Automated Grading and Feedback for DFA constructions

Rajeev Alur (Penn), Loris D’Antoni (Penn), Sumit Gulwani (MSR), Bjoern Hartmann (Berkeley), Dileep Kini (UIUC), Mahesh Viswanathan (UIUC)
Twice ab

Draw a DFA that accepts the following language over the alphabet \( \{a, b\} \): all strings in which \( 'ab' \) appears exactly twice as a substring.

Answer:

![DFA Diagram]

- States: 0, 1, 2
- Initial State: 0
- Final States: 2
- Transitions:
  - 0: a \(\rightarrow\) 1
  - 1: a \(\rightarrow\) 1, b \(\rightarrow\) 2
  - 2: a \(\rightarrow\) 1, b \(\rightarrow\) 2
Problem Syntactic Mistake

• The student misunderstood the problem
• at least 2 occurrences of ‘ab’ instead of
• exactly 2 occurrences of ‘ab’
• Tool answer
  l Feedback: The correct language is \{ s \mid \text{‘ab’ appears in } s \text{ exactly 2 times} \}
  l Grade: 5/10
• Technique
  l Synthesize a logic (MSO) description of the student attempt and the problem solution (brute force and pruning)
  l Compute tree edit distance between the two descriptions to produce grade
  l Use the logic descriptions and the edit script to produce the feedback (highlight edits)
Solution Syntactic Mistake

• The student forgot to make state 8 final, otherwise the solution is correct

• Tool answer
  Feedback: One more state should be made final
  Grade: 9/10

• Technique: compute DFA edit distance for grading
  Number of edits necessary to transform the DFA into a correct one
  An edit is
    Make a state (non)final
    Add a new state
    Redirect a transition
  Brute force based (usually DFA is small)
Solution Semantic Mistake

• The solution is wrong on most of the strings
• Tool answer
  - **Feedback:** The DFA is incorrect on the string ‘ababb’
  - **Grade:** 6/10
• Technique
  - Compute size of the set of misclassified string:
    - $S =$ correct solution, $A =$ student attempt
    - Difference: $D = S \setminus A \cup A \setminus S$
    - $\text{Size}(D,S) = \lim_{n \to \infty} \frac{D_n}{S_n}$
    - Approximate to finite $n$
    - Feedback with counterexample in $D$ (if possible synthesize language difference)
Experimental results

• Compared with human graders on 800 real student attempts
• Identical solutions receive same grades and correct attempts awarded max score (unlike human)
• 90% cases consistent with human grader (+/- 3 points)
• On disagreeing cases, human grader often realized that his assigned grade was inaccurate after reading tool’s feedback
• Always assigns full score to correct

• Tool limited to small DFAs (< 10 states) and small alphabets (< 3 symbols).
• Not a big limitation in practice.
Ongoing work

• Evaluation of quality of feedback
• Test the tool on Automata Theory courses in Fall at Penn and UIUC
• Grading and Feedback for
  • Regular expressions,
  • NFAs,
Thank you

• lorisdan@cis.upenn.edu

• Questions?