Modeling Implicit Learning : Extracting Implicit Rules from Sequences using LSTM

Ikram Chraibi Kaadoud, Nicolas P. Rougier, Frédéric Alexandre

Modeling Implicit Learning : Extracting Implicit Rules from Sequences using LSTM

Ikram CHRAIBI KAADOUD, Nicolas ROUGIER et Frédéric ALEXANDRE
i.chraibi-kaadoud@groupeonepoint.com

Trust in neural networks is often correlated with an **understanding of the predictions** of these networks. However, the latter are rightly described as **"black boxes",** an opaque set where only inputs and outputs are accessible. This work deals precisely with the **interpretability** of neural networks. In that context, we propose a generic solution for **extracting the implicit representation developed by recurrent networks equipped with Long units Short Term Memory (LSTM),** particularly in the context of learning sequences from grammars not binary. Getting our inspiration from the studies on the **implicit sequential learning in humans,** we propose a method for extracting implicitly encoded rules in the form of graphs, with different rating systems, which **clarify knowledge about the temporal arrangement and continuous state space of the hidden layer of the network**.

Keywords: Recurrent networks, LSTM, Learning of sequences, Extraction of rules, Automata, Interpretability of neural networks.

A – Introduction & Context

A.1 - Implicit sequential learning in humans

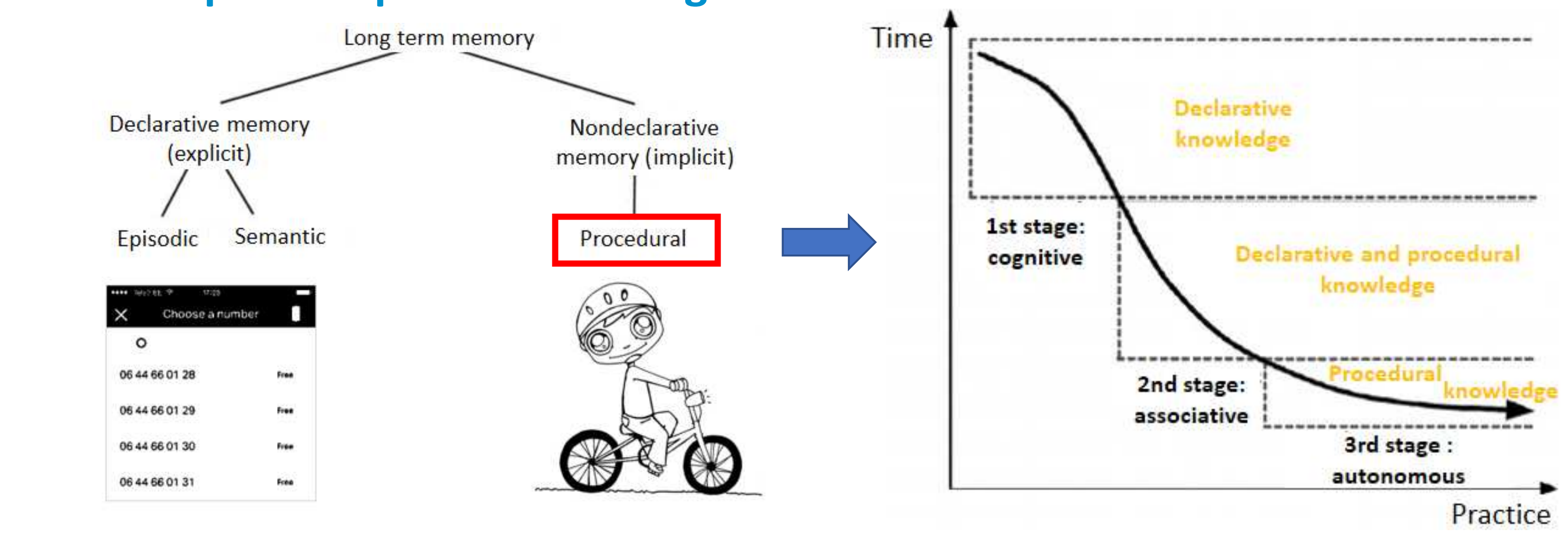


Fig.1: A partial taxonomy of the different memories (Squire and Zola, 1996) and procedural knowledge acquisition according practice and time (Kim et al,2013)

A.2 – Modeling implicit learning: the Simple Recurrent Network (SRN) approach

- Network with a feedback loop that allow to maintain a representation of the temporal context associated with the inputs (Elman, 1988)
- Framework for Modeling Sequential Learning in Humans (Cleermans et al, 1991)
- Main limitation: Attenuation of the temporal context

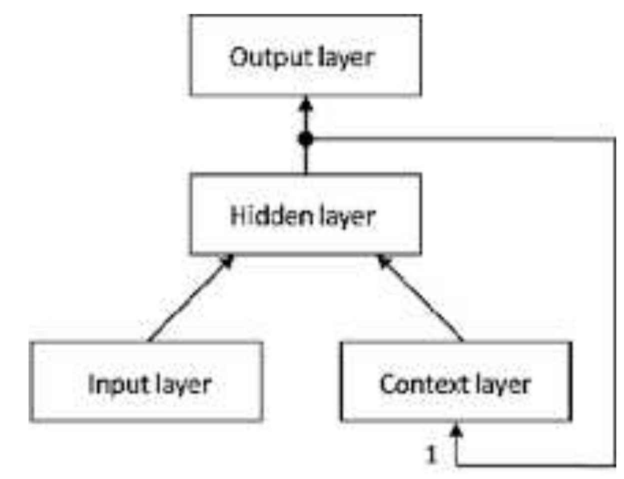
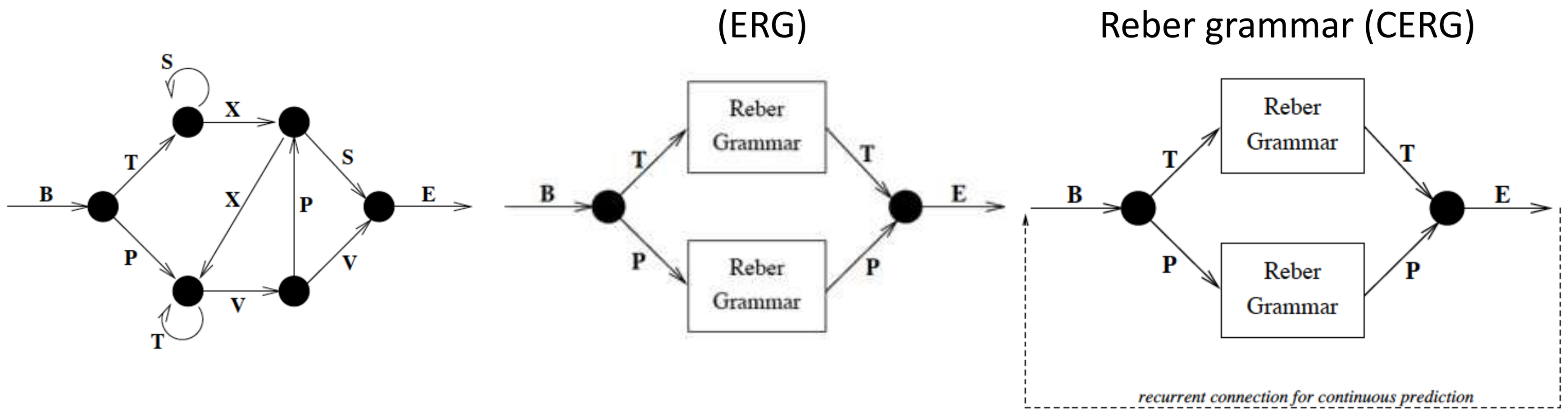


Fig.2: Architecture of the SRN

B - Grammars

Non-binary artificial grammar proposed for human experiments in measuring time respond experience in cognitive psychology (Reber, 1967)

Reber grammar (RG) — Embedded Reber grammar (ERG) — Continious Embedded Reber grammar (CERG)



C – Long Short Term Memory (LSTM) approach

Hypothesis: A network using LSTM, a model with internal and explicit representation of time, can develop an implicit representation of the rules hidden in sequences and predict according it

RNN-LSTM Architecture

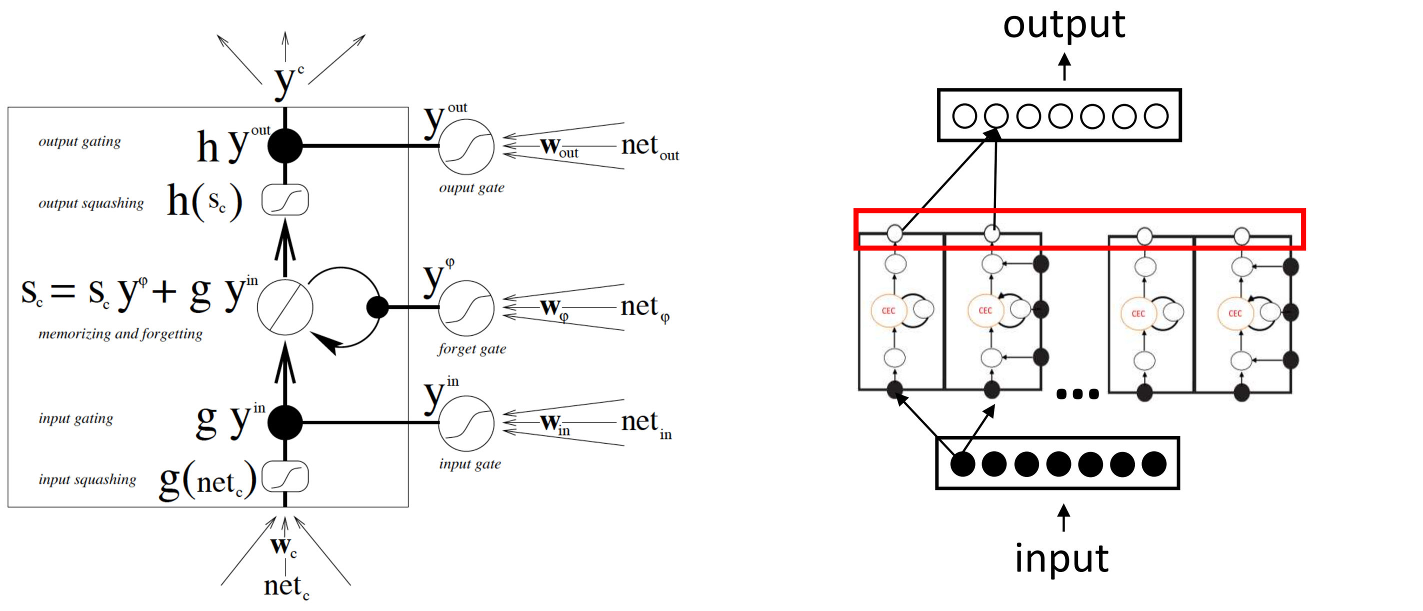


Fig.3: LSTM unit according (Gers, 1999) and its implementation in our network

3 layers:
- An input and output layer of 7 units each
- 1 hidden layer (1) of 4 blocks, of 2 cells each

Parameters : Learning Rate : 0.5 (x 0,99 each 100 time steps)

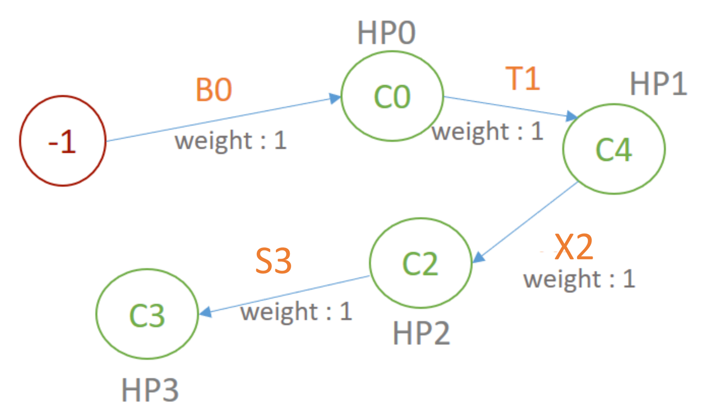
	RG & ERG	CERG
Learning	200,000 grammatical sequences	1 flow of 100,000 successive symbols
Test	10 data sets of : • 20,000 grammatical sequences • 130,000 non-grammatical sequences	10 flows of 100,000 symbols.
Results	• High recognition (close to 100%) of grammatical sequences as valid • Low recognition (close to 0%) of non-grammatical sequences as valid	100% correct predictions according (Gers, 1999) on 30,000 flows of 100,000 symbols

D – Rules extraction process from RNN-LSTM

1) Quantification using kmeans: k in [6 ; 500] on 5000 hidden patterns (10 simulations)

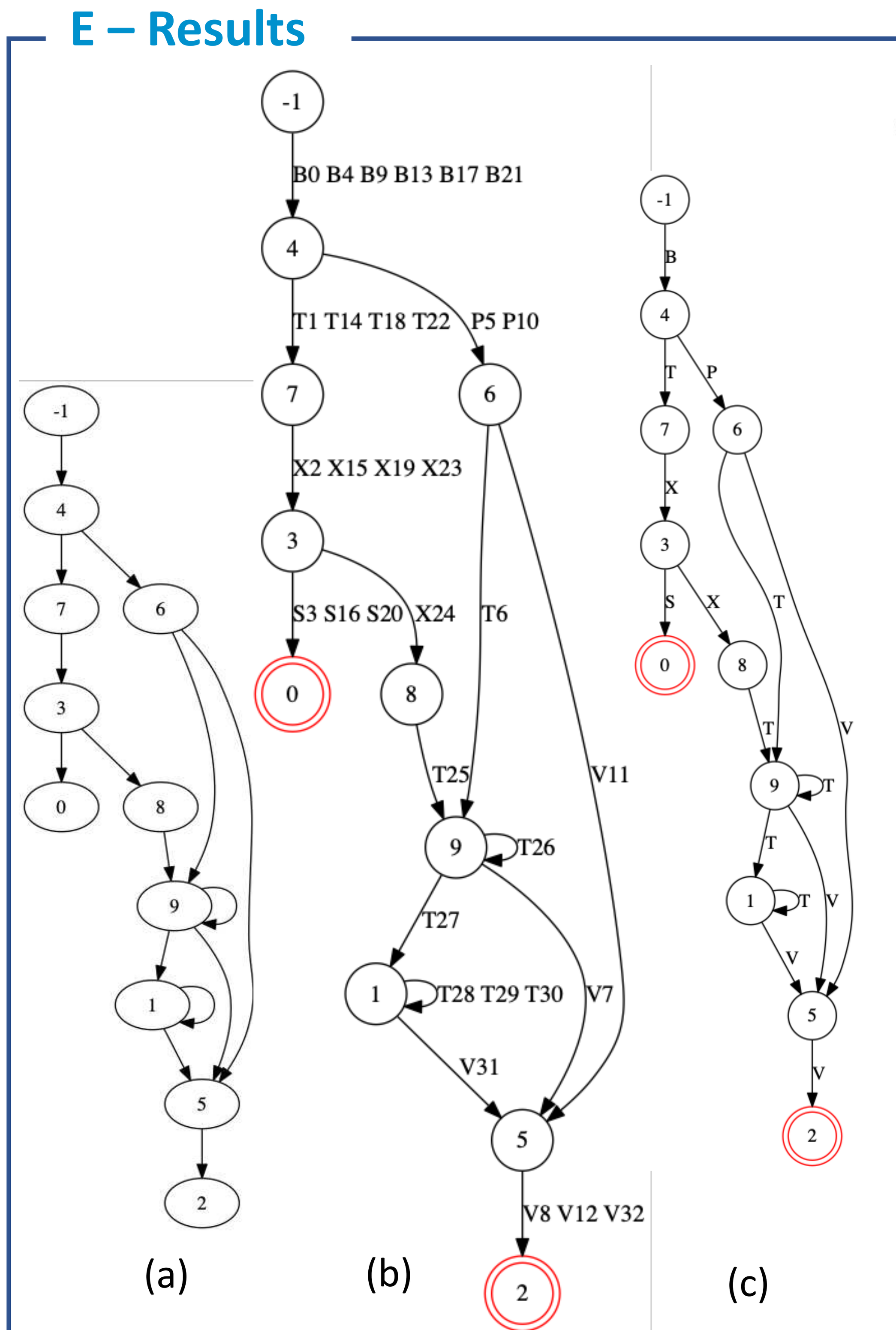
2) Rules extraction process for each k value:

Time steps	t_0	t_1	t_2	t_3
Input symbol	B	T	X	S
Index of the hidden pattern (HP)	0	1	2	3
Index of the associated cluster (C)	0	4	2	3



3) Validation of extracted automata

E – Results



(a) (b) (c)

Fig.4: Extraction in RG context with a k-means algorithm (k=10): **a** - Unlabeled automata representing the arrangement of the clusters. **b** - Long-label automata expliciting the temporal routing of patterns and "the behavior" of the network. **c** - Final automata with proposing an explicit representation of the implicit knowledge

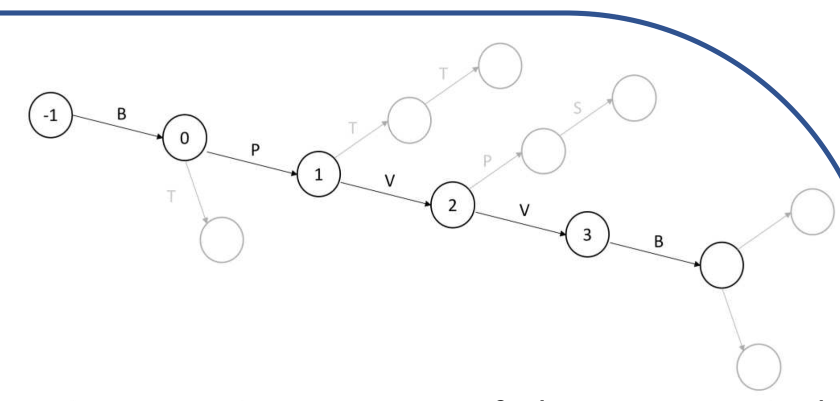


Fig.5: Testing process of the grammatical sequence BPVVE from RG

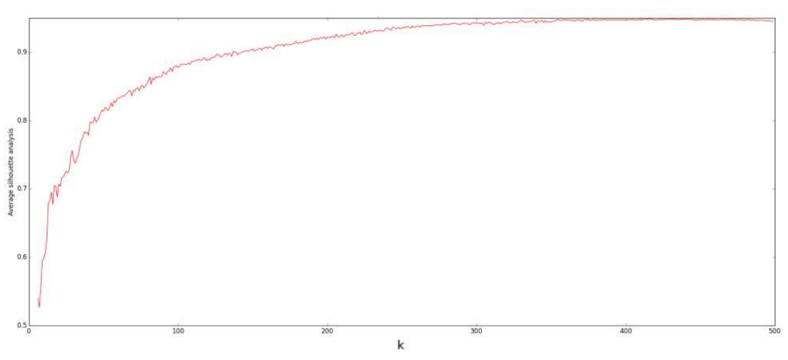


Fig.6: **E**volution of the average silhouette score computed on 5000 HD in RG context for k in [6,500] used for generation of automata in fig.4

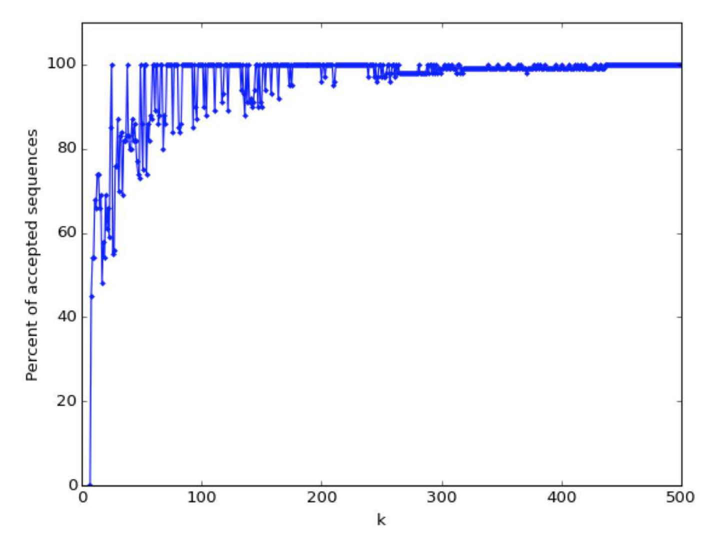


Fig.7: Percent of recognized sequences from RG. Analysis of extracted automata in fig.4

F – Perspectives

- During learning, the **network develops an implicit representation of the regularities , i.e. rules of the artificial non-binary grammars**. During testing, it **can predicts the output according to these implicit rules**. Results were confirmed in RG and ERG context
- Regarding **the question of neural networks interpretability**, it is possible to extract a representation in the form of graphs, with three different notation systems, each carrying **information on the internal functioning of the network** (explaining the prediction and behavior of the network)
- Preliminary results were obtained on real data extracted from electrical diagrams (Chraibi Kaadoud, 2018), and research is still in progress for a solid and general solution: simulations are ongoing on other artificial data, other real data, and different others grammars

References
(1) Squire, Larry R et Zola, Stuart M, 1996. Structure and function of declarative and nondeclarative memory systems. Proceedings of the National Academy of Sciences, 93(24) :13515–13522.
(2) Kim, JongW, Ritter, Frank E et Koubek, Richard J, 2013. An integrated theory for improved skill acquisition and retention in the three stages of learning. Theoretical Issues in Ergonomics Science, 14(1) :22–37.
(3) Elman, Jeffrey L, 1990. Finding structure in time. Cognitive science,14(2) :179–211.
(4) Cleeremans, A., McClelland, J.L.: Learning the structure of event sequences. Journal of Experimental Psychology: General 120(3), 235 (1991)
(5) Reber, Arthur S, 1967. Implicit learning of artificial grammars. Journal of verbal learning and verbal behavior, 6(6) :855–863.
(6) Gers, Felix A, Schmidhuber, Jürgen et Cummins, Fred, 1999. Learning to forget : Continual prediction with lstm.
(7) Chraibi Kaadoud, Ikram, 2018. "Sequence learning and rules extraction from recurrent neural networks : application to the drawing of technical diagrams." Theses. Université de Bordeaux, France