IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Modeling Implicit Learning: Extracting Implicit Rules from Sequences using LSTM

21th March 2022

Ikram Chraibi Kaadoud

Post-doctoral researcher in XAI

https://ikramchraibik.com/

https://cv.archives-ouvertes.fr/ikram-chraibi-kaadoud

TELECOM Paris

Ikram Chraibi Kaadoud

Post-Doctoral researcher in XAI

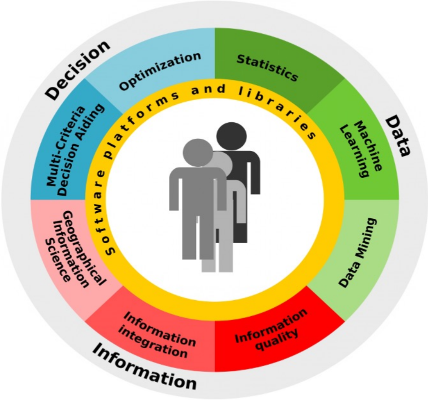
Ikram.chraibi-kaadoud@imt-atlantique.fr

XAI Interpretability Fairness Knowledge Extraction AI

Cognitive Sciences Ethics

Computational Neurosciences

DECIDE



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Lab-STICC

CNRS

UMR CNRS 6285

DECIDE Research team: https://www.labsticc.fr/en/teams/m-570-decide.htm
LAB-STICC Laboratory: https://www.labsticc.fr/en/index/
IMT-atlantique: https://www.imt-atlantique.fr/

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Outline

- Context : The role of Implicit learning in expertise
- Interpretable LSTM
- Results on different industrials contexts

Context :

The role of implicit learning

in expertise

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Implicit sequential learning in humans

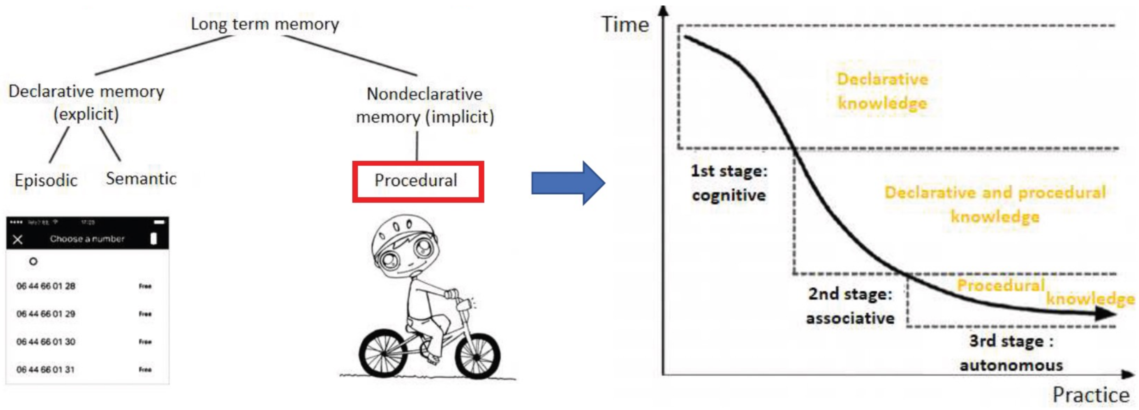


Fig 1 - A partial Taxonomy of different memories (Squire and Zola, 1996) and procedural knowledge acquisition according practice and time (Kim et al, 2013)

<u>Implicit knowledge</u> is a non-expressible knowledge of which the individual is not aware and that is acquired through implicit learning

Main characteristics of **implicit learning** are:

a) The encoded rules can not be categorized explicitly,

b) It impact the subsequent reasoning process when new rules are encoded,

c) There is **no notion of positive or negative example** learned through the implicit learning ability in the case of humans,

d) The knowledge, i.e the rules, is hidden in the temporal expression of behavior and more specifically in sequences of behaviourally significant events

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences*, *93*(24), 13515-13522.
Kim, J. W., Ritter, F. E., & Koubek, R. 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.

Interpretable LSTM

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Modeling Implicit Learning: Extracting Implicit Rules from Sequences using LSTM
Telecom Paris Seminar, March 2022
Ikram Chraibi Kaadoud

Some definitions

Common questions in Cognitive modeling using Machine Learning algorithms:
"What knowledge do they acquire? Why do they behave in a certain way ? what are the logic and aims behind their behaviour ?

Some definitions

Common questions in Cognitive modeling using Machine Learning algorithms:

"What knowledge do they acquire? Why do they behave in a certain way ? what are the logic and aims behind their behaviour ?

Interpretability

« The capacity of breaking down all the inner mechanisms of the black box (without necessarily understanding them) » (Doshi-Velez and Kim, 2017)

Explainability

« Given an audience, an explainable Arti cial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand. » (Arrieta et al.,2020).

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Some definitions

Common questions in Cognitive modeling using Machine Learning algorithms:
"What knowledge do they acquire? Why do they behave in a certain way ? what are the logic and aims behind their behaviour ?

Interpretability

« The capacity of breaking down all the inner mechanisms of the black box (without necessarily understanding them) » (Doshi-Velez and Kim, 2017)

Explainability

« Given an audience, an explainable Arti cial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand. » (Arrieta et al.,2020).

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Some definitions

Hierarchical representations in deep neural networks

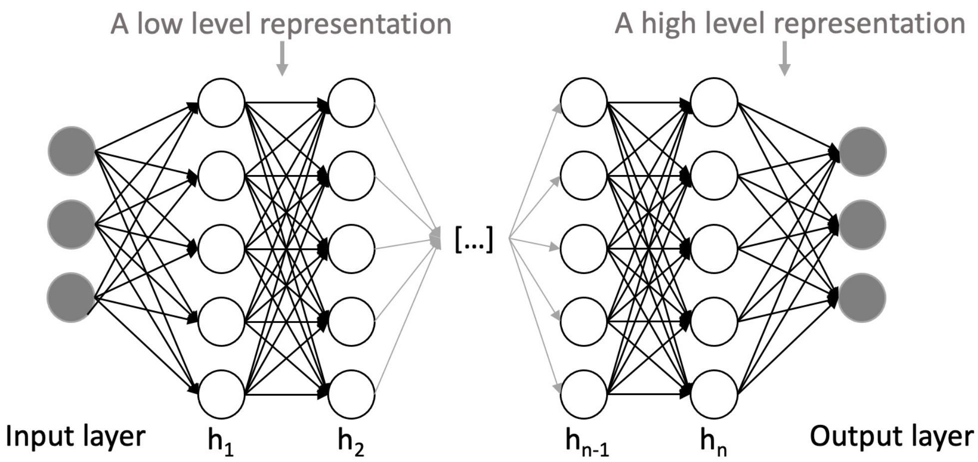


Fig 2 - Illustrative and schematic representation of the position of a low level representation and a high level representation in a deep neural network. hx refers to the xth hidden layer in the network. Image extracted from (Chraibi Kaadoud et al, 2021)

Latent space:
Abstract multidimensional space associated to each layer of a neural network where the representation of the learned data is implicitly built. Latent space contains the meaningful internal features representations of learned data, which makes it not directly interpretable.

Latent or hidden representation:
The data representation implicitly encoded by a neural network during the learning task and thus is hidden-layer dependant. It is a machine- readable data representation that contains features of the original data that have been learned by associated hidden layer.

Chraibi Kaadoud, I., Fahed, L., & Lenca, P. (2021, August). Explainable AI: a narrative review at the crossroad of Knowledge Discovery, Knowledge Representation and Representation Learning. In *Twelfth International Workshop Modelling and Reasoning in Context (MRC)@ IJCAI 2021*.

3-STEP METHODOLOGY

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

The **global experimental approach** for implicit knowledge extraction from RNN-LSTM in a task of prediction in three steps:

1) the learning phase where valid sequences generated from a grammar are used to train the network.

2) the knowledge extraction process,

3) the automata validation process where both valid and non-valid sequences are used

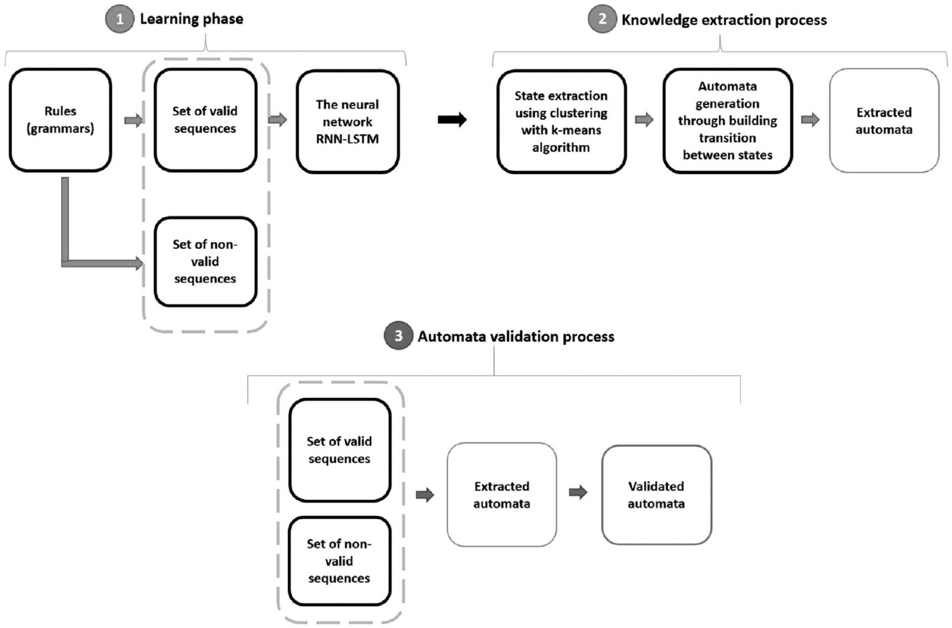
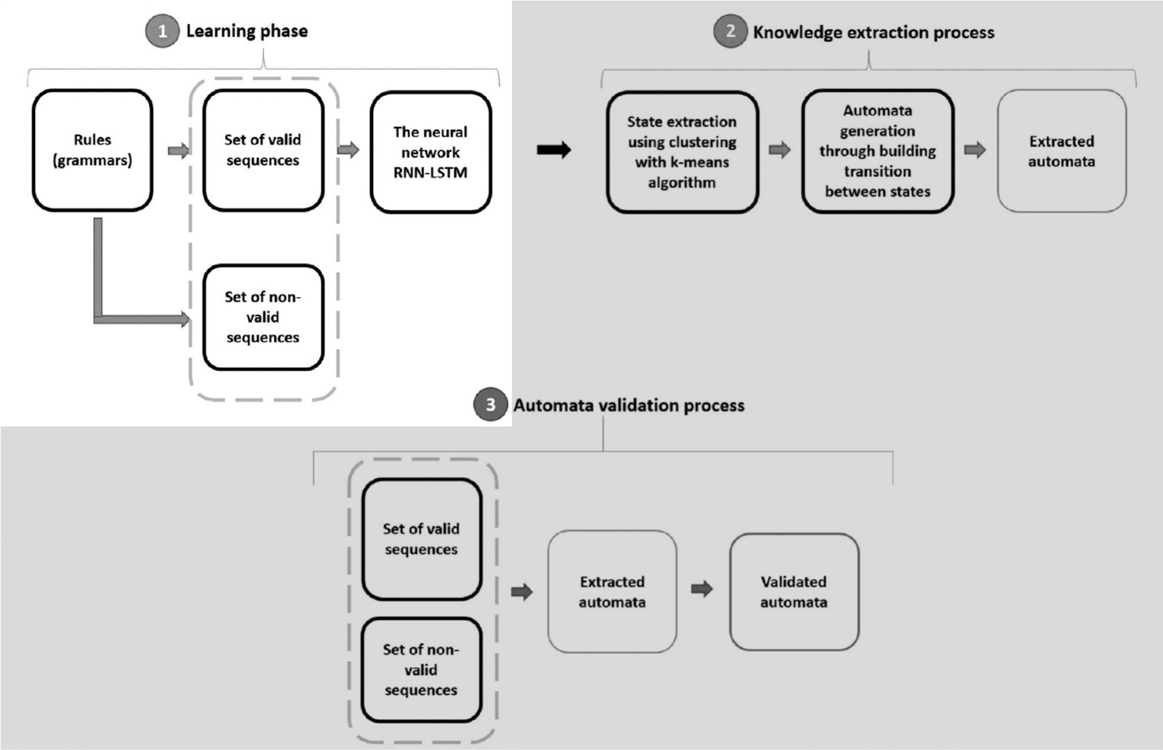


Fig 3 - The global experimental approach for implicit knowledge extraction from LSTM. Image extracted from (Chraibi Kaadoud et al, 2022)

Chraibi Kaadoud, I., Rougier, N. P., & Alexandre, F. (2022). Knowledge extraction from the learning of sequences in a long short term memory (LSTM) architecture. *Knowledge-Based Systems*, *235*, 107657.

Phase 1 – The learning phase of the RNN-LSTM



Dataset

The Reber grammar (RG), a grammar originally used in **cognitive psychology experiments** about implicit learning ability in humans, as well as its variants (ERG and CERG). (Reber, 1967)

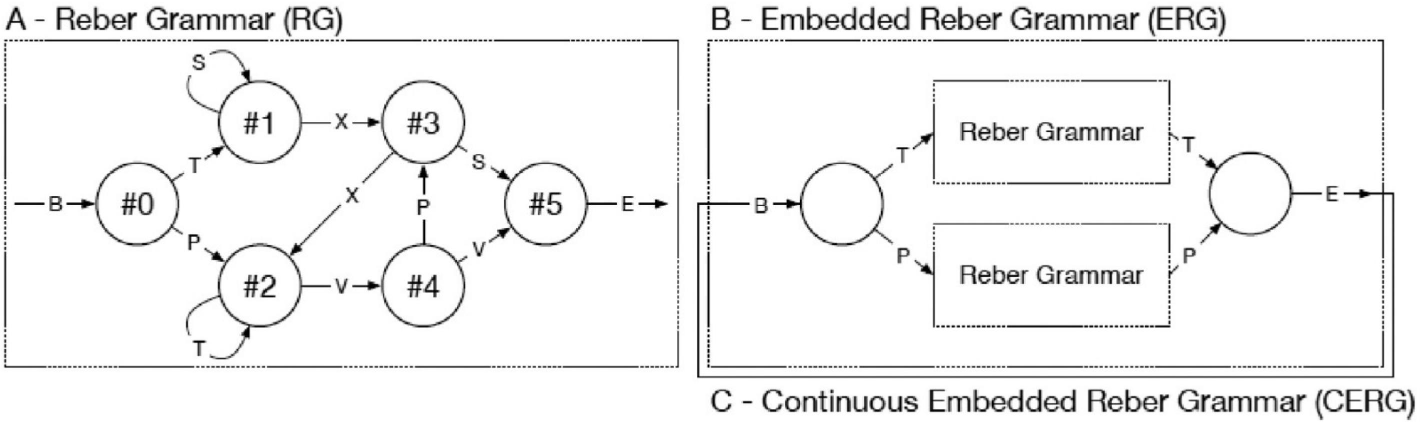


Fig 4 - The three grammars used in the experiments, represented as a Finite State Automaton including nodes representing states and bows emitting symbols. From left to right: A - Reber Grammar (RG), B - Embedded Reber Grammar (ERG), C - Continuous Embedded Reber Grammar (CERG). B means "Begin" and E means "End".

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of verbal learning and verbal behavior*, *6*(6), 855-863.

Dataset

RG and ERG grammars are used to generate sequences :

	RG	ERG
Grammatical/valid sequences Respect the rules of the gramma	BTXSE BTSXXVPSE BTSSSSSSSSSSSXSE	BPBTXSEPE BTBPVVETE BPBTXXTVVEPE
Non-grammatical/Non-valid sequences Random generation of sequences using the symbols	BE BVPXE BTPPPPPE	BPSE BSPPTTTTTTTTE

Both grammatical and non-grammatical sequences will be used to train and evaluate the RNN-LSTM performance:
Each sequence of lenght n is decomposed of n-1 pairs of symbole (current symbole-predicted symbol)
The RNN-LSTM should learn sequences and to predict the next symbol according the current one and the past ones

LSTM Unit

Steps of the forward propagation in an LSTM unit with one block, one cell :

A - The LSTM unit receives the activations of the other units of the network and then calculates the activities of the gates of the block.

B - The cell calculates the incoming activity received (input squashing) and modulates it according to the value of the input gate of the block (input gating).

C - Updating of the CEC value (memorizing) according to the modulated incoming activity of the cell and the CEC activity at the previous time modulated by the value of the block's forgetting gate (forgetting).

D - The cell calculates the activity resulting from the CEC (output squashing) and modulates it according to the value of the output gate of the block (output gating). According to the result, the cell sends an activation to the other units of the network.

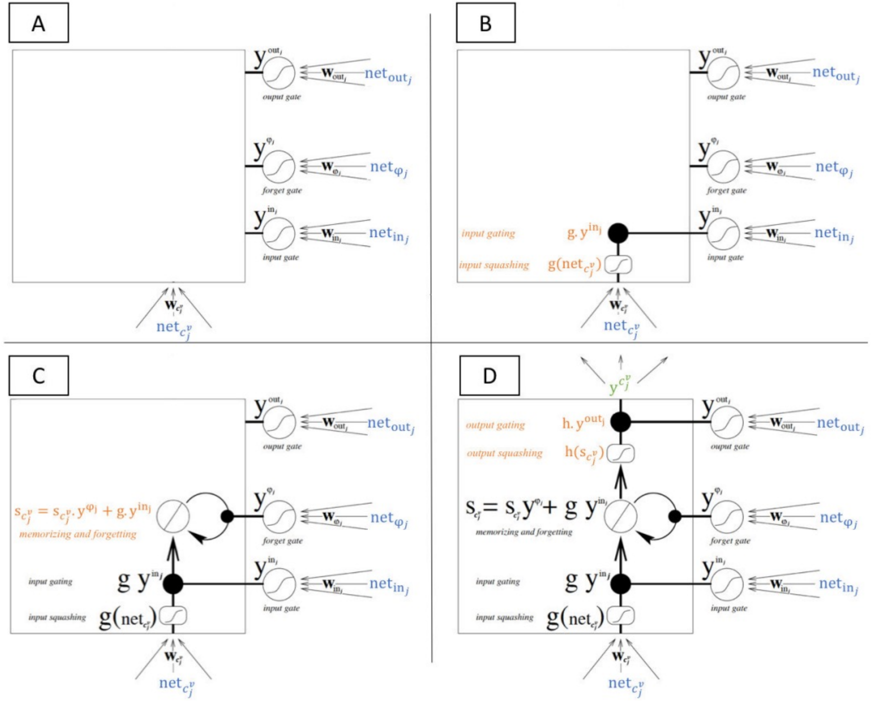


Fig 5 - Steps of the forward propagation in an LSTM unit with of a block of a cell. The incoming information in a LSTM unit is in blue, the outgoing in green. The calculations within a cell are in orange. Images taken from Chraibi Kaadoud, 2018

Chraibi Kaadoud, I. (2018). *apprentissage de séquences et extraction de règles de réseaux récurrents: application au traçage de schémas techniques* (Doctoral dissertation, Bordeaux).

The RNN-LSTM model

The RNN-LSTM model is composed of three layers :

- Input and output layers of artificial neurons
- A hidden layer four LSTM blocks with two cells and a CEC (Constant Error Carousel) in each.

In Fig. 6, all white dots outside LSTM blocks are linked to all black dots.
There are skipped connections between input and output units. The hidden layer provides a real-valued vector of size 8.

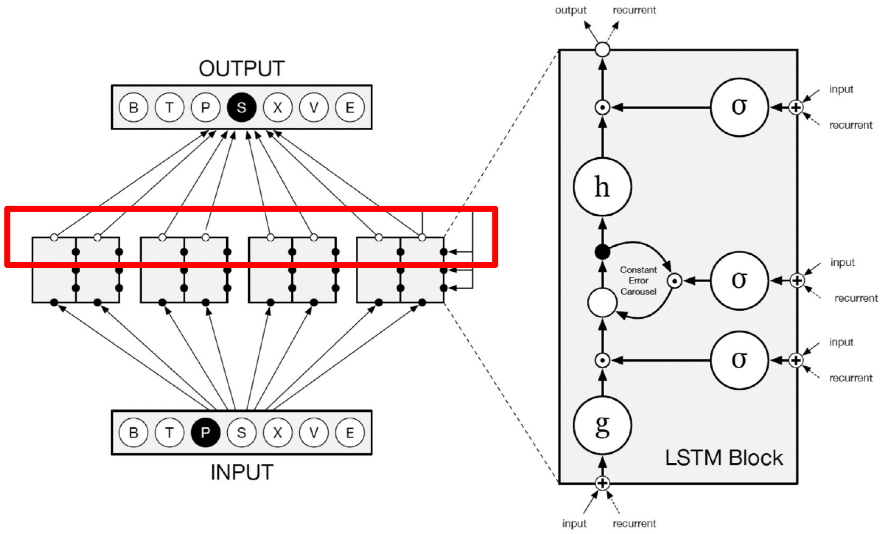


Fig 6 - The RNN-LSTM model with three layers.
Figure adapted from (Lapalme, 2006)

Lapalme, J. (2006). Composition automatique de musique à l'aide de réseaux de neurones récurrents et de la structure métrique.

The Learning process

The RNN-LSTM is trained on valid (i.e. grammatical) sequences of pairs of symbols.

During learning the model encodes hidden regularities from sequences, that corresponds to the transitions in the RG, ERG and CERG automata.

During testing, the network makes prediction according the latent representation of the grammar that it has encoded during learning.

The RNN-LSTM model learning and testing process:
1) Train the RNN-LSTM on grammatical sequences
2) Test it on grammatical sequences
3) Evaluate it on grammatical and non-grammatical sequences : only grammatical should be « accepted » by the model

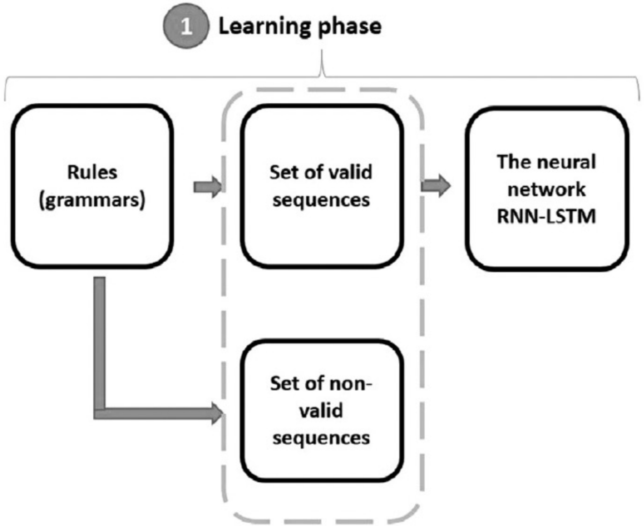


Fig 7 – Reminder of the first phase of the interpretability methodology

Test criteria & results on prediction task

A sequence is considered accepted if the network processes the entire sequence , ie, it predicts well the next symbol

A network with good performance =
High rate of accepted grammatical sequences (close to 100%)
Low rate of ungrammatical sequences accepted (close to 0%)

RG & ERG

Learning :
200 000 grammatical sequences
Test :
10 epochs of 20 000 sequences
10 epochs of 130 000 random sequences

LSTM : CERG

Test on 30 000 streams of 100 000 symbols
100 % of correct predictions

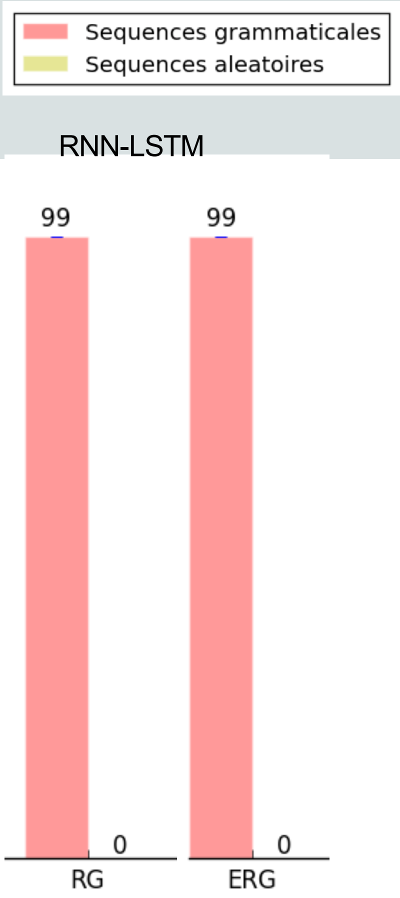
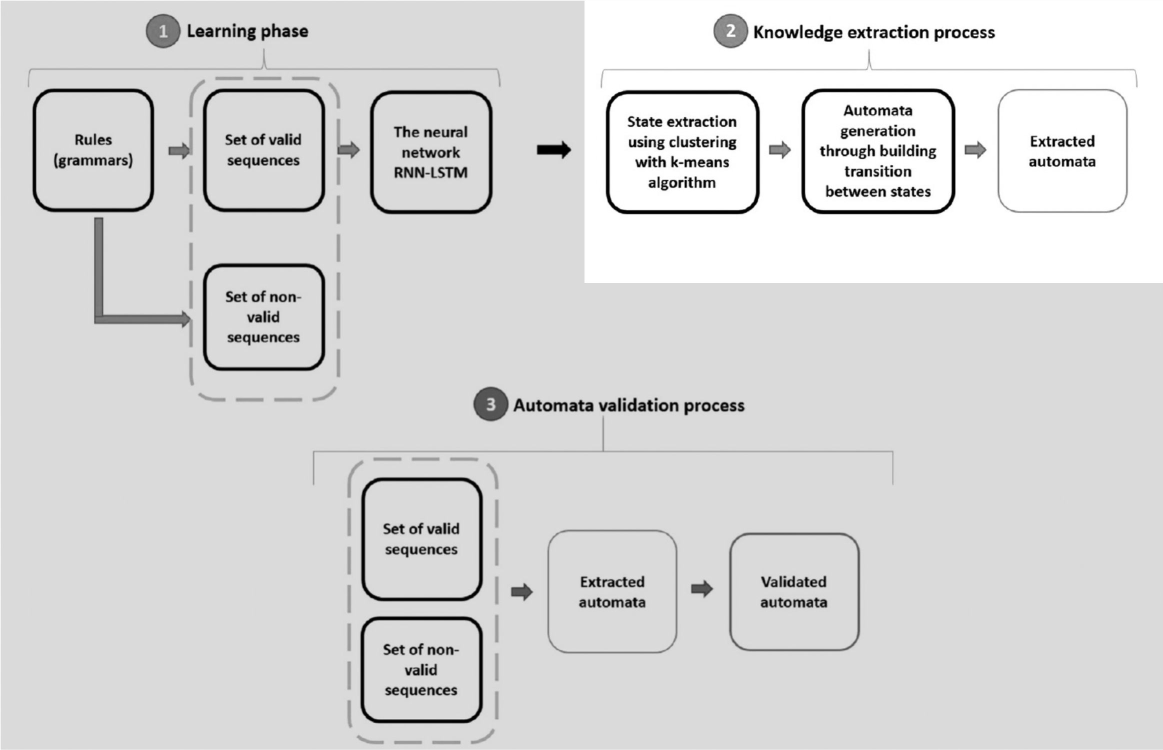


Fig 8 – Performance of the RNN-LSTM on sequence performance

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Phase 2 – The knowledge extraction process



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Automata generation

- Clustering **with k-mean algorithm** on the hidden activity patterns recorded during the test phase of the RNN-LSTM

  Result: At each **time step t,** an input lead to the generation of a hidden **pattern (P)** that is associated with a **cluster (C)**

Example :

Time	t_0	t_1	t_2	t_3	t_4
P	0	1	2	3	4
C	0	1	2	2	0

➔ Automata generation consists in the extraction of the encoded representation of the learned grammar from the latent space of the RNN-LSTM model hidden layer using the generated hidden patterns

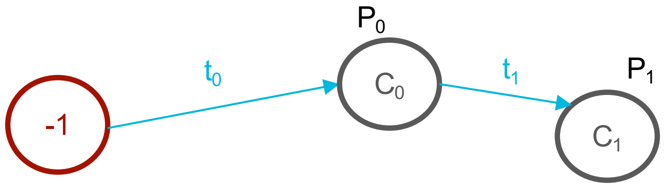
Algorithm 1 Algorithm for extracting rules in the form of a FSA with long labels, using the activity patterns of an RNN-LSTM

```
Require: # Learning and test of the RNN——
  RNN_LSTM.learning(learning_data_set)
  # labels_list : list of symbols presented to the network during tests
  activity_patterns_list, labels_list = RNN_LSTM.test(test_data_set)

  Function rules_extraction (activity_patterns_list, labels_list, k) :
  # Clustering ——————————————————
  clusters_list = k_means(k, activity_patterns_list)
  # Generation of automaton A———————————————————
  A = {} # Dictionary
  current_node= -1
  A['nodes'].add(current_node)
  A['edges'] = [ ] # list of dictionaries
  for all pattern h of index i from activity_patterns_list do
    associated_cluster = clusters_list[i]
    if associated_cluster ∉ A['nodes'] then
      A['nodes'].add(associated_cluster)
    end if
    edge= {} # Dictionary
    edge['id'] = (current_node, associated_cluster)
    if  edge ∉ GA['edges']  then
      new_edge = edge
      new_edge['weight'] = 1
      new_edge['label'] = labels_list[i]
      A['arcs'].add(nouvel_edge)
    else
      edge['weight'] = edge['weight'] +1
      edge['label'] = edge['label']+ labels_list[i]
      A['edges'].update(edge)
    end if
    # Update of the current node
    current_node= associated_cluster
  end for

  return  A

  End Function
```

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Automata generation – 1/4

Temps	t_0	t_1	t_2	t_3	t_4
P	0	1	2	3	4
C	0	1	2	2	0



P = Pattern

C = Cluster

The rule extraction process requires a simultaneous analysis of both list of atterns and list of associated:

Rule: If the associated cluster is a new one (i.e. not represented as a node in the FSA) :
a new node with its id as cluster number is added
a directed edge from the previous node to the new node is added

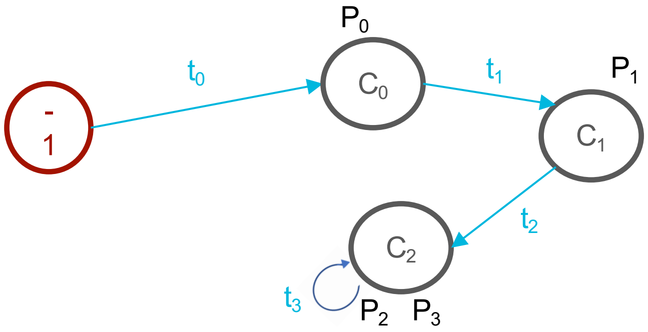
Example: P_1 generated at time t_1 and that belongs to cluster C_1 .

Automata generation – 2/4

Time	t_0	t_1	t_2	t_3	t_4
P	0	1	2	3	4
C	0	1	2	2	0

P = Pattern

C = Cluster



Rule: If two consecutive patterns belong to the same cluster, a recursive connection is added to the node representing the cluster

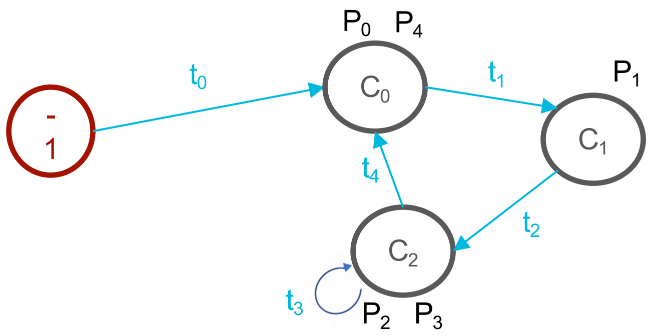
Example: P_2 and P_3 generated at time t_2 and t_3 and that belong to cluster C_2 .

Automata generation – 3/4

Time	t_0	t_1	t_2	t_3	t_4
P	0	1	2	3	4
C	0	1	2	2	0

P = Pattern

C = Cluster



Rule: If the current pattern belongs to a cluster already represented in the FSA then a directed edge between the previous node and the corresponding node is added

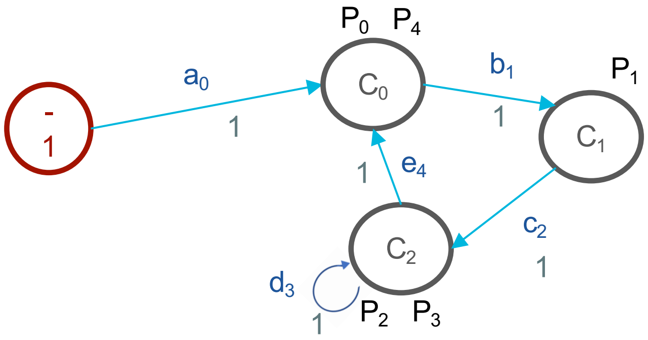
Example: P_4 generated at time t_4 and that belong to cluster C_0 .

Automata generation – 4/4

Time	t_0	t_1	t_2	t_3	t_4
symbol	a	b	c	d	e
P	0	1	2	3	4
C	0	1	2	2	0

P = Pattern

C = Cluster



Our contributions:
Addition of a -1 start node
Addition of "symbol + time step" labels on each transition
Increment the weight of the transition with each new label (+1)

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Results

Extracted automata on 33 hidden patterns with k=10:

(a) an unlabeled FSA

(b) a final FSA: a single label on each transition

(c) a long-label FSA: a long label on each transition

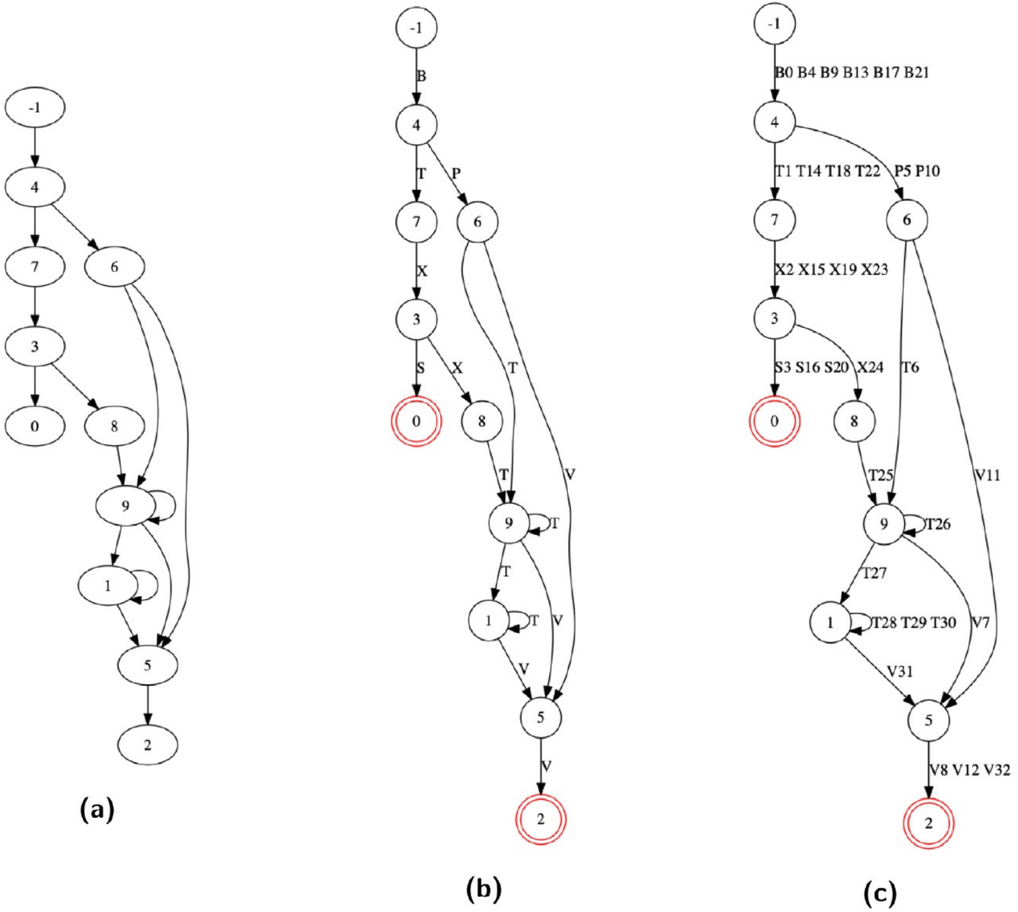


Fig 9 - Extraction in RG context on 33 hidden patterns. Final nodes, that indicate the end of sequences (i.e. that the following symbol is E), are noted with red double circles.

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Results

Important remark:

In the case of testing the model on a **small volume of data,** the extracted FSA will not represent all the implicit and encoded representation of all the learned data, **JUST** the part of the representation that corresponds to those inputs.

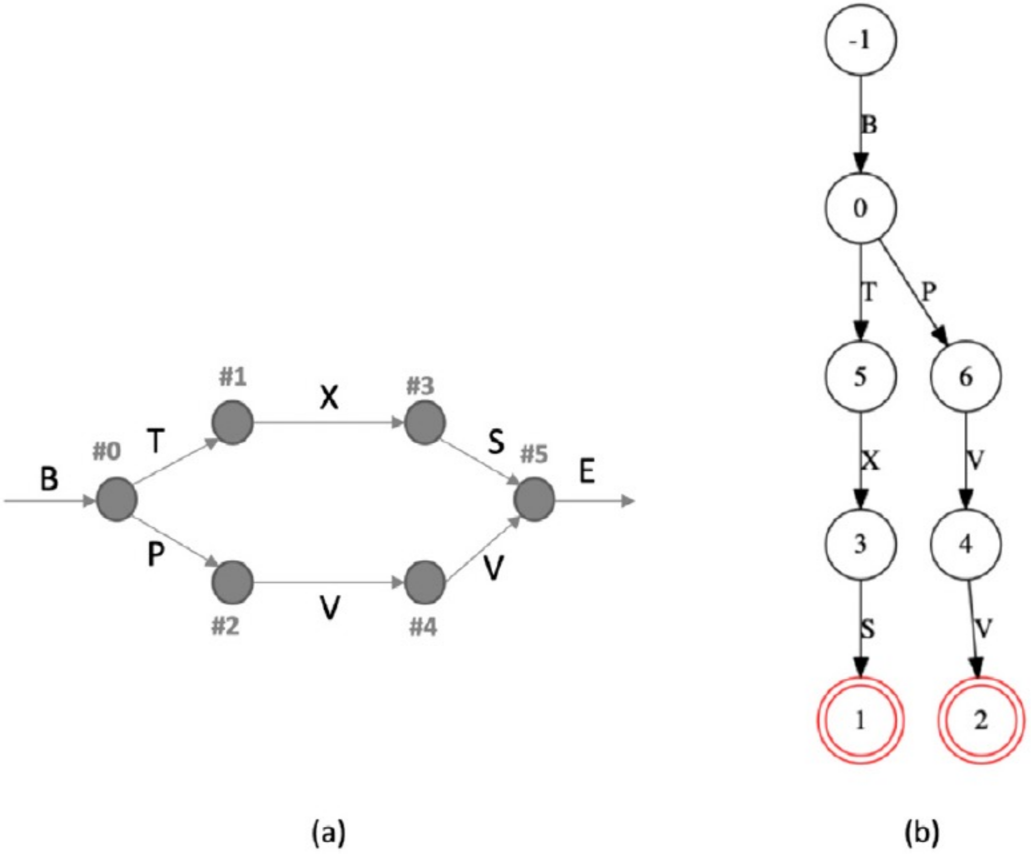
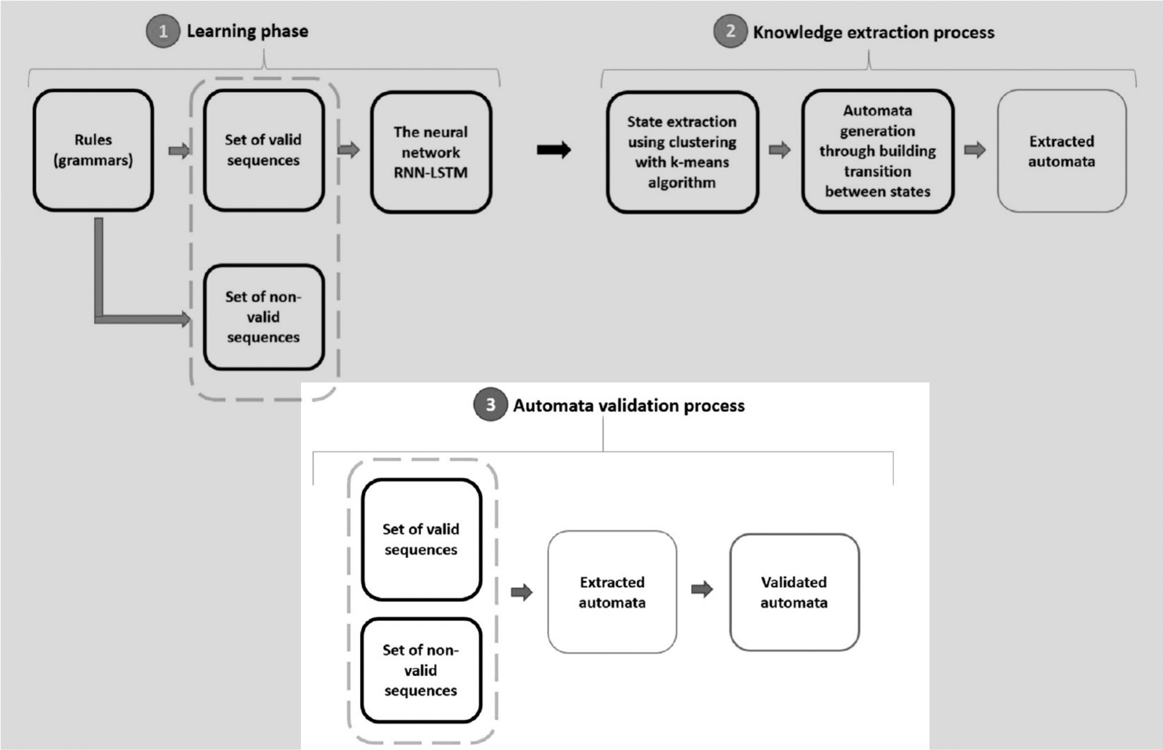


Fig 10 - Comparison of a portion of RG (a) and an extracted FSA for k=9 (b) for the 15 first time steps related to occurrences of 2 sequences: BPVVE and BTXSE. Final nodes are noted with red double circle on the extracted FSA.

Phase 3 – Validation of extracted automata



IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Validation process

The validation process follows the next steps:
For each sequence :

- the starting node is -1
- Application of the input to the extracted FSA to retrieve a new state.
- List of the neighbors (i.e. states) of this new state and their associated transitions :
  - If among these transitions, there is one corresponding to the next symbol of the sequence, the new state becomes the current state.
- The process is then repeated again, until the next symbol of the sequence is the last symbol of the sequence (i.e. symbol E).

If the FSA process the full sequence, it means that it recognize it, and that the long term dependencies of the original grammar.
If among the transitions of the neighbors, none of them corresponds to the desired next symbol, the sequence is rejected.

Difference between our validation approach and the SOTA*:
- No validation using positive and negative examples
- Verification of the preservation of the succession of symbols in a precise order and the sequential dependencies.

*State of the Art

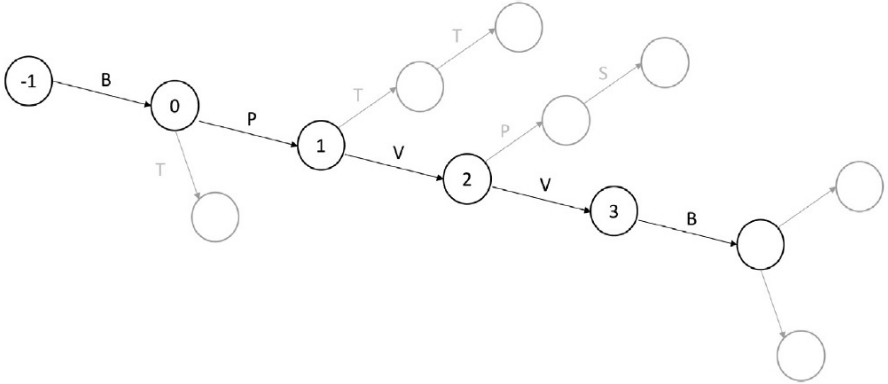


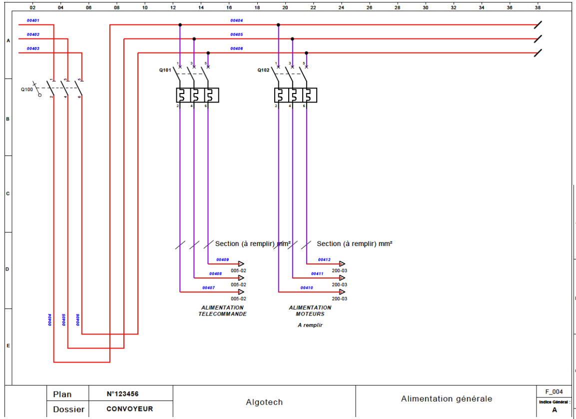
Fig 11 - Schematic representation of the testing process of the original sequence BPVVE from the Reber. Grammar on the extracted and minimized DFA. In black the selected path on the minimized DFA corresponding to the sequence, in gray the ignored ones

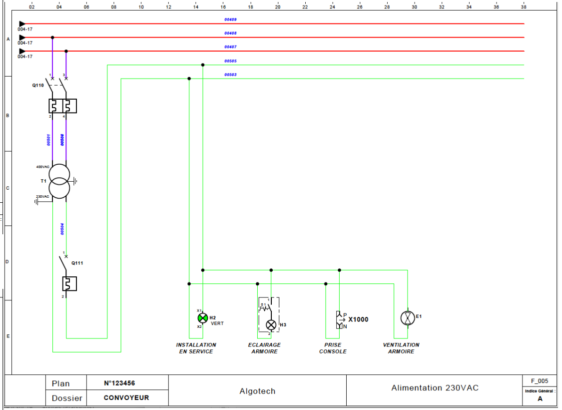
Conclusion:
➔the local context of a prediction is well learned and that the global representation of the network behavior over time is adequate with the original grammar.
➔Validation of the implicit encoding power of LSTMs.

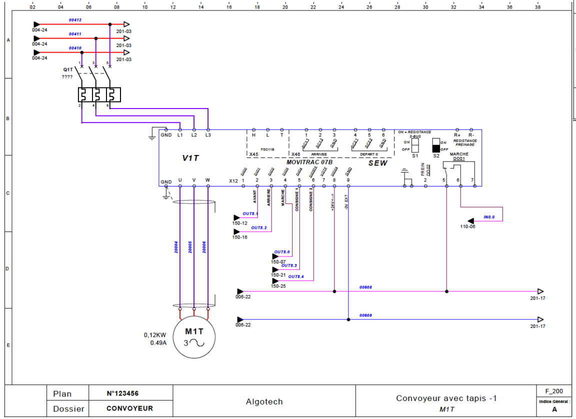
Results on differents industrial context

Expertise extraction from electrical diagrams

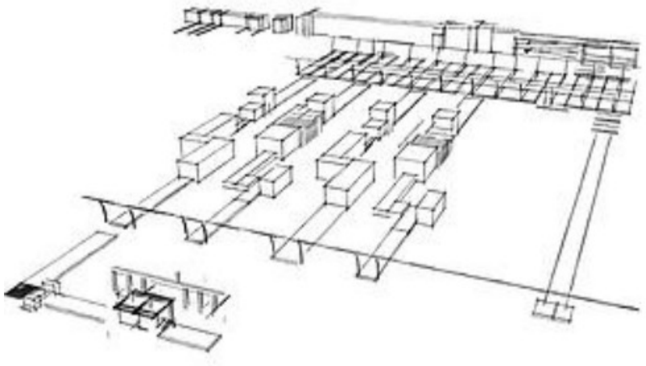
Exlactrical diagram : PDF files with one to more than 100 pages





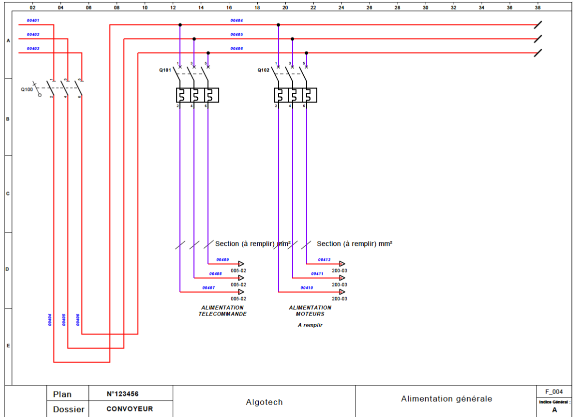


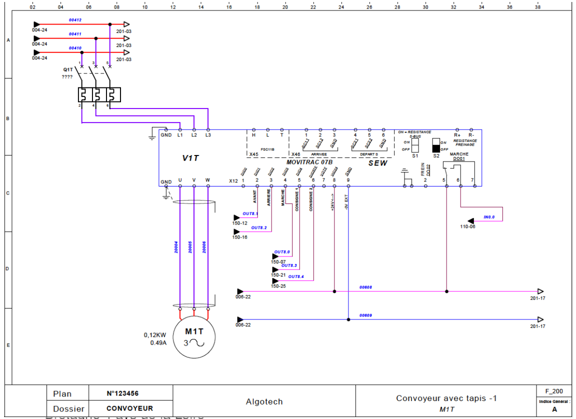
Un convoyeur



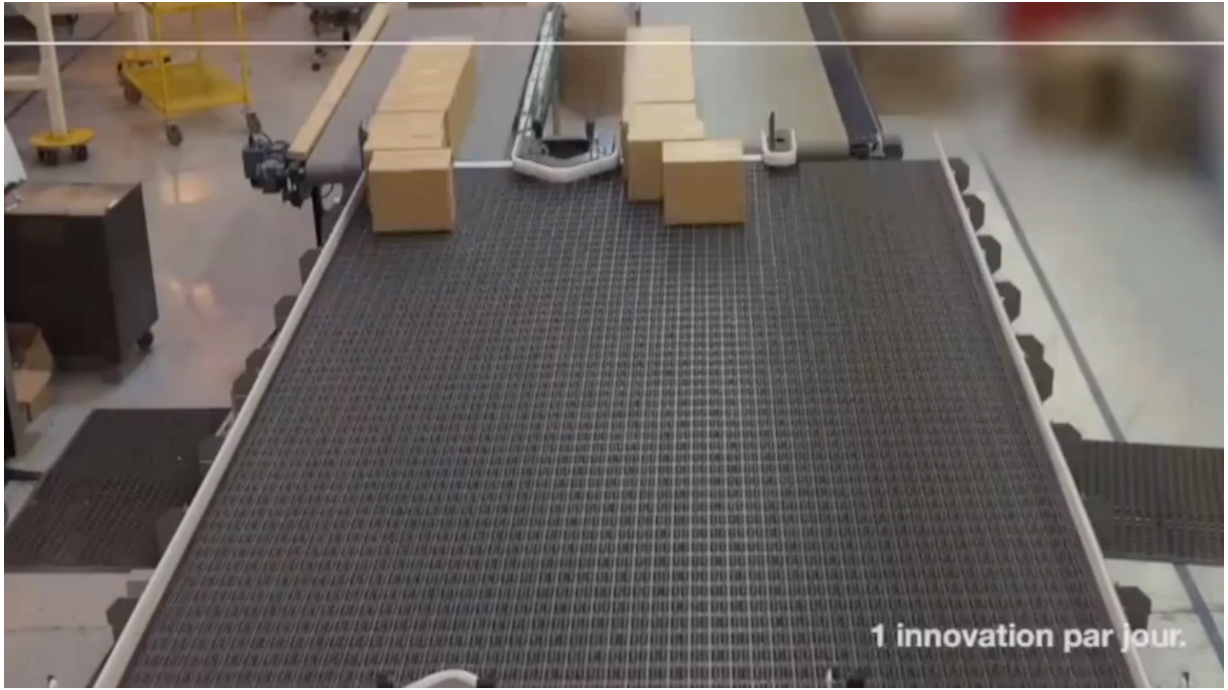
Expertise extraction from electrical diagrams

Exlactrical diagram : PDF files with one to more than 100 pages





A conveyor

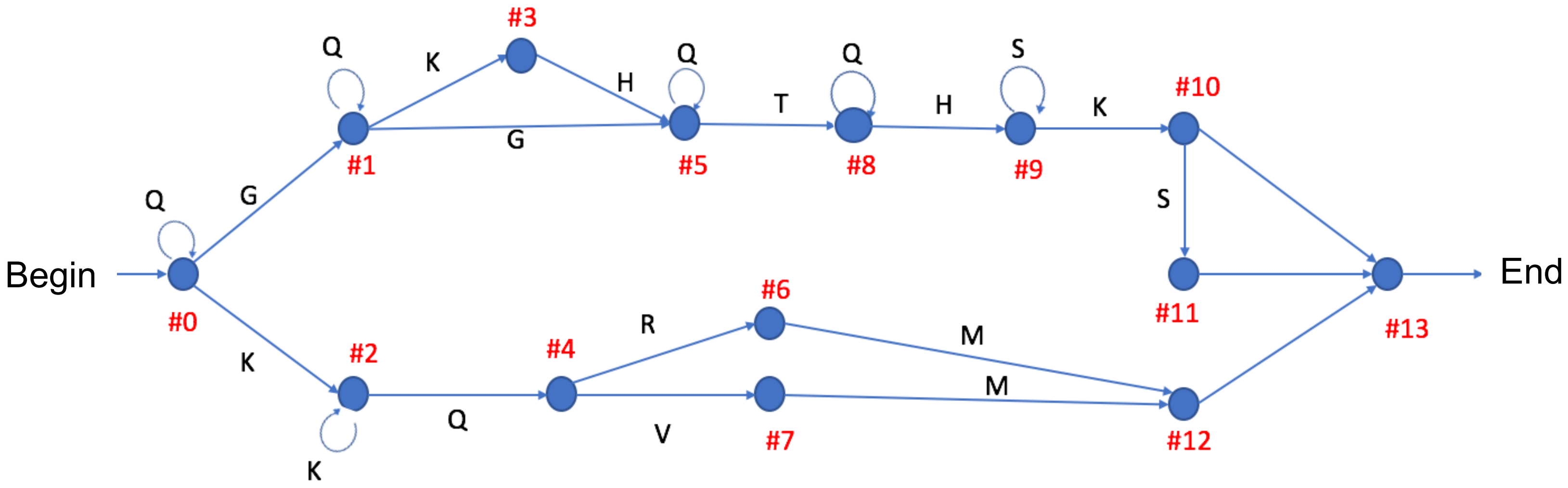


Source vidéo : https://www.youtube.com/watch?v=Ue7h_jMr2Is

Proposition of an electrical grammar

- New domain with unknown grammar
- Manual study of 3 separate diagrams (real cases):
- Scheme A (30 pages), Scheme B (31 pages) and Scheme C (86 pages)

- Manual generation of an electrical grammar (submitted and validated by Algo'Tech experts)



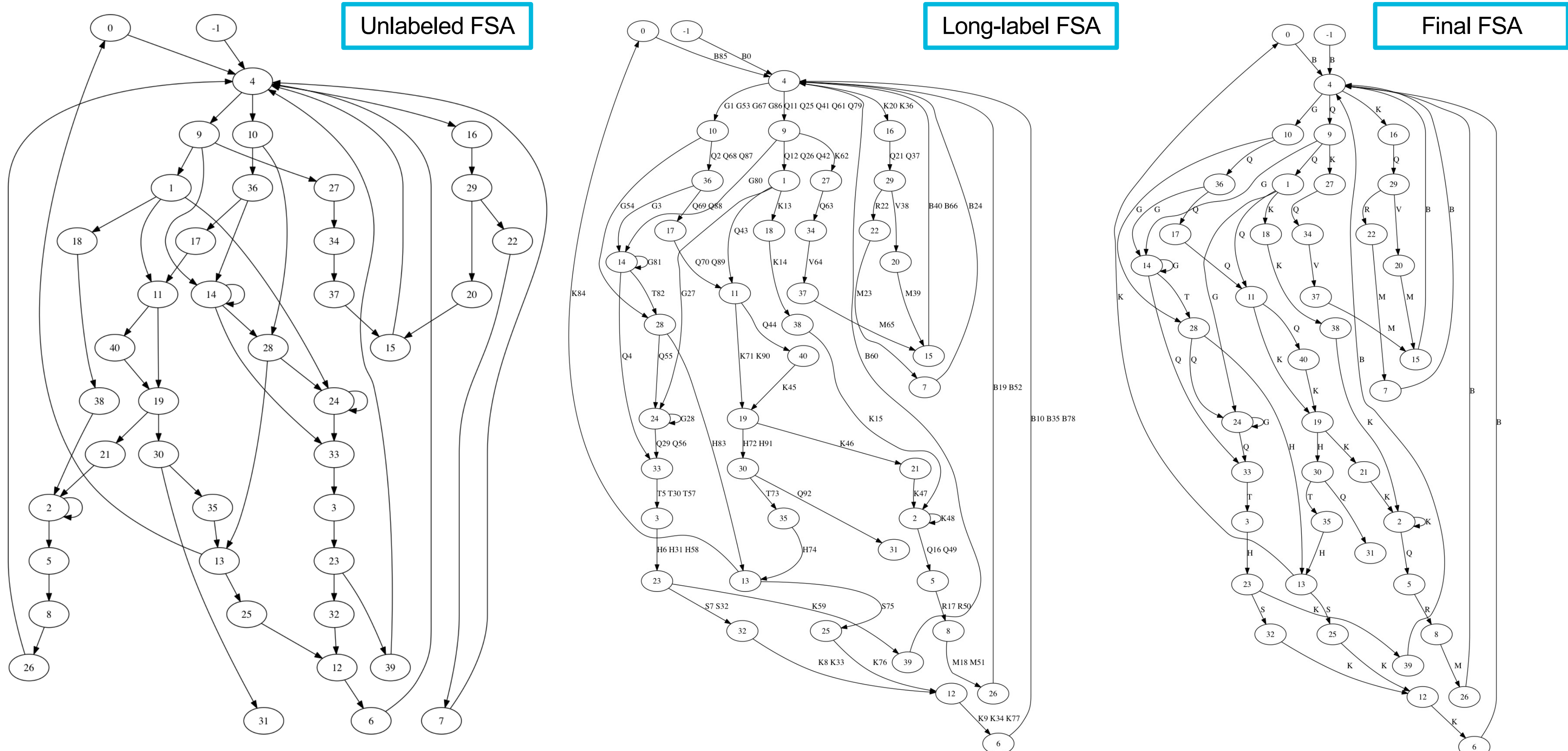
15 nodes & 25 edges
k-means in [2, 50]

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

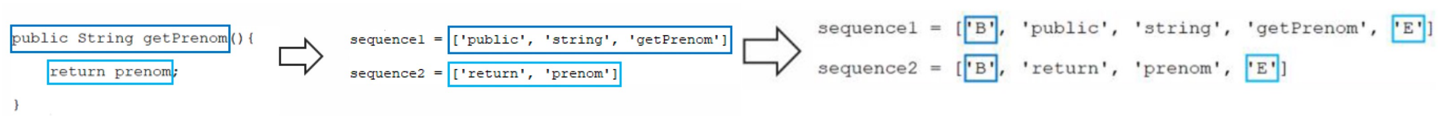
Chraibi Kaadoud, I., Rougier, N. P., & Alexandre, F. (2022). Knowledge extraction from the learning of sequences in a long short term memory (LSTM) architecture. *Knowledge-Based Systems*, *235*, 107657.
Chraibi Kaadoud, I. (2018). *apprentissage de séquences et extraction de règles de réseaux récurrents: application au traçage de schémas techniques* (Doctoral dissertation, Bordeaux).

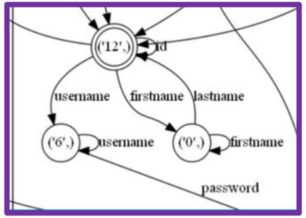
The construction of the electrical automaton

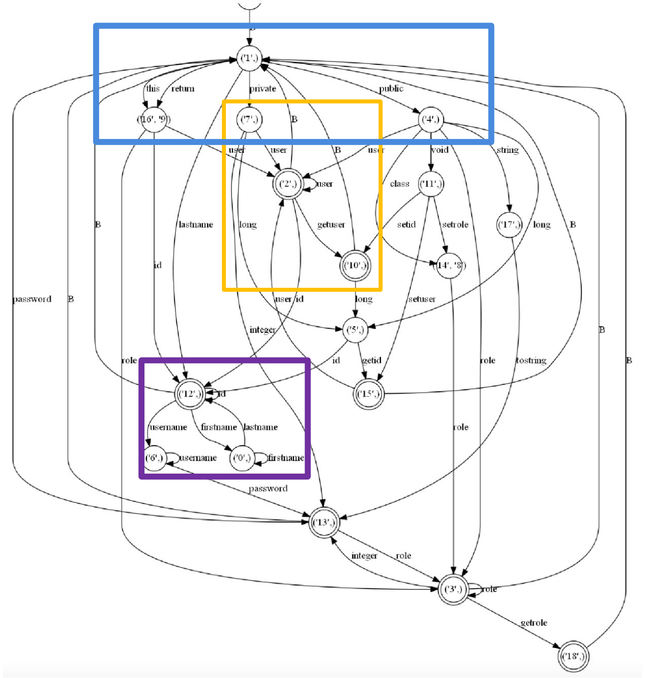
Extraction on the first 79 time steps with k=41 for the k-means

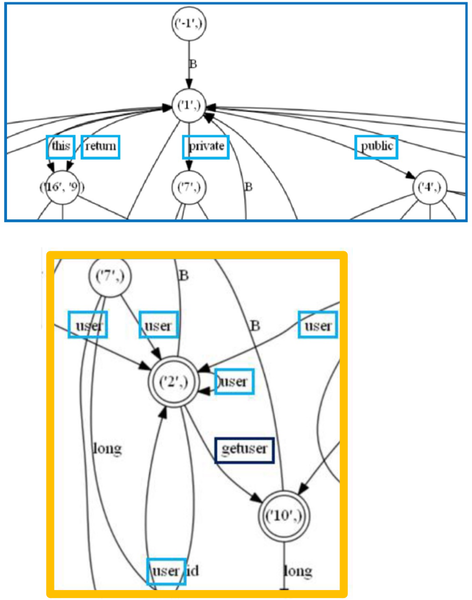


Expertise extraction from Java code









Link to the manuscrit of the student : https://github.com/MarineLH/interpretability_of_neural_networks/blob/master/M20_MARINELHUILLIER.pdf

Fig 12 – The extracted minimized automaton from sequences of java code

Thank you for your attention!

src: https://www.newyorker.com/cartoon/a19697

*“Does your car have an idea why my car pulled it over?”**

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

For further questions: ikram.chraibi-kaadoud@imt-atlantique.fr

Main references for this presentation :

- Chraibi Kaadoud, I., Rougier, N. P., & Alexandre, F. (2022). Knowledge extraction from the learning of sequences in a long short term memory (LSTM) architecture. *Knowledge-Based Systems*, *235*, 107657.
- Chraibi Kaadoud, I., Fahed, L., & Lenca, P. (2021, August). Explainable AI: a narrative review at the crossroad of Knowledge Discovery, Knowledge Representation and Representation Learning. In *Twelfth International Workshop Modelling and Reasoning in Context (MRC)@ IJCAI 2021*.
- Chraibi Kaadoud, I. (2018). *apprentissage de séquences et extraction de règles de réseaux récurrents: application au traçage de schémas techniques* (Doctoral dissertation, Bordeaux).

References:

- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Kim, J. W., Ritter, F. E., & Koubek, R. 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.
- Lapalme, J. (2006). Composition automatique de musique à l'aide de réseaux de neurones récurrents et de la structure métrique.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of verbal learning and verbal behavior*, *6*(6), 855-863.
- Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences*, *93*(24), 13515-13522.