Evaluation and Comparison of Inferred Regular Grammars

Neil Walkinshaw, Kirill Bogdanov and Ken Johnson

Department of Computer Science, University of Sheffield, UK

ICGI'08

Overview

- Aim is to infer regular grammar from (potentially sparse) sample of strings
  - Regular grammar can be represented as a Deterministic Finite Automaton (DFA)
- Perceived accuracy of these models has a major impact
  - Used for empirical comparison with other techniques (publications and competitions)
  - Genetic inference algorithms rely on accuracy as a fitness-function
- Conventional accuracy measure is flawed:
  - Test set is usually a random sample $\longrightarrow$ non-uniform coverage
  - Single value provides no insight into the strenghts/weaknesses of the technique

Overview

- Aim is to infer regular grammar from (potentially sparse) sample of strings
  - Regular grammar can be represented as a Deterministic Finite Automaton (DFA)
- Perceived accuracy of these models has a major impact
  - Used for empirical comparison with other techniques (publications and competitions)
  - Genetic inference algorithms rely on accuracy as a fitness-function
- Conventional accuracy measure is flawed:
  - Test set is usually a random sample $\longrightarrow$ non-uniform coverage
  - Single value provides no insight into the strenghts/weaknesses of the technique

Model Evaluation

- Models are evaluated by computing a test set, and working out the proportion of tests that are correctly classified

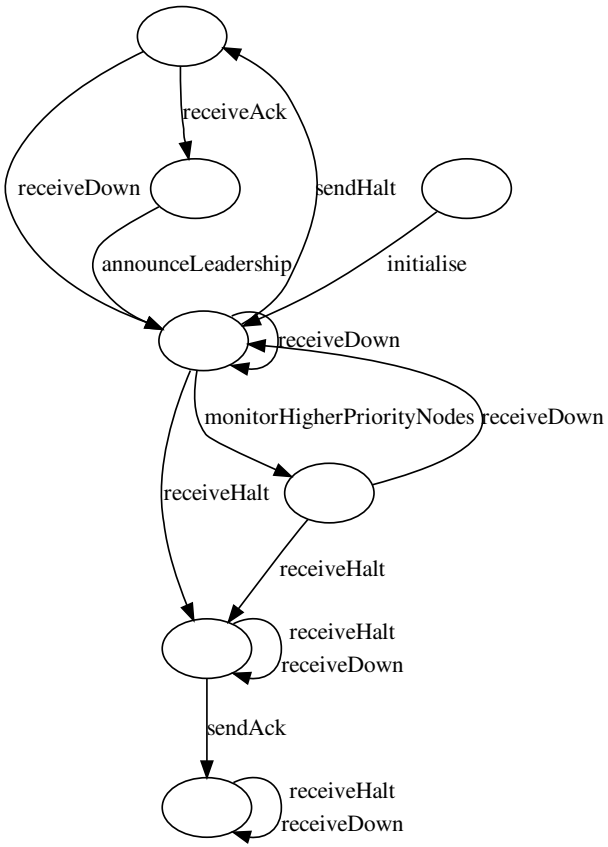
Test sampling

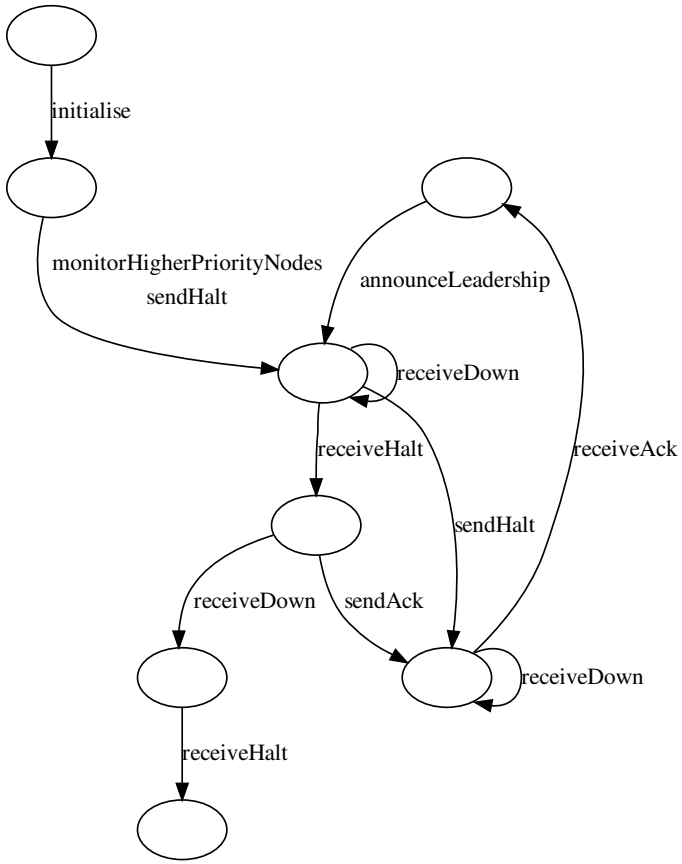
- Tests taken from **training set** or **random traces** over target model
  - Ensuring even split between accepting and rejecting sequences
  - Number of tests is related to the number of target states
  - Length of tests is often restricted to a uniform distribution related to the depth of the target model

Measuring accuracy

$$accuracy = \frac{correctly\ classified}{total\ classified}$$

Model of a software system





Model of a software system

Random test set

initialise receiveAck
initialise announceLeadership
initialise, monitorHigherPriorityNodes, receiveHalt, sendAck, receiveDown, receiveHalt
initialise, monitorHigherPriorityNodes, receiveHalt, sendAck, receiveDown, receiveDown
initialise, monitorHigherPriorityNodes, receiveDown, receiveHalt, sendAck
initialise, monitorHigherPriorityNodes, receiveDown, sendHalt, receiveAck receiveAck
initialise, receiveHalt, sendAck, receiveHalt, receiveHalt, receiveHalt, receiveHalt
initialise, receiveHalt, sendAck, receiveHalt, receiveHalt, receiveDown, receiveHalt, receiveHalt
initialise, receiveHalt, sendAck, receiveHalt, receiveDown announceLeadership
initialise, receiveHalt, receiveHalt, receiveDown, receiveHalt sendHalt
initialise, receiveHalt, receiveDown, sendAck, receiveDown, receiveHalt
initialise, sendHalt, receiveAck
initialise, sendHalt monitorHigherPriorityNodes
initialise, sendHalt, receiveDown, receiveHalt, receiveHalt receiveAck
initialise, sendHalt, receiveDown, sendHalt, receiveAck
initialise, receiveDown, monitorHigherPriorityNodes, receiveDown receiveAck
initialise, receiveDown, receiveHalt, sendAck sendHalt
initialise, receiveDown, receiveHalt, sendAck, receiveDown, receiveDown
initialise, receiveDown, sendHalt
initialise, receiveDown, receiveDown, monitorHigherPriorityNodes, receiveHalt initialise
initialise, receiveDown, receiveDown, monitorHigherPriorityNodes, receiveHalt, receiveDown receiveAck
initialise, receiveDown, receiveDown, sendHalt, receiveAck, announceLeadership
initialise, receiveDown, receiveDown, sendHalt, receiveDown, sendHalt, receiveAck
announceLeadership

Model of a software system

Result

41.6%

Problems

Validity

- To be valid the test set must be characteristic of the target model
- Unlikely to be the case with random tests
  - Random samples more likely to exercise certain aspects of a grammar than others
  - Only of 2.4% of random strings positive in Tomita1 (Bongard & Lipson)
  - Random traces required for approximate inference grows exponentially with target

Interpretation

- Interpretation is problematic
  - Is model under or over generalised?
  - Does it produce too many false positives or negatives?

Model-Based Testing

Background

- Popular for testing network protocols (and other systems)
- Compare two models (implementation and specification)
  - Use specification to generate test sequences that will identify discrepencies with the implementation model

Common assumptions

- The models are minimal and deterministic
- The number of extra states in the implementation can be guessed
- There is a reliable reset in the implementation

W-Method

- Large number of model-based testing techniques
  - Each technique caters for a different set of assumptions
  - W-Method, HSI, UIO...
- W-Method (Chow'78 and Vasilevskii'73)
  - Guaranteed to find any missing/extra states or transitions
  - Tests reach every state, and then ensure that the behaviour of that state is correct
    - Majority of tests expected to fail - ensure that failure occurs at right point
  - Produces large test sets
    - Negligible when tracing a path in an inferred grammar

Summary

W-Method test sets are absolutely authoritative, as opposed to random samples

W-Method

- Large number of model-based testing techniques
  - Each technique caters for a different set of assumptions
  - W-Method, HSI, UIO...
- W-Method (Chow'78 and Vasilevskii'73)
  - Guaranteed to find any missing/extra states or transitions
  - Tests reach every state, and then ensure that the behaviour of that state is correct
    - Majority of tests expected to fail - ensure that failure occurs at right point
  - Produces large test sets
    - Negligible when tracing a path in an inferred grammar

Summary

W-Method test sets are absolutely authoritative, as opposed to random samples

W-Method

- Large number of model-based testing techniques
  - Each technique caters for a different set of assumptions
  - W-Method, HSI, UIO...
- W-Method (Chow'78 and Vasilevskii'73)
  - Guaranteed to find any missing/extra states or transitions
  - Tests reach every state, and then ensure that the behaviour of that state is correct
    - Majority of tests expected to fail - ensure that failure occurs at right point
  - Produces large test sets
    - Negligible when tracing a path in an inferred grammar

Summary

W-Method test sets are absolutely authoritative, as opposed to random samples

Precision and Recall

Background

- Used in Information Retrieval to measure overlap of **REL**evant and **RET**rieved documents
  - Exactness: $Precision = \frac{|REL \cap RET|}{|RET|}$
  - Completeness: $Recall = \frac{|REL \cap RET|}{|REL|}$

To evaluate a regular grammar

- Measure overlap of test classifications in target and inferred grammar
- What do we add to *RET* and what do we add to *REL*?
  - Want to count both tests that are correcly accepted and correctly rejected - need two sets: RET^+, REL^+, RET^-, REL^-.
  - For a single machine this will produce positive and negative precision and recall

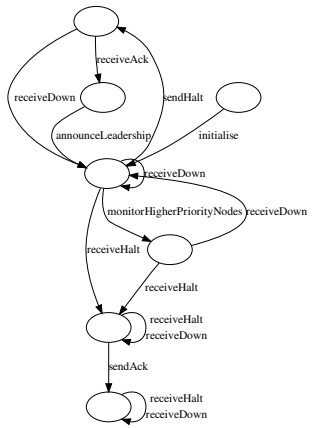
Precision and Recall

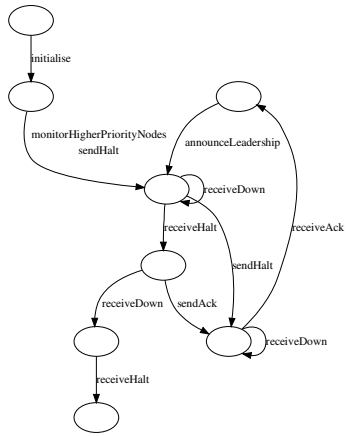
Background

- Used in Information Retrieval to measure overlap of **REL**evant and **RET**rieved documents
  - Exactness: $Precision = \frac{|REL \cap RET|}{|RET|}$
  - Completeness: $Recall = \frac{|REL \cap RET|}{|REL|}$

To evaluate a regular grammar

Hypothesis	Target	RET^+	REL^+	RET^-	REL^-
accept	accept	×	×		
accept	reject	×			×
reject	accept		×	×	
reject	reject			×	×

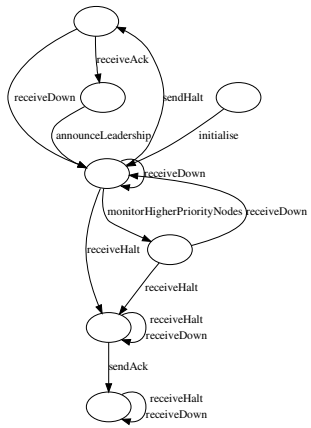


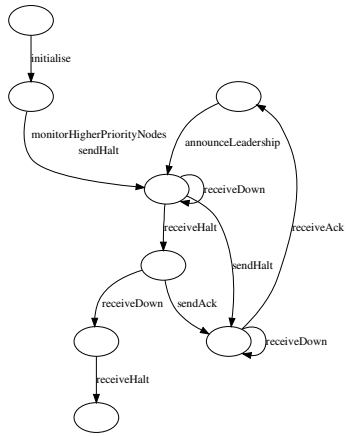


Comparison

Accuracy: 41.6%

	Precision	Recall
+	78.7	29.8
-	65.9	94.3



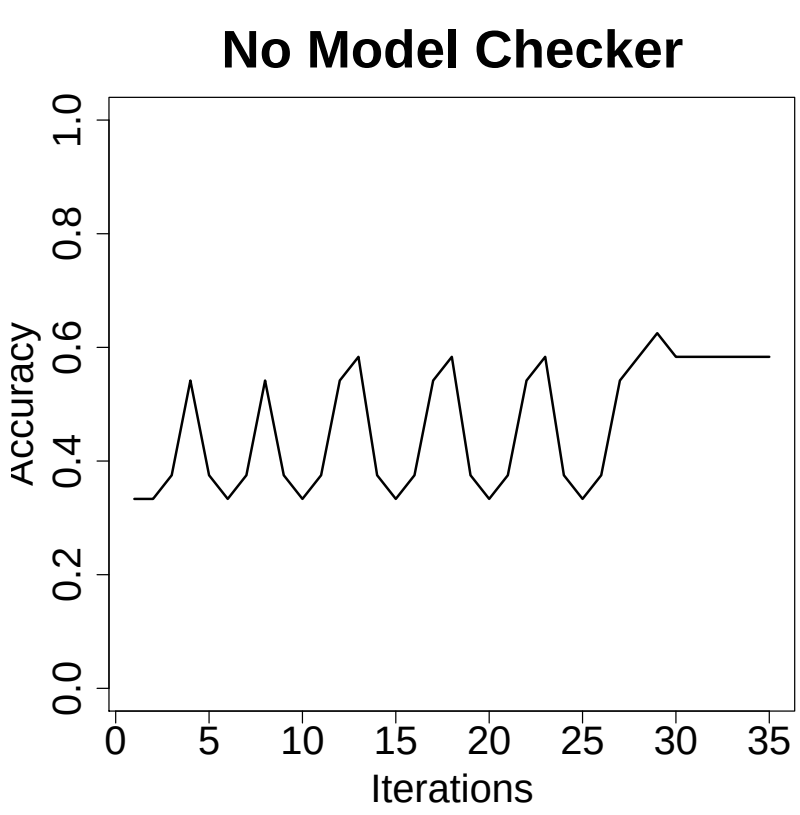


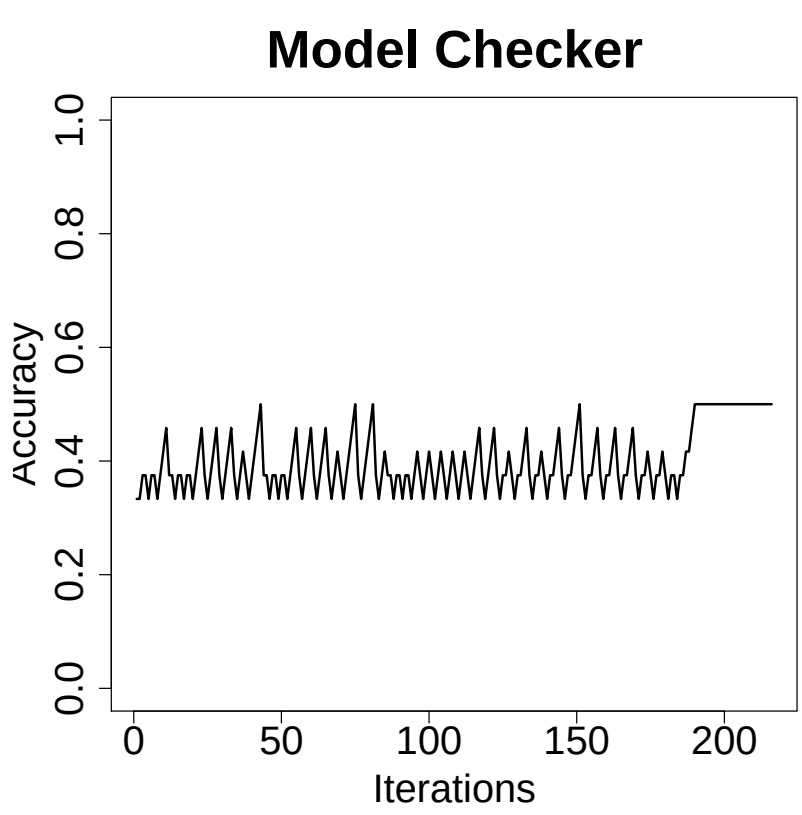
Comparison

Accuracy: 41.6%

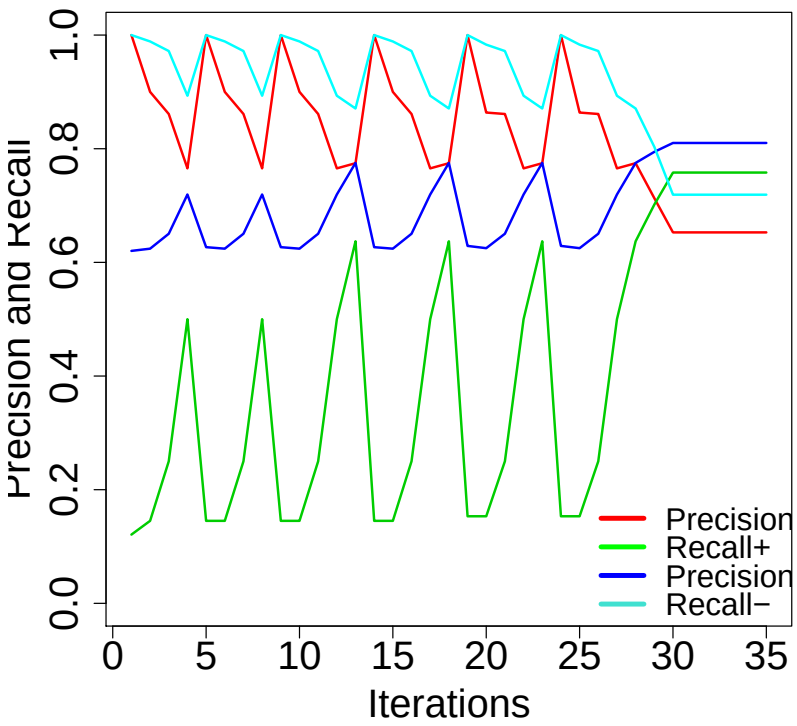
	Precision	**Recall**
+	78.7	29.8
-	65.9	94.3

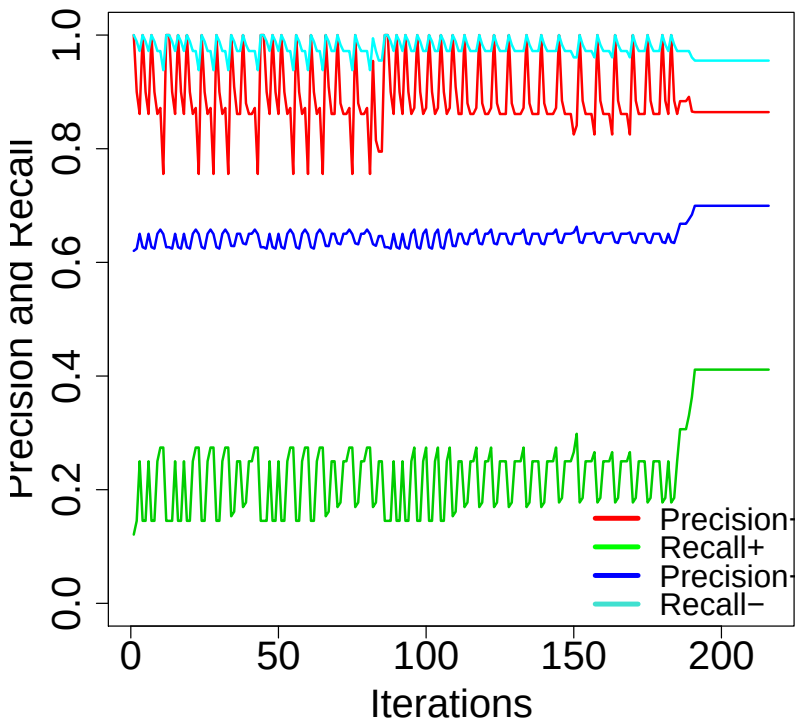
Tracking the Accuracy of the Inference Process





Tracking the Accuracy of the Inference Process





Discussion

Why this is useful

- Authoritative - Does not rely on sampling to produce a suitable test set
- More descriptive - helps to explain why a hypothesis is (in-)accurate
- Precision and Recall does not rely on an even split between positive / negative strings

Future Work

- Using PR in fitness functions for genetic grammar inference algorithms
- Investigate evaluation in terms of state machine structure as opposed to language-based measures