Recurrent Neural Network Learning for Text Routing

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

1 Abstract

This paper describes new recurrent plausibility networks with internal recurrent hysteresis connections. These recurrent connections in multiple layers encode the sequential context of word sequences. We show how these networks can support text routing of noisy newswire titles according to different given categories. We demonstrate the potential of these networks using an 82 339 word corpus from the Reuters newswire, reaching recall and precision rates above 92%. In addition, we carefully analyze the internal representation using cluster analysis and output representations using a new surface error technique. In general, based on the current recall and precision performance, as well as the detailed analysis, we show that recurrent plausibility networks hold a lot of potential for developing learning and robust newswire agents for the internet.

2 Introduction

With the recent exponential expansion of the internet, a need has arisen to design more sophisticated learning agents which are capable of classifying relevant information. Much initial work in the field of internet agents has used manual encoding techniques or simple techniques from information retrieval [12].
However, it becomes increasingly apparent that automatic adaptation, learning, dealing with incompleteness and robustness are necessary requirements [15, 14]. Recently, there has been a new focus on neural network learning techniques and text processing, for instance for newswires and world wide web documents [3]. However, most internet agents, such as classifiers, search engines and extractors still use ad hoc heuristic coding.
In this paper, we will present an analysis of learning agents with an emphasis on agents supported by neural network learning. We will describe experiments from a Hybrid Neural agent for Text routing (HyNeT). In particular, we explore recurrent plausibility networks with a dynamic short-term memory which allows the processing of sequences in a robust manner. Furthermore, we show a detailed analysis of the internal representation. The work presented here shows the potential of using neural techniques within an intelligent agent for semantic text routing.

3 Learning Agents

Several interesting proposals have been put forward to encourage a formalization of meta-data information in web pages [3, 5]. The main idea is to start from web pages that already have some form of meta-data about other pages. A system that automatically updates such knowledge has been constructed by making use of the special structure of hyperlinks and words. However, as has been pointed out [10], there are several limitations to manually encoded web meta-data that is a formal part of a page: this approach is difficult to implement and needs a long time to become established as part of general web-building practice [10]. Also, such a meta-data system will rely on manually encoded symbolic knowledge for classification.

Statistical techniques have been shown to perform successfully in the classification of text [2]. 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 [1]; however, these statistical methods are very data-intensive.
A self-organizing map forms a non-linear projection from a high-dimensional space onto low-dimensional space and has been used in the WEBSOM project [8]. 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 grid.

4 Recurrent Networks

In this paper, we will examine the use of recurrent neural networks for text routing. Partially recurrent networks, such as simple recurrent networks, have recurrent connections between the hidden layer and context layer [4]; Jordan networks have connections between the output and context layer [6]. However, using just simple recurrent processing, the information about states decays rapidly.
For the type of networks described in [6], self-recurrent connections within each of the context layers are used to control the decay within the network. Other work [13] has focused on the introduction of different decay memories by using distributed recurrent delays over the separate context layers that represent the contexts at various time steps. The current input and the incremental contexts from the preceding $n-1$ time steps are processed by such a network with n hidden layers.
The general structure of our recurrent plausibility network is shown in Figure 1. Supervised learning [7] is used to train this recurrent network. The distributed context layers and self-recurrent connections of the context layers are combined with the general features of recurrent networks.
The underlying layer L_{n-1}, as well as the incremental context layer C_n both constrain the input to a hidden layer L_n. The activation of a unit $L_{ni}(t)$ at time t is computed on the basis of the weighted activation of

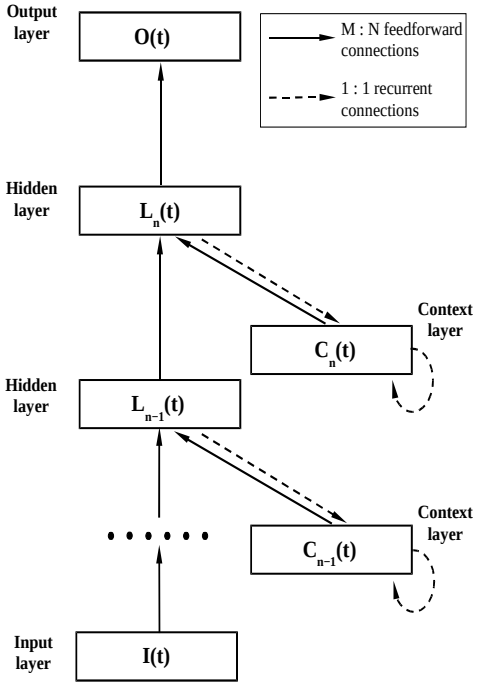

Figure 1: Recurrent plausibility network.

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.

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

A time-averaging of the information is performed by the units in the context layers, 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. The hysteresis value φ_n represents the self-recurrent connections. The context layer C_{n-1} has a hysteresis value which is lower than the hysteresis value of the next context layer C_n. Hence, the context layers nearer the input layer act as memory with a more dynamic context. As the upper context layers have hysteresis values which are higher, being nearer the output layer, they incrementally build more stable sequential memories. This results in a larger context which arises from the more recent and dynamic one.

Semantic Category	Title
money-fx	U.K. Money Market Deficit Forecast Revised Upwards
ship	Japanese Shipyards to Form Cartel, Cut Output
interest	Top Discount Rate at U.K. Bill Tender Falls to 9.1250 pct
economic	Korea Plans to Open Markets to Ease Won Pressure
currency	Japan Business Leaders Say G-7 Accord is Worrying
corporate	Corling <GLW>, Hazleton <HLC> Set Exchange Ratio
commodity	EC Commission Details Sugar Tender
energy	EIA Says Distillate, Gas Stocks off in Week
ship & energy	Kuwait May Re-register Gulf Tankers - Newspaper
money-fx & economic	Taiwan Said Considering Currency Liberalization
money-fx & interest & currency	J.P. Morgan <JPM> Says DLR may Prevent FED Easing

Table 1: Example titles from the corpus.

5 Text Routing for a Newswire

The Reuters text categorization test collection [9] contains real-world documents which have appeared on the Reuters newswire. All news titles in the Reuters corpus belong to one or more of eight main categories: Money/Foreign Exchange (**money-fx**, **MFX**), Shipping (**ship**, **SHP**), Interest Rates (**interest**, **INT**), Economic Indicators (**economic**, **ECN**), Currency (**currency**, **CRC**), Corporate (**corporate**, **CRP**), Commodity (**commodity**, **CMD**), Energy (**energy**, **ENG**). Some examples of typical titles from these categories are shown in Table 1.

The titles are from the so-called ModApte split; all 10 733 titles are used and belong to at least one topic category; the overall number of words is 82 339, with the different number of words being 11 104. We use 1 040 for the training set, the first 130 titles of each of the 8 separate categories. The remaining 9 693 titles are kept for testing the generalisation performance with unseen examples.

The words in the corpus are represented using semantic significance vectors. These vectors are determined based on the frequency of a word in different semantic categories. Each word w is represented with a *vector* $(v(w,c_1), v(w,c_2), \cdots, v(w,c_n)$, where c_i represents a certain semantic category. A *value* $v(w,c_i)$ is computed for each dimension 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 *normalized* frequency of occurrences of word w in the corpus (the normalized corpus frequency). That is:

$$v(w,c_i) = \frac{Norm.\ freq.\ of\ w\ in\ c_i}{\sum_j Norm.\ freq.\ of\ w\ in\ c_j}$$

for $j \in \{1, \cdots n\}$, and where

$$Norm.\ freq.\ of\ w\ in\ c_i = \frac{Freq.\ of\ w\ in\ c_i}{No.\ of\ titles\ in\ c_i}$$

The type of normalization applied here ensures that the word representation is independent of the number of examples observed in each category. Therefore, the vectors represent the plausibility of a word occurring in a particular semantic category.

In our experiments, we use the recurrent plausibility network described in Figure 1 with two hidden and two context layers. Input to the network is the word representation, one word at a time. Output is the desired semantic routing category. Training is performed until the sum squared error does not decrease anymore, typically after 900 epochs of training. Different networks were 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.

Table 2 shows the recall and precision rates [11] obtained with the recurrent plausibility network. In general, the recall and precision rates were fairly high for this noisy real-world corpus. The generalization performance for new and unseen news titles has been even better than the performance on the training data. This is a very desirable

effect and demonstrates that overfitting on the training set does not exist. The classification of the training set is actually harder to learn than the one for the large test set, since even titles from less frequent categories were presented relatively often in the training set compared to the test set.

Category	Training set		Test set	
	recall	prec.	recall	prec.
MFX	87.34	89.47	86.03	76.70
SHP	84.65	89.21	82.37	90.29
INT	85.24	87.77	88.19	86.05
ECN	90.24	91.77	81.89	83.80
CRC	88.89	91.36	89.64	89.86
CRP	92.31	92.66	95.55	95.43
CMD	92.81	93.14	88.84	90.29
ENG	85.27	87.74	87.69	92.95
All titles	89.05	90.24	**93.05**	**92.29**

Table 2: Results using recurrent plausibility network

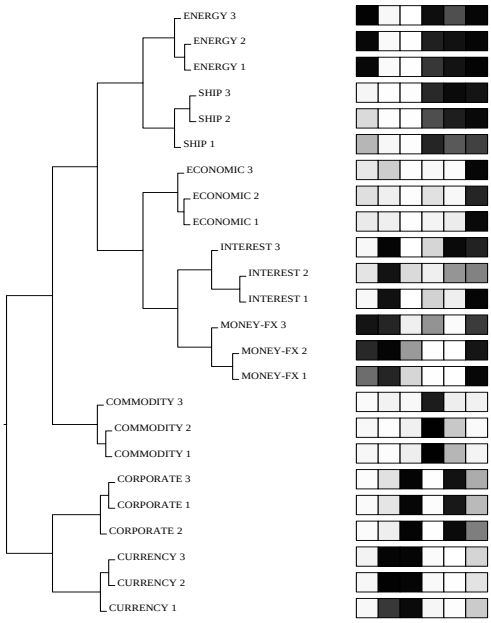

Figure 2: The cluster dendrogram and internal representations for 24 example titles from 8 categories at the end of the title

6 Analysis of the Representations

To illustrate the network's internal representation, we show an analysis of a set of 24 correctly classified titles from the test set, 3 titles from each semantic category. Examples of the titles from this set are shown in Table 1. After processing a title, we take the second hidden layer as an activation vector containing the internal network representation of this title. The cluster analysis of these vectors is presented in Figure 2.

This analysis represents the complexity of the corpus and the relationships between the separate semantic categories. According to the Reuters corpus, certain categories such as "money-fx", "interest" and "economic" are very close in their semantic interpretation, as they have a lot of common terms and a large number of shared documents. Similarly, as a result of the significant number of documents about oil tankers, the network considers the categories "ship" and "energy" as close. Most important however, we can observe that the network clearly distinguishes between these categories in its internal representation: titles from the same category belong to closely related clusters.

We have also developed a new method for visualizing the error over sequences and training time and show some representative examples of this detailed analysis.

In Figure 3, we present the processing of the title "Bank of Japan Intervenes Shortly after Tokyo Opens" during the training of a plausibility neural network for 900 epochs. The sum squared error of the output preferences is shown for different epochs of training and for all words of a title. We can see how the error quickly decreases over the earlier epochs for all words. Furthermore, at the beginning of the title, the error is higher since it is not a significant start for the desired semantic categories "money-fx" and "currency".

The error surface of another title is shown in Figure 4. At the beginning of the training, the network has not yet learned the strong preferences in the corpus although the overall value of the error is relatively low at the beginning of this title. Between epochs 300 and 700, we can observe that the network has learned the most significant regularities and has a strong preference for a particular semantic category. At the term "Soviet Union", it classifies the title into a wrong category. Then, the word "crude" causes the network preference to switch to the correct category. However, during this part of

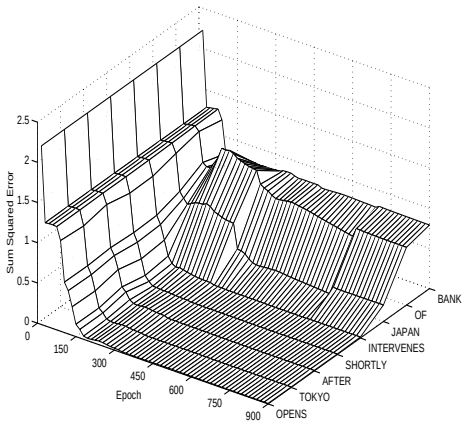

Figure 3: The error surface of the title "Bank of Japan Intervenes Shortly after Tokyo Opens"

the training, the network has not yet learned to deal with the context. In the last part of the training, the network has learned the context and gives the correct classification at the end of the title.

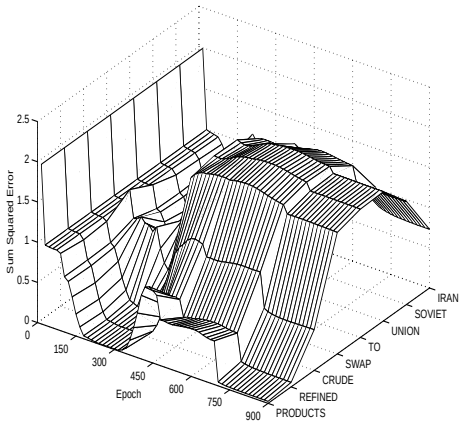

Figure 4: The error surface of the title "Iran, Soviet Union to Swap Crude, Refined Products"

Figure 5 shows the surface error of a long title from the "economic" category. Some words at the beginning of the title are associated with different semantic categories and increase the error. During the first part of the training, the network is more unstable and readily fluctuates in its output preference. However, the trained network supports the incremental context and has a more stable output preference.

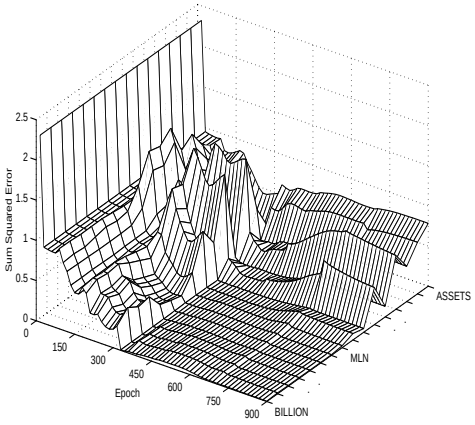

Figure 5: The error surface of the title "Assets of Money Market Mutual Funds Fell 35.3 mln Dlrs in Latest Week to 237.43 billion"

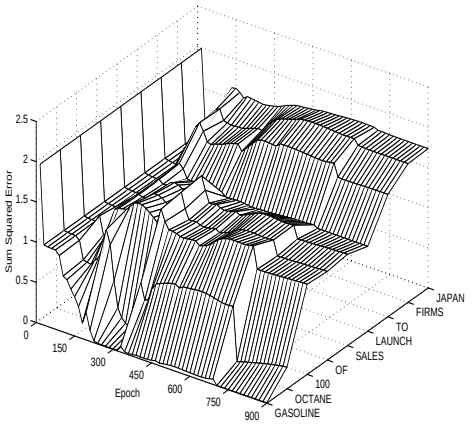

Figure 6: The error surface of the title "Japan Firms to Launch Sales of 100 Octane Gasoline"

In Figure 6, we present the title "Japan Firms to Launch Sales of 100 Octane Gasoline". This title belongs to the "energy" category. The overall error is high at the beginning of the sentence. The categorisation preference is initially improved but the word "sales" causes an incorrect categorisation. However, the words "octane" and "gasoline" cause a switch to the correct category as the correct context is learned.

7 Conclusion

We have described a hybrid neural agent HyNeT for routing news headlines. We demonstrate that carefully developed neural network architectures can deal with larger training and test sets. On an 82 339 word corpus, the recall and precision rates for recurrent plausibility networks were fairly high (92% and 93%), given the degree of noise and ambiguity. In comparison, a bag-of-words approach, to test performance on sequences without order, reached 86.6% recall and 83.1% precision.
Furthermore, we have also carefully examined the error of the network at each epoch and for each word of a training headline. These surface error figures allow a clear, comprehensive evaluation of training time, word sequence and overall classification error.
To date, recurrent neural networks have not been developed for a new task of such size and scale, in the design of title routing agents. HyNeT is robust, classifies noisy arbitrary real-world titles, processes titles incrementally from left to right, and shows better classification reliability towards the end of a title based on the learned context.

References

[1] T. Mitchell A. McCallum, R. Rosenfeld 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.

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

[3] 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*, 1998.

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

[5] D. Freitag. Information extraction from html: Application of a general machine learning approach. In *Proceedings of AAAI/IAAI*, pages 517–523, Madison, Wisconsin, 1998.

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

[7] M. I. Jordan and D. E. Rumelhart. Forward models: supervised learning with a distal teacher. In Y. Chauvin and D. E. Rumelhart, editors, *Backpropagation: theory, architectures and applications*, pages 189–236. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995.

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

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

[10] M. Marchiori. The limits of web metadata, and beyond. In *Proceedings of the Seventh International World Wide Web Conference*, 1998.

[11] G. Salton. *Automatic Text Processing*. Addison-Wesley, New York, 1989.

[12] 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.

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

[14] S. Wermter and R. Sun. *Hybrid Neural Symbolic Systems*. Springer, Heidelberg, 1999 (to appear).

[15] S. Wermter and V. Weber. SCREEN: Learning a flat syntactic and semantic spoken language analysis using artificial neural networks. *Journal of Artificial Intelligence Research*, 6(1):35–85, 1997.