Rethinking Rule Extraction from Recurrent Neural Networks

Towards an Artificial Scientific Intelligence

Henrik Jacobsson & Tom Ziemke

{henrik.jacobsson,tom.ziemke}@his.se
School of Humanities and Informatics,
University of Skövde, Sweden

NeSy @ IJCAI-05, Aug 1, 2005, Edinburgh

Outline

The "Golden Properties" of Simulated Systems

e.g. Recurrent Neural Networks (RNNs)

- Simulated systems are obviously suitable for studies. We can:
  - Duplicate them.
  - Replicate experiments as many times as we want.
  - Study effects of arbitrary perturbations.
  - Do nonperturbative studies.
  - Need more data? Simulate some more!
- Do these properties oblige researchers to utilize them?
- *Yes, each individual system can and should be studied empirically!*
- But... individual reseacherers will drown in the data!
- How do we solve this?
- *By exploiting these properties through rule extractor/model builder with a closed empirical loop!*

Problems of Earlier Algorithms

- Common constituents:
  1. *Quantisation* of state space.
  2. State and output *observation*.
  3. Rule *construction*.
  4. Rule *minimization*.
- But... no integration of the constituents.
- No tailor-made state space analysis.
- New approach: The Crystallizing Substochastic Sequential Machine Extractor, `CrySSMEx`.
  - Efficient.
  - Deterministic.
  - Parameter free.
  - Handles missing data.

Some preliminary results using `CrySSMEx`

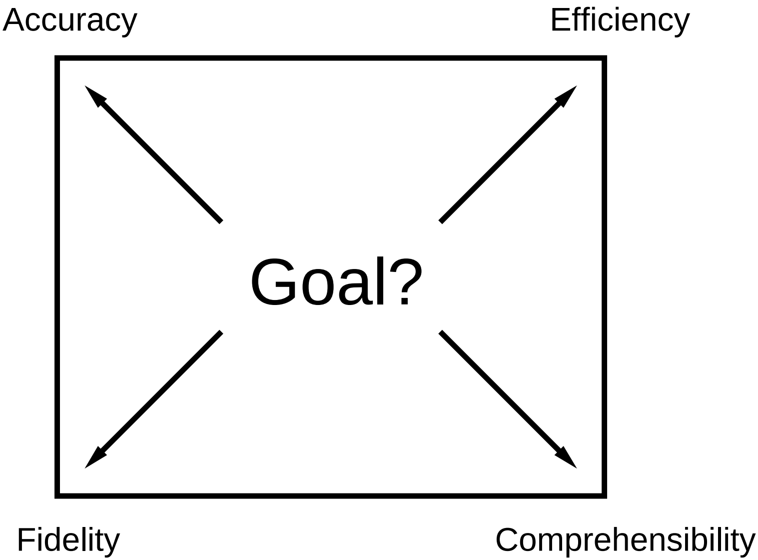
- Extraction from RNNs operating in regular language domains trivial.
- Extraction from RNNs predicting context free language $\mathbf{a}^n\mathbf{b}^n$ possible.
- Extraction of stochastic rules from chaotic systems possible.
- Extraction from high-dimensional RNNs possible (10^3 state nodes tested).
- Extraction results in a finite state machine, a hierarchical organisation of states, and a topological structure of RNN state space.

Conclusion: the empirical loop approach (partially implemented in `CrySSMEx`) seems to work...

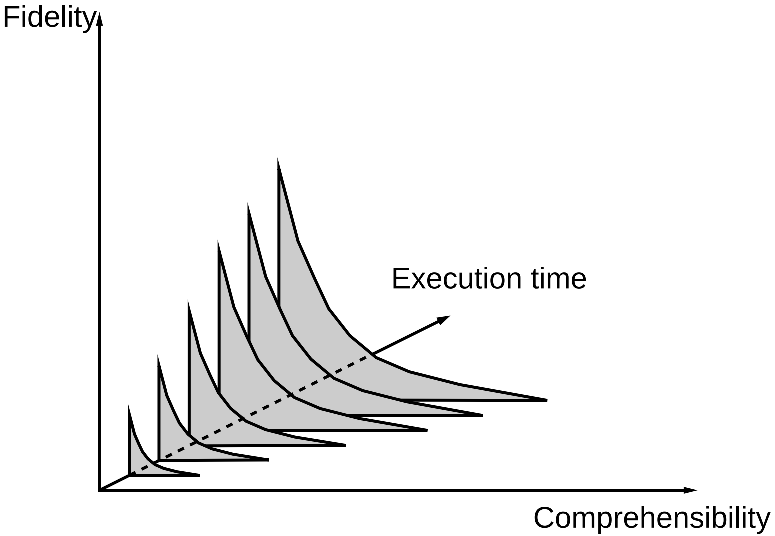
A “Wish-list” for Future Research

1. Focus not on RNNs but simulated systems in general.
2. User freedom: Comprehensibility, Fidelity, Accuracy & Efficiency a matter of choice.
3. Consistency over parameters.
4. Any-time extraction, gradual approximation.
5. Distance measure between systems.
6. Automatic subsystem identification.
7. Rules that can be queried: The power of a model is to be a proxy for queries.
8. Empirical Machines: complete the empirical loop, select or collect relevant data automatically.
9. Popperian Machines: generate falsifiable theories over models generated by one or more empirical machines, plan experiments that attempt falsification.

Competing Goals
Choice of the User

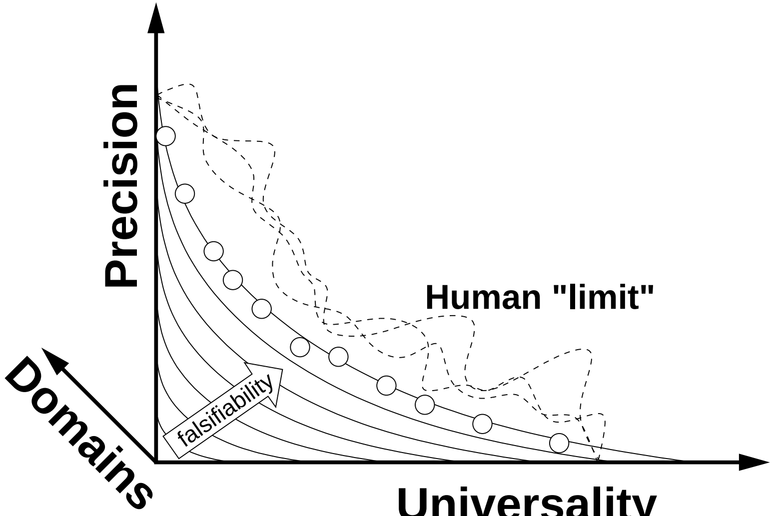


Fidelity/Comprehensibility/Time Tradeoff
The Revenue of Invested Time

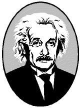


The Popperian Pareto Front

A Suggested Strategy for Generating Theories



Final Words



"A scientific theory should be as simple as possible, but no simpler"

—

Albert Einstein

- – — — — — — — – -

"For every complex problem, there is a solution that is simple, neat, and wrong"

—

Henry Louis Mencken

An Extracted Machine Example