Machine Translation using Neural Networks and Finite-State Models*

M. Asunción Castaño[1] Francisco Casacuberta[2] Enrique Vidal[2]

(1) Unidad Predepartamental de Informática
Campus Penyeta Roja, Univ. Jaume I
12071 Castellón de la Plana (Spain)

(2) Dpto. de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
46071 Valencia (Spain)

email: castano@inf.uji.es

Abstract. Both Neural Networks and Finite-State Models have recently proved to be encouraging approaches to Example-Based Machine Translation. This paper compares the translation performances achieved with the two techniques as well as the corresponding resources required. To this end, both Elman Simple Recurrent Nets and Subsequential Transducers were trained to tackle a simple pseudo-natural machine translation task.

1 Introduction

Example-Based (EB) Machine Translation (MT) systems have recently led to successful limited-domain applications [Brown et al., 1993, Vogel et al., 1996]. *Subsequential Transducers* (SSTs) [Berstel, 1979], which are also within the EB framework and are a class of Finite-State Models (FSMs), have become an interesting approach to both Language Understanding [Castellanos et al., 1993] (considered as a particular case of translation), and MT [Castellanos et al., 1994, Oncina et al., 1994, Vilar et al., 1995]. The appeal of transducer learning is in part due to the fact that very accurate translation models can be obtained when enough training examples are presented [Oncina et al., 1993]. However, the amount of data required is sometimes quite large.

Moreover, *Neural Networks* (NNs), so-called *Connectionist Models,* can also be considered as an encouraging approach to EB MT. On that score, NNs have also shown empirical success dealing with Language Understanding tasks [Stolcke, 1990, Castaño et al., 1995]. However, only a few connectionist MT systems have been developed in the literature. PARSEC [Jain, 1991], which was used in the JANUS project [Waibel et al., 1991], follows this approach. Another effective and more simple EB connectionist translator for text-to-text applications has been recently introduced in [Castaño & Casacuberta, 1997]. The preliminary results presented in that paper indicated that translations from the source to the target language can be automatically and successfully approached. In addition, these findings suggested that small corpora are required to train the neural models. FSMs had also been previously applied to the same task considered in these pilot experiments [Castellanos et al., 1994, Oncina et al., 1994]. However, SSTs were not trained and tested on the same data as those employed for NNs. Consequently, both connectionist and transduction models could not be precisely compared. Appropriate experiments which compare both MT methodologies are carried out and are discussed in this paper.

The paper is organized as follows: First, the MT task in which NNs and SSTs are compared is described. Section 3 presents the neural architecture employed, as well as the procedure used to train the net. Section 4 briefly describes some concepts related to SSTs and the different experiments which were carried out with these models. In Section 5, the experimental results achieved with the two techniques are reported. The conclusions of the comparative experiments are discussed in Section 6.

* Work partially funded by the Spanish C.I.C.Y.T. (project TIC95-084-CO2)

2 The Experimental Machine Translation Task

With the objective of comparing Connectionist and Finite-State approaches to MT, both models were tested on a pseudo-natural task called *Miniature Language Acquisition* (MLA). This task was originally introduced by Feldman and his collaborators and involved descriptions of simple two-dimensional visual scenes [Feldman et al., 1990]. Later, it was adequately reformulated as an MT task in [Castellanos et al., 1994], so that descriptions of scenes in a given source language had to be associated with corresponding sentences in another target language. In that paper, English, Spanish and German were considered. However, only the first two languages are taken into account in this paper. An example of paired sentence of this task, named *Descriptive MLA-MT,* is shown in Figure 1.

Since this Descriptive task involved fairly-simple syntax, a more complex *Extended MLA-MT task,* (which increased the degree of input-output "asynchrony") [Castellanos et al., 1994] was also considered in our experiments. This last task included the possibility of adding or removing objects to or from a scene. Figure 1 shows an example of these sentences.

Following the directions of the original task formulation [Feldman et al., 1990] and its corresponding extension [Castellanos et al., 1994], a large corpus of paired sentences was generated in a semi-automatic way for each pair of languages and for each of the two above tasks.

Spanish:	un *cuadrado mediano y claro y un círculo claro tocan a un círculo y un cuadrado mediano*
English:	*a medium light square and a light circle touch a circle and a medium square*
Spanish:	*se elimina el círculo grande que está encima del cuadrado mediano y oscuro y del triangulo*
English:	*the large circle which is above the medium dark square and the triangle is removed*

Figure 1. Two examples of Spanish-English paired sentences from the Descriptive (above) and Extended (below) MLA-MT task, respectively. Vocabulary sizes are about 30 words and trigram test-set perplexities are about 3.

3 Learning Neural Networks

3.1 Network Architecture

In accordance with the nature of the MT tasks, a connectionist model with an explicit representation of time was required. A *Simple Recurrent Network* (SRN) presented in [Elman, 1990] was adopted as the basic architecture in the preliminary experiments [Castaño & Casacuberta, 1997]. In order to increase the performance of the model, preceding and following contexts of the input signal were presented to the SRN. Figure 2 illustrates the resulting neural topology. The specific number of hidden units and the topology of the delayed inputs employed to tackle each MLA-MT task coincided with those adopted in the preceding work [Castaño & Casacuberta, 1997] and are reported in Section 5.

The input units and the output layer were designed according to a *local representation* of the source and target vocabularies, respectively. This means that input and output words were encoded by orthogonal vectors. An additional output neuron was created specifically to mark the end of the translated sentence.

3.2 Training Procedure

The neural architecture described above was trained using an on-line version of the *Backward-Error Propagation* algorithm [Rumelhart et al., 1986]; that is, a gradient-truncated version of a

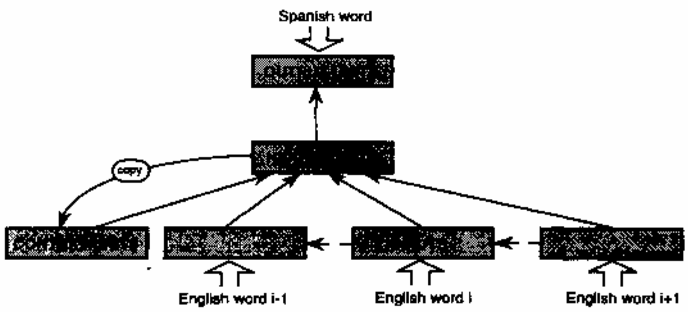

Figure 2. Delayed Elman Simple Recurrent Network.

full step-descendent procedure[1]. The words of every input message were presented sequentially at the input layer of the SRN, while the model had to provide the successive words of the corresponding translated sentence. After inputs and target units were updated, the forward step was computed, the error was back-propagated through the net and the weights were modified. Later, the hidden unit activations were copied onto the corresponding context units. This *time cycle* was continuously repeated until the target value of the corresponding output neuron identified the end of the translated sentence. A sigmoid function (0,1) was assumed as the non-linear activation function and, consequently, context activations were initialized to 0.5 at the beginning of every input-output pair. Training stopped when some established criterion was verified. In this paper, we assume the best values estimated for the learning rate and momentum in the preliminary experiments [Castaño & Casacuberta, 1997].

With regard to the translated message provided by the net, the SRN continuously generated (at each time cycle) output activations. Since only one of the output neurons should be activated at a time, we considered that the output word was the word associated to the neuron with a maximum activation.

4 Learning Finite-State Models

In order to compare Finite-State Models (FSMs) with the previous connectionist translator, subsequential transducers were inferred and tested using the same training and test data, respectively, as those which were considered for the SRNs. Details of this kind of model and the corresponding learning process are given in [Amengual et al., 1997] and only the basic ideas will be reviewed here below.

4.1 Subsequential Transducer Learning: Basic Concepts

A *Subsequential Transducer* (SST) is a deterministic finite-state network that accepts sentences from a given input language and produces associated sentences of an output language. Each edge of the network has an input symbol and an output string associated to it. Every time an input symbol is accepted, the corresponding string is output and a new state is reached. After the whole input is processed, additional output may be produced from the last (final) state reached in the analysis of the input [Berstel, 1979].

[1] All connectionist experiments presented in the paper were trained and tested using the SNNS neural simulator [Zell et al., 1995].

An example of SST can be seen in Figure 3. This transducer correctly translates the four Spanish phrases *un triangulo, un cuadrado, un triangulo grande,* and *un cuadrado grande* into the corresponding *a triangle, a square, a large triangle,* and *a large square.*

Given a set of training pairs of sentences from a translation task, a SST is learned (induced) using a very efficient algorithm called *Onward Subsequential Transducer Inference Algorithm* (OSTIA) [Oncina et al., 1993].

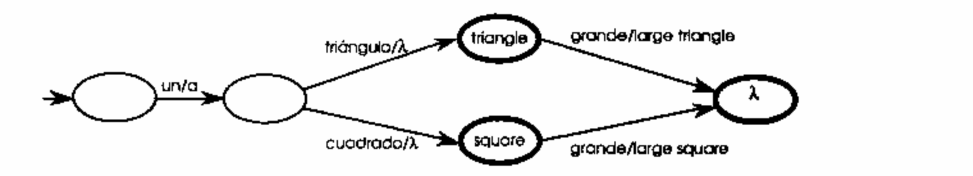

Figure 3. An example of Subsequential Transducer. The initial state has an arrow pointing to it and final states are marked by double-circling. The empty string is represented by A.

4.2 Transducer Learning using Input-Output Language Models

The learning strategies followed by OSTIA attempt to generalize the training pairs as much as possible. This often leads to very compact transducers that accurately translate input text. However, this compactness often entails undesirable "overgeneralization" of the input and output languages, which can be overcome by imposing adequate Language Model constraints: the learned SSTs should not accept input sentences or produce output sentences which are not consistent with given Language Models of the input (Domain) and output (Range) languages. Learning with Domain and Range (DR) constraints can be carried out with a version of OSTIA called OSTIA-DR [Oncina et al., 1994].

In our experiments we modelized both input and output language using simple FSMs; namely "4-Grams" [Jelinek, 1996]. They were obtained from the same input-output training sentences as those which were employed later to infer the SST with the 4-Gram DR constraints.

4.3 Transducer Learning using Error Models

SSTs inferred by OSTIA tend to produce large failures for test input sentences whose structure has not been exactly captured in the learned models. This behaviour can be improved with an error-correcting extension which includes an stochastic error model [Amengual et al., 1997].

In summary, the following three experiments related to FSMs were tested in the paper: OSTIA learning without DR constraints and OSTIA-DR learning with and without the error model.

5 Comparative Experimental Results

Three tasks were considered in this work: Both the Spanish-to-English and English-to-Spanish translations for the Descriptive MLA-MT task, and the Spanish-to-English Extended task only.

5.1 Training and Recognition Data

For each of the three tasks considered, an increasing and accumulative sample was employed to train both the neural and the finite-state models. Successive blocks of 500, 1,000, 2,000, ... input-output pairs were tried until the test translation rates were close to perfect performances.

Test set accuracies were obtained by evaluating the learned models on 6,000 sentences (for each task) which were different to those employed for training. Both training and test sets were randomly drawn from the original, large corpus mentioned in Section 2.

5.2 Criterion Assessing Correct Translations

A *source test sentence* supplied to a connectionist or finite-state model was considered to be *correctly translated* if the output provided by the model exactly coincided with the expected translation for this source sentence. In order to determine *word accuracy,* the obtained and expected translations corresponding to every source sentence in the test sample were compared using a conventional Edit-Distance (Dynamic Programming) procedure. In this way, the minimum number of insertions, deletions and substitution errors were obtained. The word accuracies reported here correspond to the ratio of the total number of no-errors with respect to the total number of edit (total error + correct) operations.

5.3 Results for the English-to-Spanish Descriptive MLA-MT Task

Previous results for the English-to-Spanish Descriptive MLA-MT task [Castaño & Casacuberta, 1997] revealed that the best Elman SRN for approaching this task had 22 input units, 24 outputs, 3+3 delayed input words and 60 hidden units (and, consequently, 344 neurons and 17,160 trainable connections in all). This network was trained on the accumulative learning sample, requiring from 1,000 to 3,000 epochs to converge. Figure 4 shows the evolution of the (averaged) test performances achieved at both word and whole-sentence levels as a function of the size of the training set.

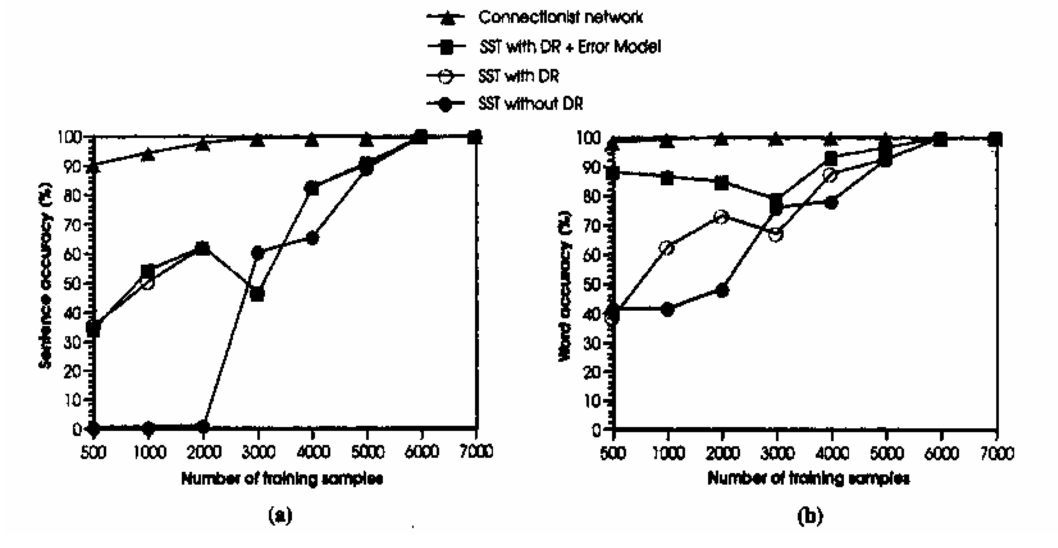

Figure 4. Sentence translation rates (a) and word accuracies (b) achieved using Connectionist Models and different Finite-State Models trained on an incremental training set for the *English-to-Spanish Descriptive MLA-MT task.*

The same accumulative training data considered for these neural models was employed to infer successive SSTs using the different versions of OSTIA (including and not including DR constraints and the error model as described in Section 4). The resulting transducers were not large: typically from 450 to 1,250 states and from 1,100 to 3,200 edges, when the SST was integrated with the input and output language models. The induced FSMs were then evaluated on the same test corpus employed in the connectionist approach and the corresponding (averaged) performances are shown in Figure 4. These results clearly indicate that NNs require less training data to converge than FSMs.

5.4 Results for the Spanish-to-English Descriptive MLA-MT Task

In the Spanish-to-English Descriptive MLA-MT task, only 500 and 1,500 learning samples were taken into account, since neural models require large computation time to be trained. The connectionist architecture adopted was an Elman SRN with 23 input units, 23 outputs, 2+7 delayed input words and 80 hidden units (and, consequently, with 413 neurons and 26,800 trainable weights). Previous experiments for this task [Castaño & Casacuberta, 1997] suggested this network as the best (delayed) Elman configuration for approaching the task. Both connectionist and finite-state models were trained on the two preceding corpora and tested on the same test corpus. The (averaged) translation accuracies achieved (see Table 1) again show that NNs outperformed FSMs. It should be noted that a SST inferred with DR constraints only provides complete sentences; and, consequently, the translation word accuracy could be decreased (as it can be observed in Table 1) when such constraints are included in the transducer. The induced SSTs for the Spanish-to-English Descriptive MLA-MT task had less than 1,000 states and 2,000 edges[1].

Table 1. Sentence translation rates (SA) and word accuracies (WA) obtained using Connectionist Models and different Finite-State Models trained on two incremental corpora from the *Spanish-to-English Descriptive MLA-MT task.*

Number of training samples	SSTs without DR		SSTs with DR		SSTs with DR Error Model		Neural networks	
	SA	WA	SA	WA	SA	WA	SA	WA
500	0.6%	43.2%	18.8%	23.4%	18.9%	84.6%	86.7%	98.4%
1,500	25.2%	59.0%	58.6%	70.2%	58.7%	86.7%	97.9%	99.8%

5.5 Results for the Spanish-to-English Extended MLA-MT Task

Finally, the neural approach and the approach based on FSMs were compared for the more complicated Spanish-to-English Extended MLA-MT task. The connectionist architecture adopted had 29 input units, 26 outputs, 6+7 delayed input words and 140 hidden units (and, consequently, 712 neurons and 80,360 trainable connections in all). This Elman configuration had been the best for tackling this task in the previous work [Castaño & Casacuberta, 1997]. Both approaches considered larger training sets (with 500 and 3,000 samples) since this Extended task was relatively more complex than the previous Descriptive one. The transducers inferred had less than 1,600 states, and less than 3,500 edges. The learned neural and finite-state models were later evaluated on the test corpus. Table 2 shows the (averaged) translation rates, which clearly indicates that NNs outperformed FSMs.

[1] However, SSTs with less than 100 states and a few hundred edges were inferred using 50,000 training samples [Vidal, 1997].

Table 2. Sentence translation rates (SA) and word accuracies (WA) obtained using Connectionist Models and different Finite-State Models trained on two incremental corpora from the *Spanish-to-English Extended MLA-MT task.*

Number of training samples	**SSTs without DR**		**SSTs with DR**		**SSTs with DR Error Model**		**Neural networks**	
	SA	WA	SA	WA	SA	WA	SA	WA
500	0.3%	37.6%	1.0%	11.1%	1.1%	71.8%	53.1%	93.3%
3,000	1.1%	45.4%	48.6%	69.6%	48.7%	85.9%	98.4%	99.9%

6 Conclusions and Future Work

NNs and FSMs have been compared on two MLA-MT tasks [Castellanos et al., 1994]. When analyzing the comparative results obtained, the following general conclusions can be drawn: First, both the connectionist and finite-state models are able to successfully approach the tasks. Second, performances of the (neural and finite-state) learned models improve with the amount of training data. Third, the translation rates achieved on small training sets are clearly better using NNs. Fourth, transducer learning is usually better when DR constraints are included, as was expected; the difference becomes more noticeable for small training sets. And finally, it is advantageous for the word translations to include an error model to the SST inferred using OSTIA-DR.

Moreover, the more exhaustive experiments carried out on the Spanish-to-English Descriptive task confirm a hint suggested in recent work [Castaño & Casacuberta, 1997]: the number of samples required by FSMs to achieve a sufficiently good performance level are clearly higher than those needed by neural models; as an example, the performances reached for a NN trained using only 500 samples were similar to those obtained for FSMs trained using 5,000 paired-sentences. However, the NNs learned were bigger than the FSMs inferred. In addition, NNs required large amounts of training-time (days and even weeks), in contrast with the low learning-time (minutes) required by FSMs. Destructive methods [Omlin & Giles, 1993] and a more compact (distributed) representation of the input and output alphabets should be explored to attempt to decrease the size of the NNs and, consequently, the learning time. Word categorization for both the input and output languages [Vilar et al., 1995] and injection of "a priori" knowledge into the net [Castaño & Casacuberta, 1997] can also be tried. Finally, new architectures or training methods which continue lowering this training time should also be considered.

6.1 Acknowledgements

The authors would like to thank J.M. Vilar and J.C. Amengual for their insightful suggestions, comments and advice.

References

[Amengual et al., 1997] J.C. AMENGUAL ET AL.: "Error-Correcting Parsing for Text-to-text Machine Translation using Finite State Models". *In this Proceedings.* (1997)

[Berstel, 1979] J. BERSTEL: *Transductions and Context-Free Languages,* Teubner. (1979)

[Brown et al., 1993] P.P.BROWN, S.A. DELLA PIETRA, V.J. DELLA PRIETA AND R.L. MERCER: "The Mathematics of Statistical Machine Translation: Parameter Estimation". *Computational Linguistics,* 19, pp. 263-311. (1993)

[Castaño et al., 1995] M.A. CASTAÑO, E. VIDAL AND F. CASACUBERTA-. "Preliminary Experiments for Automatic Speech Understanding through Simple Recurrent Networks". *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH-95),* vol. 3, pp. 1673-1676, Madrid, Spain. (1995)

[Castaño & Casacuberta, 1997] M.A. CASTAÑO AND F. CASACUBERTA: "A Connectionist Approach to Machine Translation". *To appear in Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH-97),* Rhodes, Greece. (1997)

[Castellanos et al., 1993] A. CASTELLANOS, E. VIDAL AND J. ONCINA: "Language Understanding and Subsequential Transducer Learning". *Proceedings of the 1st International Colloquium on Grammatical Inference (ICGI-93),* pp.11/1-11/10, Colchester, UK. (1993)

[Castellanos et al., 1994] A. CASTELLANOS, I. GALIANO AND E. VIDAL: "Application of OSTIA to Machine Translation Tasks". In *Lecture Notes in Computer Science-Lecture Notes in Artificial Intelligence: Grammatical Inference and Applications,* vol 862, pp.93-105, R.C.Carrasco and J. Oncina (Eds), Springer-Verlag, Berlin. (1994)

[Elman, 1990] J.L. ELMAN: "Finding Structure in Time". *Cognitive Science,* 2, pp. 279-311. (1990)

[Feldman et al., 1990] J.A. FELDMAN, G. LAKOFF, A. STOLCKE AND S.H. WEBER: "Miniature Language Acquisition: A Touchstone for Cognitive Science". Technical Report no. TR-90-009, Int. Computer Science Institute, Berkeley, California. (1990)

[Jain, 1991] A.N. JAIN: "Parsing Complex Sentences with Structured Connectionist Networks". *Neural Computation,* 3, pp. 110-120. (1991)

[Jelinek, 1996] F. JELINEK: "Language Modelling for Speech Recognition". *Proceedings of the ECAI-96 Workshop: Extended Finite State Models of Language,* pp. 26-32, Budapest, Hungary. (1996).

[Omlin & Giles, 1993] C.W. OMLIN AND C.L. GILES: "Pruning Recurrent Neural Networks for Improved Generalization Performance". Technical Report no. 93-6, Computer Science Department, Rensselaer Polytechnic Institute, Troy, N.Y. (1993)

[Oncina et al., 1993] J. ONCINA, P. GARCIA, AND E. VIDAL: "Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 15, pp. 448-458. (1993)

[Oncina et al., 1994] J. ONCINA, A. CASTELLANOS, E. VIDAL, AND V.M. JIMENEZ: "Corpus-Based Machine Translation through Subsequential Transducers". *Proceedings of the 3rd International Conference on the Cognitive Science of Natural Language Processing,* Dublin, Ireland. (1994)

[Rumelhart et al., 1986] D.E. RUMELHART, G. HINTON AND R. WILLIAMS: "Learning Sequential Structure in Simple Recurrent Networks". In *Parallel Distributed Processing: Experiments in the Micro structure of Cognition,* vol. 1. Rumelhart D.E., McClelland J.L. and the PDP Research Group (Eds), MIT Press, Cambridge. (1986)

[Stolcke, 1990] A. STOLCKE: "Learning Feature-based Semantics with Simple Recurrent Networks". Technical Report no. TR-90-015, International Computer Science Institute, Berkeley, California. (1990)

[Vidal, 1997] E. VIDAL: "Finite-State Speech-to-Speech Translation". *Proceedings of the 1997 International Conference on Acoustics, Speech and Signal Processing (ICASSP-97),* pp. 111-114, Munich, Germany. (1997)

[Vilar et al., 1995] J.M. VILAR, A. MARZAL, AND E. VIDAL: "Learning Language Translation in Limited Domains using Finite-State Models: some Extensions and Improvements". *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH-95),* pp. 1231-1234, Madrid, Spain. (1995)

[Vogel et al., 1996] S. VOGEL, H. NEY, AND C. TILLMANN: "HMM-Based Word Alignment in Statistical Translation". *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96),* pp. 836-841, Copenhagen, Denmark. (1996)

[Waibel et al., 1991] A. WAIBEL, A.N. JAIN, A.E. MCNAIR, H. SAITO, A.G. HAUPTMANN AND J. TEBELSKIS: "JANUS: A Speech-to-Speech Translation System using Connectionist and Symbolic Processing Strategies". *Proceedings of the 1991 International Conference on Acoustics, Speech and Signal Processing (ICASSP-91),* pp. 793-796, Toronto, Canada. (1991)

[Zell et al., 1995] A. ZELL ET AL.: "SNNS: Stuttgart Neural Network Simulator". User manual, Version 4.1. Technical Report no. 6195, Institute for Parallel and Distributed High Performance Systems, University of Stuttgart. (1995)