# A CONNECTIONIST APPROACH TO MACHINE TRANSLATION *

*M.A. Castaño*[†], *F. Casacuberta*[††]

[†]Dpto. de Informática. Universitat Jaume I de Castellón. Spain.
[††]Dpto. Sistemas Informáticos y Computación. Universidad Politécnica de Valencia. Spain.
e-mail: castano@inf.uji.es

**Abstract**

Connectionist Models can be considered as an encouraging approach to Example-Based Machine Translation. However, the neural translators developed in the literature are quite complex and require great human effort to classify and prepare training data. This paper presents an effective and more simple text-to-text connectionist translator with which translations from the source to the target language can be directly, automatically and successfully approached. The neural system, which is based on an Elman Simple Recurrent Network, was trained to tackle a simple pseudo-natural Machine Translation task.

## 1. INTRODUCTION

In comparison with traditional *Knowledge-Based* Machine Translation (MT) systems, *Example-Based* (EB) techniques have recently led to successful limited-domain applications. Under this paradigm, systems are automatically built from training sets of examples, resulting in lower development costs. There are a number of works that directly aim at placing MT within the EB framework [1] [12] [16]. In this direction, *Neural Networks* (so-called *Connectionist Models*) can be considered as an encouraging approach to MT. In fact, they have demonstrated empirical success in tackling Language Understanding tasks [14] [2] [3], which can be considered as a particular case of translation. However, only a few connectionist MT systems were developed in the literature. One of these is PARSEC [9] which is part of the JANUS project [17]. This neural translator employs an intermediate (pivot) language for each source-target pair of languages considered. In addition, the connectionist system separately approaches the syntactic and semantic features associated to a language, resulting in a translation model which is quite complex.

This paper presents a simple EB Connectionist Translator for text-to-text, limited-domain applications which directly carries out the translation between both the input and the output languages (with no intermediate items). At the same time, the neural translator automatically learns the semantic and syntax implicit in both languages.

* Work partially funded by the Spanish CICYT (project TIC95-084-CO2)

The paper is organized as follows: Section 2 describes the basic connectionist architecture of the connectionist translator employed, as well as the procedure used to train it. Section 3 presents the MT task with which the neural translator was evaluated. The obtained performances on this task are later reported in Section 4. Finally, Section 5 discusses the conclusions of the experimental process.

## 2. MACHINE TRANSLATION THROUGH SIMPLE RECURRENT NETWORKS

### 2.1. Network architecture

In accordance with the nature of the task, a connectionist model with an explicit representation of time is required. Therefore, the basic neural architecture adopted in the experimentation of this paper is a Simple Recurrent Network (SRN) introduced in [6]. In order to increase the performance of the model, the preceding and the following contexts of the input signal were presented to the net. Figure 1 illustrates the resulting neural topology called *Basic Elman SRN*.

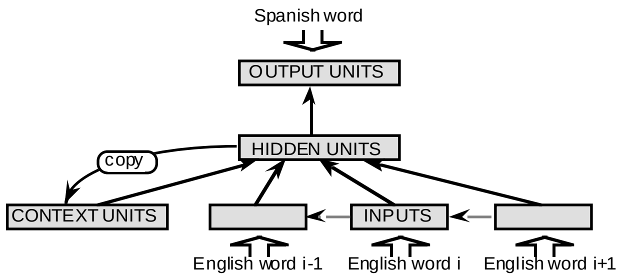


***Figure 1.*** Basic Elman Simple Recurrent Network.

Moreover, previous results on Language Understanding tasks [3] indicated that the performance of an Elman SRN could sometimes be improved when the outputs of the net were also fed backwards in time. The results seemed to be more advantageous when the net did not have enough hidden units. Consequently, an *Extended Elman SRN*, which also fed back the output activations into the hidden layer (see Figure 2), was also considered in the MT experiments.

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 included to mark the end of the translated sentence.

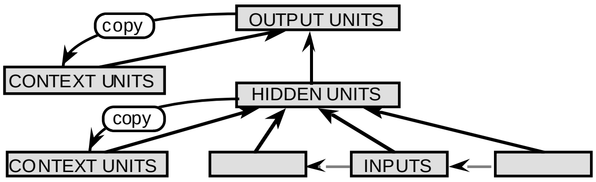


***Figure 2.*** Hybrid Elman Simple Recurrent Network.

### 2.2. Training procedure

The neural architecture described above was trained using an on-line version of the *Backward-Error Propagation* algorithm [13]; that is, a gradient-truncated version of a full-descent procedure. The words of every 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 (see Figure 1). 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 (and also the outputs for the Extended architecture) 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. The updating of the weights required estimating appropriate values for the learning rate and momentum. With this objective in mind, the net was trained for 10 random presentations of the complete learning corpus (10 epochs). Training continued for the learning rate and momentum which led to the lowest mean squared error. And the learning process stopped when a certain established criterion was verified.

With regard to the translated message provided by the net, the SRN continuously generated (at each time cycle) output activations. Because of the local representation of the target lexicon, it was determined that only one of the output neurons should be activated at a time. We considered that the net supplied the output word which was associated to the neuron with the maximum activation.

## 3. THE EXPERIMENTAL MACHINE TRANSLATION TASK

The connectionist translator described in the previous section was tested with a pseudo-natural task called *Miniature Language Acquisition* (MLA), which had been originally introduced in [7] and adequately reformulated later as a MT task [5]. This task consisted in translating descriptions of simple two-dimensional visual scenes from Spanish into English and vice versa. Small lexicons (about 22 words) were taken into account. An example of this task, named *Descriptive MLA-MT task* is shown in Figure 3.

Since this Descriptive MLA-MT task involved fairly simple syntax, a more complex *Extended MLA-MT task* (presented in [5]) was also considered in the experiments. This last task, which included the possibility of adding or removing objects to or from a scene, increased the degree of input-output asynchrony. The lexicons were also (slightly) increased to 30 words. Figure 3 shows an example of this Extended task.

## 4. EXPERIMENTAL RESULTS

While both English-to-Spanish and Spanish-to-English translations for the Descriptive task were approached in the paper, only the Spanish-to-English Extended task was taken into account. In addition, it should be noted that all connectionist experiments presented in the paper were trained and tested using the SNNS neural simulator [18].

### 4.1. Training and Recognition Data

The corpora adopted in each of the three tasks considered were sets of text-to-text pairs each of which consisted in a sentence in the source language and the corresponding translation in the target language. Two training samples of 500 and 1,500 pairs were employed to learn each Descriptive task; 500 and 3,000 training pairs were adopted for the Extended task.

In order to provide robust test accuracies, rates were obtained by evaluating the learned models on three different test sets (for each of the tasks). Each of these test corpora consisted of 2,000 sentences which were generated independently of those employed for training.

### 4.2. Criterion assessing correct translations

A *source test sentence* supplied to a connectionist architecture 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

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

***Figure 3.*** Two Spanish-English sentences from the Descriptive and Extended MLA-MT task, respectively.

the test sample were compared using a conventional Edit-Distance (Dynamic Programming) procedure. In this way, the number of insertions, deletions and substitution errors was obtained. The word accuracies reported here correspond to the ratio of the total number of non-errors with respect to the total number of edit (total error + correct) operations.

### 4.3. Results for the English-to-Spanish Descriptive MLA-MT Task

The first step in exploring the capabilities of the above EB connectionist translator for the English-to-Spanish Descriptive task was to estimate adequate values for the topology of the net. The Basic and Extended Elman SRNs had 22 input units and 24 outputs. Networks with a single hidden layer ranging from 40 to 200 units were employed. The delayed inputs ranged from 6 to 13 English words, balancing the right and left contexts of the input word. Each connectionist network was trained on each of the two learning samples (of 500 and 1,500 pairs, respectively) for 2,000 random epochs. The resulting learned models were then tested on three different sets, each of which consisted of 2,000 samples. The best test performances for this task were usually obtained training SRNs with 60 hidden units and 6 (3+3) delayed input words. Table 1 summarizes the best averaged sentence accuracy translation rates and word accuracies achieved for the topologies with such features. Looking at these results it can be observed that translation performances close to 100% were achieved for Basic Elman architectures. The results were slightly worse for Extended models.

Incremental training heuristics, which first present the shortest strings to the net, usually make the training easier and increase the convergence of neural models [8]. For testing this heuristic, additional experimentation was carried out so that the nets were first trained to translate the subjects and the direct objects of 500 input sentences; in a second step, 2,000 presentations of the training set with 1,500 complete sentences were fed to the nets. Table 1 reports the resulting translation rates on the 3x2,000 test sentences. Nevertheless, these findings did not outperform those previously obtained when the SRNs had only been trained with complete sentences.

Previous experiments on other topics related to neural networks revealed that the convergence time required to learn a task was sped up when some *a priori information* was injected into the net [11]. By assuming that some consecutive source words in the English-to-Spanish Descriptive task were always translated into a different number of consecutive (fixed) target words (*touches/toca a*, *below/debajo de*), some "empty filler words" can be included in these pieces of sentences in order to make their corresponding lengths equal. We then repeated the previous experiments. And the results revealed that a Basic SRN with 40 hidden units, 6 (3+3) input delays and 20 complete presentations of 500 training pairs was enough to perfectly approach the English-Spanish Descriptive MLA-MT task. Table 1 shows the translation performances obtained for the network with such features.

### 4.4. Results for the Spanish-to-English Descriptive MLA-MT Task

In the Spanish-to-English Descriptive MLA-MT task Basic and Extended Elman SRNs with 23 inputs and 23 outputs were trained. Since the complexity of this task is similar to the preceding English-to-Spanish Descriptive task, the size and the topology of the connectionist translator considered were similar to those employed in the previous Section. Thus, the hidden layer of the nets ranged from 60 to 100 neurons and the input layer had among 6 to 13 delayed words. However, at that point, both balanced and unbalanced right and left contexts of the input Spanish word were employed. Each of the two training corpora (with 500 and 1,500 pairs, respectively) were randomly presented to each resulting net for 2,000 epochs. The performance of the neural translator were then evaluated on three different test corpora each of which consisted of 2,000 Spanish sentences. As in the preceding English-to-Spanish experiments, translation accuracies close to 100% were also achieved for the Spanish-to-English Descriptive task. And the best results were obtained using a Basic SRN with 80 hidden units and 9 (2+7)-not balanced- delayed input words. Table 2 shows these (best) averaged word and sentence translation rates.

### 4.5. Results for the Spanish-to-English Extended MLA-MT Task

Finally, the performance of the connectionist translator proposed in the paper was evaluated on the Spanish-to-

| ARCHITECTURE | TRAINING PAIRS | HIDDEN UNITS | DELAYED INPUTS | SAR | WAR |
|---|---|---|---|---|---|
| **Basic SRN** | 500 sentences | 60 | 6 (3+3) | 90.3% | 98.3% |
| | 1,500 sentences | 60 | 6 (3+3) | 98.8% | 99.8% |
| | 500 subsentences + 1,500 sentences | 60 | 6 (3+3) | 85.7% | 98.2% |
| | 500 sentences with filler words | 40 | 6 (3+3) | 100% | 100% |
| **Extended SRN** | 500 sentences | 60 | 6 (3+3) | 81.2% | 96.3% |
| | 1,500 sentences | 60 | 6 (3+3) | 85.8% | 97.5% |
| | 500 subsentences + 1,500 sentences | 60 | 6 (3+3) | 78.3% | 96.0% |

***Table 1.*** Sentence accuracy translation rates (SAR) and word accuracy rates (WAR) for the English-to-Spanish Descriptive task using different Elman architectures.

| MLA-MT TASK | TRAINING PAIRS | SAR | WAR |
|---|---|---|---|
| **Descriptive Spanish-to-English** | 500<br>1,500 | 86.7%<br>97.9% | 98.4%<br>99.8% |
| **Extended Spanish-to-English** | 500<br>3,000 | 53.1%<br>98.4% | 93.3%<br>99.9% |

***Table 2.*** Sentence accuracy translation rates (SAR) and word accuracy rates (WAR) for the Spanish-to-English Descriptive and Extended MLA-MT tasks.

English Extended MLA-MT task. According to a local representation of the source and target vocabularies, the SRNs trained in this experiment had 29 input units and 26 outputs. Since the Extended task is more complex than the preceding Descriptive one and it also involves greater lexicons, bigger Basic and Extended Elman architectures were considered at the time. The number of hidden units ranged from 100 to 140 neurons and the delayed input Spanish words, from 9 to 13. Two learning corpora with 500 and 3,000 Spanish-to-English pairs, respectively, were employed at this time. Training was carried out for 500 random presentations of each of the two sets. The learned neural translators were then evaluated on 3 different test sets of 2,000 sentences. A Basic Elman network with 140 hidden neurons, 13 (6+7) delayed input words and trained using 3,000 pairs provided the best test translation performances. These averaged sentence and word accuracy rates are reported in Table 2. They show that a nearly perfect connectionist translators can also be obtained for the Spanish-to-English Extended MLA-MT task.

## 4. CONCLUSIONS AND FUTURE WORK

A simple Example-Based Connectionist Translator for text-to-text, limited-domain applications is presented in this paper. This neural system has been tested with a pseudo-natural task called *Miniature Language Acquisition* [5]. The translation accuracies achieved on this task were close to 100% in both simple and more complex Spanish-English translations (two-way translations). However, perfect translators were obtained by injecting some "a-priori" information about the task into the net. These models also required smaller neural architectures and less training time in order to converge.

In addition, the results obtained seem to suggest that our neural approach requires less training data than other (non-connectionist) promising Example-Based techniques [4].

Based on these encouraging performances, future work dealing with more complex limited-domain translations seems to be feasible. However, the size of the neural nets required for such applications (and consequently, the learning time) can be prohibitive. To this end, destructive methods [10] and a more compact (distributed) representation of the input and output alphabets should be explored. Word categorization for both the input and output languages [15] can be also tried. Finally, new architectures or training methods which continue to lower this training time should be also be considered.

## 5. REFERENCES


[1] P.F. Brown, S.A. Della Pietra, V.J. Della Prieta, R.L. Mercer. *The Mathematics of Statistical Machine Translation: Parameter Estimation.* Computational Linguistics, vol. 19, no. 2, pp. 263--311, 1993.

[2] M.A. Castaño, E. Vidal, F. Casacuberta. *Learning Direct Acoustic-to-Semantic Mapping through Simple Recurrent Networks*. Procs. of the 3rd European Conference on Speech Communication and Technology (EUROSPEECH-93), vol. 2, pp. 1017--1020, Berlin, Germany, 1993.

[3] M.A. Castaño, E. Vidal, F. Casacuberta. *Preliminary Experiments for Automatic Speech Understanding through Simple Recurrent Networks*. Procs. of the 4th European Conference on Speech Communication and Technology (EUROSPEECH-95), vol. 3, pp. 1673--1676, Madrid, Spain, 1995.

[4] M.A. Castaño, F. Casacuberta, E. Vidal. *Machine Translation using Neural Networks and Finite-State Models.* To appear in Procs. of the 7th Int. Conference on Theoretical and Methodological Issues in Machine Translation (TMI-97), Santa Fe, July 1997.

[5] A. Castellanos, I. Galiano, 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, 1994.

[6] J.L. Elman. *Finding Structure in Time.* Cognitive Science, vol. 2, no. 4, pp. 279--311, 1990.

[7] J.A. Feldman, G. Lakoff, A. Stolcke, S.H. Weber. *Miniature Language Acquisition: A Touchstone for Cognitive Science*. Technical Report no. TR-90-009, Int. Computer Science Institute, Berkeley, California, 1990.

[8] C.L. Giles, B.G. Horne, T. Lin. *Learning a Class of Large Finite State Machines with a Recurrent Neural Network.* Neural Networks, vol. 8, no. 9, pp. 1359, 1995.

[9] A.N. Jain. *Parsing Complex Sentences with Structured Connectionist Networks.* Neural Computation, vol. 3, pp. 110--120, 1991.

[10] C.W. Omlin, 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.

[11] C.W. Omlin, C.L. Giles. *Rule Revision with Recurrent Neural Networks.* IEEE Trans. on Knowledge and Data Engineering, no. 8, vol. 1, pp. 183--188, 1996.

[12] J. Oncina, P. García, E. Vidal. *Learning Subsequential Transducers for Pattern Recognition Interpretation Tasks.* IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 15, no. 5, pp. 448--458, 1993.

[13] D.E. Rumelhart, G. Hinton, R. Williams. *Learning Sequential Structure in Simple Recurrent Networks*. In "Parallel distributed processing: Experiments in the microstructure of cognition", vol. 1. Rumelhart D.E., McClelland J.L. and the PDP Research Group (Eds.), MIT Press. Cambridge, 1986.

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

[15] J.M. Vilar, A. Marzal, E. Vidal. *Learning Language Translation in Limited Domains using Finite-State Models: some Extensions and Improvements.* Procs. of the 4th European Conference on Speech Communication and Technology (EUROSPEECH-95), pp. 1231--1234, Madrid, Spain, 1995.

[16] S. Vogel, H. Ney, C. Tillmann. *HMM-Based Word Alignment in Statistical Translation*. Procs. of the Int. Conference on Computational Linguistics, pp. 836--841, Copenhagen, Denmark., 1996.

[17] A. Waibel, A.N. Jain, A.E. McNair, H. Saito, A.G. Hauptmann, J. Tebelskis. *JANUS: A Speech-to-Speech Translation System using Connectionist and Symbolic Processing Strategies.* Procs. of the 1991 Int. Conference on Acoustics, Speech and Signal Processing (ICASSP-91), pp. 793--796, 1991.

[18] 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.