Network Analysis in a Neural Learning Internet Agent

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

Abstract

In recent years, with the massive increase in the amount of available online information on the Internet, a need has arisen for being able to organize and access that data in a meaningful and directed way. In this paper, special emphasis is placed on neural network approaches in implementing a learning agent for text routing. A brief summary of the various important approaches used is presented. An approach for neural learning internet agents is outlined, one that uses recurrent neural networks for the learning of classifying a textual stream of news information. In particular, we analyze the context of the internal memory and how it is relevant in making classification decisions.

Introduction

Recently there has been a lot of interest in agent development for internet access. The most general explanation of an agent is that it is a software system, to some degree autonomous, that is designed to perform a specific task (Balabanovic, Shoham, & Yun 1997; Menczer, Belew, & Willuhn 1995).

Internet agents can be designed to perform various tasks, whether they be textual classification (Nigam *et al.* 1998; Fuernkranz, Mitchell, & Riloff 1998), information retrieval/extraction (Craven *et al.* 1998; Freitag 1998), automated web browsing (Balabanovic & Shoham 1995), and learning web-agents (Liere & Tadepalli 1996; Wermter, Panchev, & Arevian 1999; Wermter 1999).

As the Internet has got larger, and the contents of the information environment have become more complex (Menczer & Belew 1999), the need to search, filter, organize and update all the information available in a meaningful and directed way has gained greater importance. However, most of the work on internet agents has not focused on learning. Learning internet agents is the topic of this paper and it focuses on neural network learning techniques and how they can be used in an internet agent.

Recent applications of specific neural network and connectionist approaches have been successfully applied to the problem of classification of textual data; two contrasting examples are the HyNeT agent (Wermter, Panchev, & Arevian 1999) that uses a recurrent neural network, and the WEBSOM agent (Kohonen 1995; Kaski *et al.* 1998), that makes use of an unsupervised Kohonen network. In this paper we present an update and new results from our learning agent HyNeT for classifying news articles. In particular we focus on a new context analysis and offer an explanation why the networks perform well.

Neural Routing Machinery

One interesting area for exploring learning internet agents is the routing and classification of newswire titles. Neural network agents are adaptive to a changing environment, are robust against noise and are able to learn nonlinear representations of an external environment - qualities attractive for a system that needs to interact dynamically to any changes, such as ambiguous or corrupted messages.

The specific neural network explored here is a more developed version of the simpler recurrent neural network, namely a Recurrent Plausibility Network (Wermter 1995). Recurrent neural networks are able to map both previous internal states and external states to a desired output - essentially acting as short-term incremental memories that take time and context into consideration. The recurrent plausibility network combines the features of recurrent networks with distributed context layers and self-recurrent connections of the context layers. It consists of an input layer, multiple hidden and context layers and an output layer. Recurrent connections exist for the hidden layers and can be spatially modeled in context layers (for more details see (Wermter 1995)).

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.

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

In our particular case here, 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.

Experiments

In order to get a comparison of performance, several experiments were conducted using different vector representations of the words in the Reuters corpus (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. We use all 10 733 titles with 82 339 words. The number of different words in the titles is 11 104. For training we use 1 040 news titles, the first 130 of each of the 8 categories. All other 9 693 news titles are used for testing.

The variously derived vector representations were fed into the input layer of simple recurrent networks, the output being the desired semantic routing category. The preprocessing strategies are briefly outlined and explained below. The recall/precision results are presented below in Table 1 for each experiment.

Complete News Titles and Significance Vectors In the initial experiment, words were represented using significance vectors; these were obtained by determining the frequency of a word in different semantic categories using the following operation:

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

If a *vector* $(x_1 x_2 \ldots x_n)$ represents each word w, and x_i is a specific semantic category, then $v(w, x_i)$ is calculated for each dimension of the word vector, as the frequency of a word w in the different semantic categories x_i, divided by the number of times the word w appears in the corpus. The computed values are then presented to input layer in the form $(v(w, x_1), v(w, x_2), \ldots, v(w, x_n))$.

Complete News Titles and Semantic Vectors An alternative preprocessing strategy was to represent 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, using the following operation:

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

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 again 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*).

Complete News Titles with Recurrent Plausibility Network In the final experiment, a recurrent plausibility network was trained (Wermter 1995; Wermter, Panchev, & Arevian 1999). The actual architecture used for the experiment was one with two hidden and two context layers. After trying various combinations of settings for the values of the hysteresis value for the activation function of the context layers, it was found that the network performed optimally with a value of 0.2 for the first context layer, and 0.8 for the second. The remaining parameters were all set to those of the previous experiment using the semantic vector representation to allow the comparison of the various performances.

The results in Table 1 show the clear improvement in the overall recall/precision values from the first experiment using the significance vectors, to the last using the plausibility network. The experiment with the semantic vector representation showed an improvement over the first. The best performance was found for the plausibility network. Though it may be argued that the overall classification improvement from the first set of experiments to the last was around 2%, in a corpus of 10,000 titles, this is equivalent to about 200 correctly classified titles, a significant number in real terms.

Analysis of the Output Representations

For a clear presentation of the network's behavior, the results are illustrated and analyzed below; the error surfaces show plots of the sum-squared error of the output preferences, plotted against the number of training epochs and each word of a title.

Figure 1 shows the surface error of the title "Miyazawa Sees Eventual Lower US Trade Deficit"; in the Reuters Corpus this is classified under the "economic" category; as can be seen, the network does learn the correct category classification. The first two words, "Miyazawa" and "sees", are initially given several possible preferences to other categories and the errors are high early on in the training. However, the subsequent words "eventual", "lower", etc. cause the network to increasingly favor the correct classification, and at the end, the trained network has a very strong preference (shown by the low error value) for the incremental context of the desired category.

Type of Vector Representation Used in Experiment	Training set		Test set	
	recall	precision	recall	precision
Significance Vectors with Simple Recurrent Network	85.15	86.99	**91.23**	**90.73**
Semantic Vectors with Simple Recurrent Network	88.57	88.59	**92.47**	**91.61**
Semantic Vectors with Recurrent Plausibility Network	89.05	90.24	**93.05**	**92.29**

Table 1: Recall/precision results from various experiments

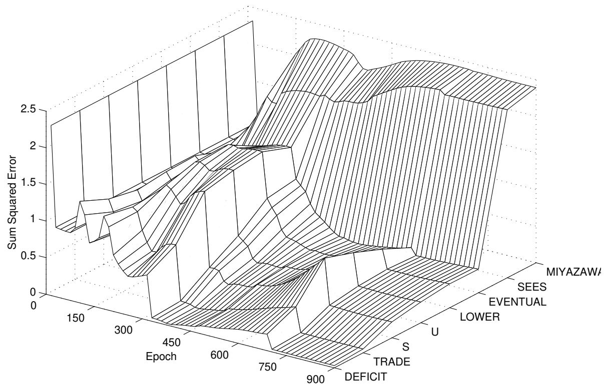

Figure 1: The error surface of the title "Miyazawa Sees Eventual Lower US Trade Deficit"

Analysis of Context in Plausibility Networks

Based on five representative titles from each category, we define the *class-activity* of a particular context unit for a particular category as the activation of this unit when a title from this category is presented to the network. That is, for example, at the end of a title from the category "money-fx" the units with the higher activation will be classified as more class-active, and the units with lower activation as less class-active.

For the analysis of the context building in the plausibility networks, we recorded the activation of the context units while processing the title "Assets of money market mutual funds fell 35.3 mln dlrs in latest week to 237.43 billion". This is a title from the category "economic". The recorded data was sorted with a key which is in our example the activity of the neurons for the category "economic". The results are shown in Figures 2 and 3.

The most active unit for the class "economic" is specified as unit 1 in the figure and the unit with the lowest class-activity as unit 6. Thus, the ideal curve at a given word step for the title to be classified to the correct category will be a monotonically decreasing function starting from the units with the highest class-activity to the units with lower class-activity. As we can see, most of the units in the first context layer (closer to the input) are more dynamic. They are highly dependent on the current word. Therefore the first context layer does not

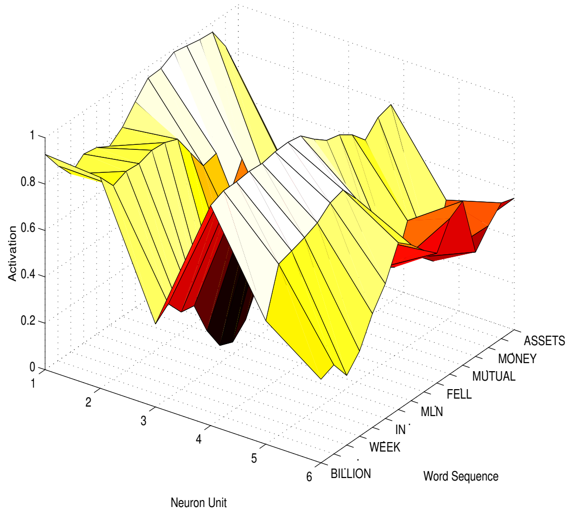

Figure 2: The first context layer

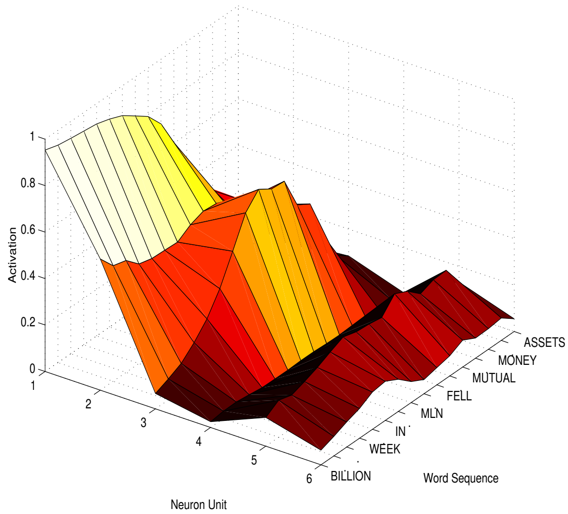

Figure 3: The second context layer

build a representative context for the required category at the end of the title. It rather responds to the incoming words, building a short dynamic context. However, the second context layer is incrementally building its context representation for the particular category. It is the context layer which is most responsible for a stable output and does not fluctuate so much with the incoming different words.

Conclusion

We analyzed the internal dynamics of HyNeT, a news agent based on recurrent neural networks. 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. We analyzed the underlying context mechanisms that extend previous recurrent neural network mechanisms, in particular multiple hidden layers in recurrent networks. Such detailed network analysis allows a better understanding of the internal nonlinear dynamics of recurrent neural networks.

References

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, AAAI Press.*

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

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.*

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

Fuernkranz, J.; Mitchell, T.; and Riloff, E. 1998. A case study in using linguistic phrases for text categorization on the www. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorisation.*

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

Kohonen, T. 1995. *Self-Organizing Maps.* Berlin, Heidelberg: Springer. (Second Extended Edition 1997).

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

Liere, R., and Tadepalli, P. 1996. The use of active learning in text categorisation. Technical report, AAAI Spring Symposium on Machine Learning in Information Access.

Menczer, F., and Belew, R. 1999. *Adaptive Retrieval Agents: Internalizing Local Context and Scaling up to the Web.* Boston, MA; Manufactured in The Netherlands: Kluwer Academic Publishers.

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

Nigam, K.; McCallum, A.; Thrun, S.; and Mitchell, T. 1998. Learning to classify text from labeled and unlabeled documents. In *Proceedings of the American Association for Artificial Intelligence.*

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

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

Wermter, S. 1999. Preference moore machines for neural fuzzy integration. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 840–845.