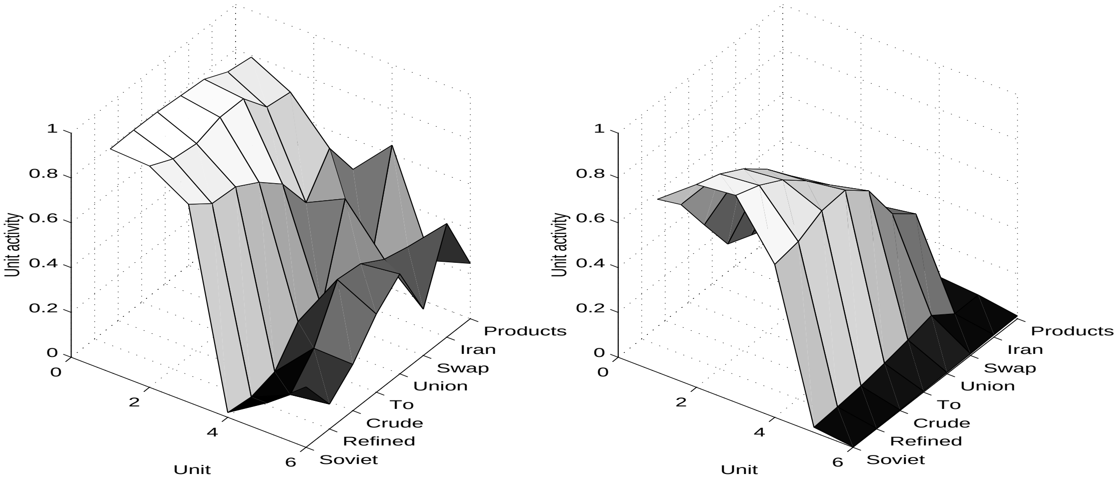

Figure 4: First and second context layers with *random word-order* plotted against activation of units which have been ordered from highest to lowest activity

context and can be very useful for classification tasks of word sequences, such as those from our real-world Reuters corpus.

The recurrent plausibility network is shown to be very robust against input variations such as ordering of the input vectors that represent words in a sequential title: all the experiments demonstrate that the various configurations of word-vector representations do not cause a degradation in the classification accuracy. This indicates that the network is picking up context between various keywords, which in conjunction with other weighted keywords, allow the formation of contextual information from the input streams. The incremental previous context, with clear preferences, leads to good final preferences for this task.

We also demonstrate that the first context layers in all cases are more dynamic and that the second context layers act as more stable, long-term memory. It is indeed the context layers that allow incremental context to be built-up during training. Randomizing the input vector representations did not have any significant effect on classification accuracy, showing that the actual grammatical word-order is less important; parallel experiments using an SRN and an RPN with their context layers removed showed that the RPN performed better. In conclusion, we have shown that RPN networks can be used effectively, especially for robust classification.

References

[Balabanovic and Shoham, 1995] Balabanovic, M. and Shoham, Y. (1995). 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.*

[Cohen, 1996] Cohen, W. (1996). Learning rules that classify e-mail. In *AAAI Spring Symposium on Machine Learning in Information Access*, Stanford, CA.

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

[Craven et al., 1998] 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*, Madison, WI.

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

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

[Giles and Omlin, 1993] Giles, 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, 1991] Hendler, J. (1991). Developing hybrid symbolic/connectionist models. In Barnden, J. A. and Pollack, J. B., editors, *Advances in Connectionist and Neural Computation Theory, Vol.1: High Level Connectionist Models*, pages 165–179. Ablex Publishing Corporation, Norwood, NJ.

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

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

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

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

[Liere and Tadepalli, 1996] Liere, R. and Tadepalli, P. (1996). The use of active learning in text categorisation. In *AAAI Spring Symposium on Machine Learning in Information Access*, Stanford, CA.

[McCallum et al., 1998] McCallum, A., Rosenfeld, R., Mitchell, T., and Ng, A. Y. (1998). 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.

[Miikkulainen, 1993] Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing.* MIT Press, Cambridge, MA.

[Moisl, 1992] Moisl, H. (1992). Connectionist finite state natural language processing. *Connection Science*, 4:67–91.

[Niki, 1997] Niki, K. (1997). Self-organizing information retrieval system on the web: SirWeb. In Kasabov, N., Kozma, R., Ko, K., O'Shea, R., Coghill, G., and Gedeon, T., 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.

[Wermter, 1995] Wermter, S. (1995). *Hybrid Connectionist Natural Language Processing.* Chapman and Hall, Thomson International, London, UK.

[Wermter, 2000] Wermter, S. (2000). The hybrid approach to artificial neural network-based language processing. In Dale, R., Moisl, H., and Somers, H., editors, *A Handbook of Natural Language Processing.* Marcel Dekker.

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