Disambiguation of Regular Expressions with Backreferences via Term Rewriting

PSSV
November 3th-4th, 2023

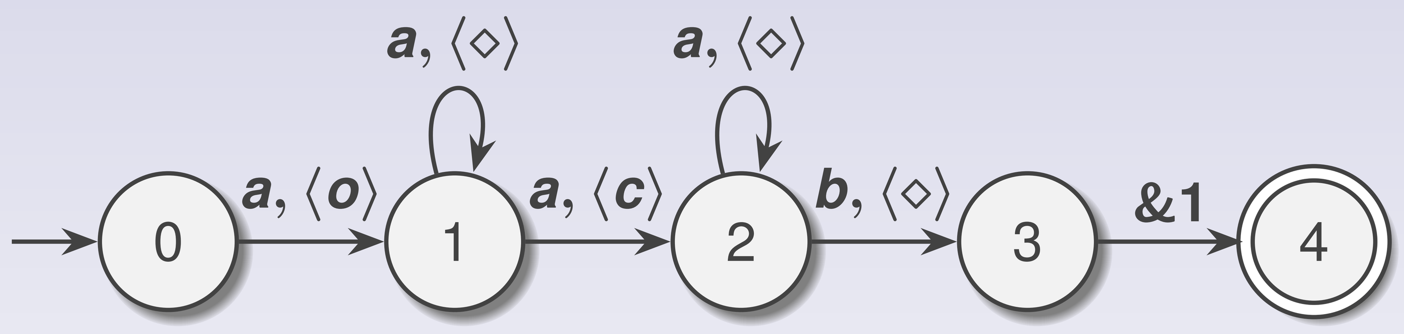
Daria Ismagilova *Antonina Nepeivoda*
Bauman Moscow State Technical University

Memory Finite Automaton

Backref-regex (ref-words, by Shmid) operations:

$\begin{cases} [_k\tau]_k & \text{(named capturing)} \\ \&k & \text{(reading memory cell)} \end{cases}$

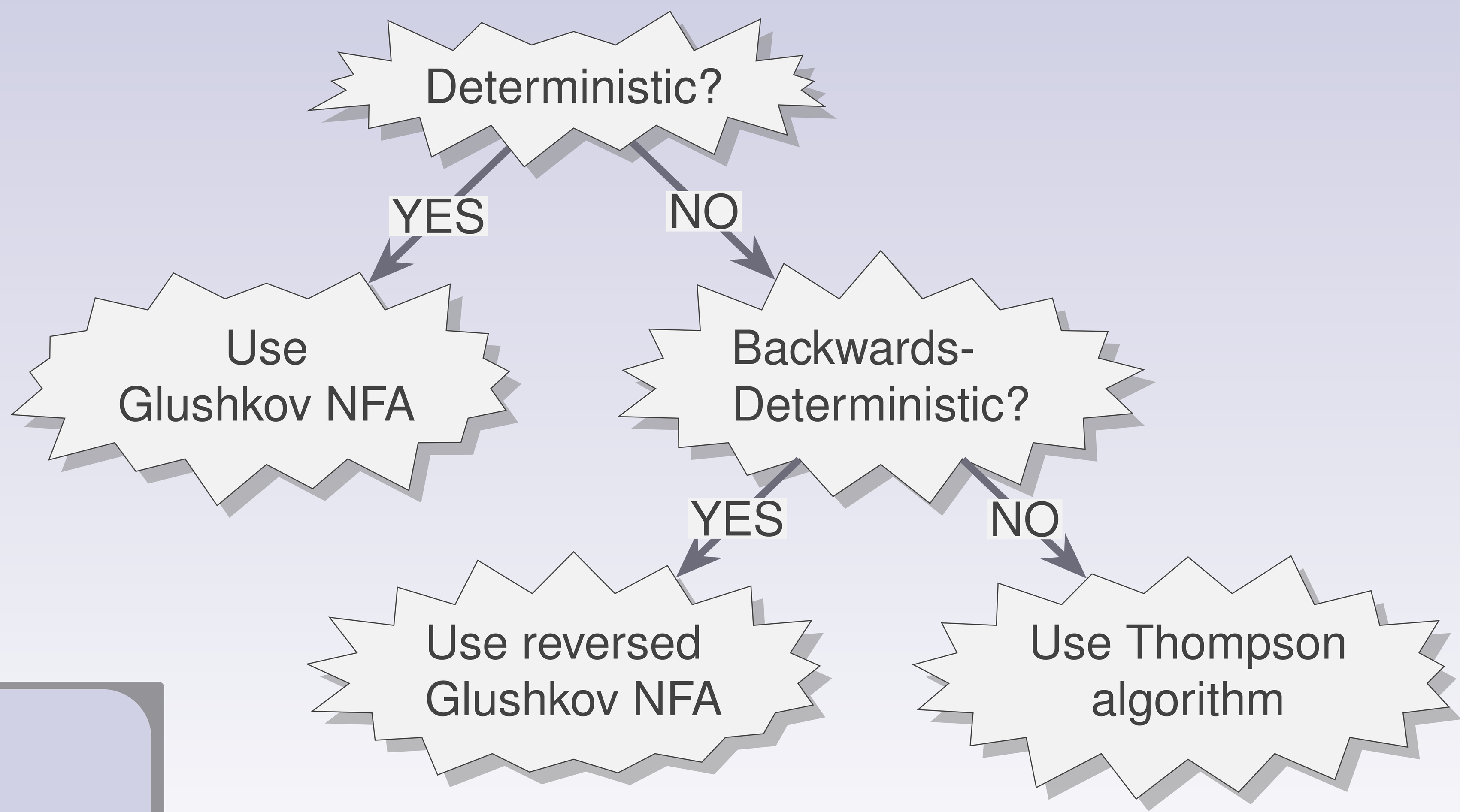
Example: $[_1a^*]_1a^+b\&1$ defines $\{a^mba^n \mid m > n\}$



Memory Finite Automaton by Shmid (extended Glushkov automaton)

Memory operations: $\begin{cases} o & \text{(open)} \\ \diamond & \text{(retain)} \\ c & \text{(close)} \end{cases}$

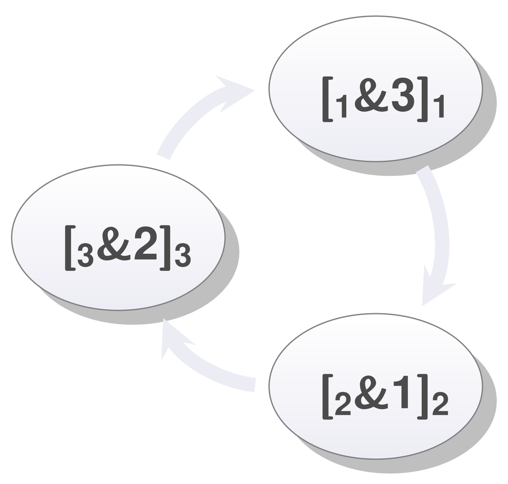
RE2 Matching Strategy



Problem

Improve matching for ref-words

Cyclic Memories



- Destroy closure properties
- Hard to restrict syntactically

Rewriting Rules

- Semiring properties
  - $x(yz) = (xy)z$
  - $x(y + z) = xy + xz$
- Conway–Crob transformations
  - $x^*(yx^*)^* = (x + y)^*$
  - $x(yx)^* = (xy)^*x$

Suggested Approach

ACREG class

- Preserving semiring & Conway–Crob rules
- Closed under sound memory renaming

Sound renaming

$[_1a^*]_1b\&1[_2b^*]_2\&2 \equiv [_1a^*]_1b\&1[_1b^*]_1\&1$

Unsound renaming

$[_1a^*]_1b[_2b^*]_2\&1\&2 \equiv [_1a^*]_1b[_1b^*]_1\&1\&1$

Main Idea

Memory can be disambiguated, leading to Backref-Normal Form

Theorem

1. **ACREG languages are closed under reversal;**
2. **Shmid ref-word languages are not closed under reversal.**

Experimental Matching Algorithm