DIPLOMA THESIS

Peter Černo

Clearing Restarting Automata

Supervised by RNDr. František Mráz, CSc.

Clearing Restarting Automata

- Represent a new restricted model of restarting automata.
- Can be learned very efficiently from positive examples and the extended model enables to learn effectively a large class of languages.
- In the thesis we relate the class of languages recognized by these automata to Chomsky hierarchy and study their formal properties.

Diploma Thesis Outline

- **Chapter 1** gives a short introduction to the theory of automata and formal languages.
- **Chapter 2** gives an overview of several selected models related to our model.
- **Chapter 3** introduces our model of clearing restarting automata.
- **Chapter 4** describes two extended models of clearing restarting automata.

Selected Models

- **Contextual Grammars** by Solomon Marcus:
  - Are based on adjoining (inserting) pairs of strings/contexts into a word according to a selection procedure.
- **Pure grammars** by Mauer et al.:
  - Are similar to Chomsky grammars, but they do not use auxiliary symbols – nonterminals.
- **Church-Rosser string rewriting systems**:
  - Recognize set of words which can be reduced to an auxiliary symbol Y. Each maximal sequence of reductions ends with the same irreducible string.
- **Associative language description** by Cherubini et al.:
  - Work on so-called stencil trees which are similar to derivation trees but without nonterminals. The inner nodes are marked by an auxiliary symbol Δ.

Selected Models

- **Restarting automata**:
  - Introduced by Jančar et al. in 1995 in order to model the so-called `analysis by reduction’.
  - Analysis by reduction is a technique used in linguistics to analyze sentences of natural languages that have free word order.

Formal Definition

- Let k be a *positive integer*.
- k-clearing restarting automaton (k-cl-RA-automaton for short) is a couple $M = (\Sigma, I)$:
  - Σ is a finite nonempty *alphabet*, $¢, \$ \notin \Sigma$.
  - I is a finite set of *instructions* (x, z, y), $z \in \Sigma^+$,
    - $x \in LC_k = \Sigma^k \cup ¢.\Sigma^{\leq k-1}$ (left context)
    - $y \in RC_k = \Sigma^k \cup \Sigma^{\leq k-1}.\$$ (right context)
  - The special symbols: $¢$ and $\$$ are called *sentinels*.

Formal Definition

- A word $w = uzv$ can be *rewritten* to uv (denoted $u\underline{z}v \vdash_M uv$) if and only if there exist an instruction $i = (x, z, y) \in I$ such that:
  - x is a *suffix* of $¢.u$
  - y is a *prefix* of $v.\$$
- A word w is *accepted* if and only if $w \vdash^*_M \lambda$ where $\vdash^*_M$ is reflexive and transitive closure of $\vdash_M$.
- The $k\text{-}cl\text{-}RA$-automaton M *recognizes* the language $L(M) = \{w \in \Sigma^* \mid M \text{ accepts } w\}$.

Example

- Language $L = \{a^n b^n \mid n \geq 0\}$.
- Can be recognized by the *1-cl-RA*-automaton $M = (\{a, b\}, I)$, where the instructions I are:
  - $R1 = (a, \underline{ab}, b)$
  - $R2 = (¢, \underline{ab}, \$)$
- For instance:
  - $aaa\underline{ab}bbb \vdash^{R1} aa\underline{ab}bb \vdash^{R1} a\underline{ab}b \vdash^{R1} \underline{ab} \vdash^{R2} \lambda$.
- Now we see that the word *aaaabbbb* is accepted.

Regular Languages

- **Theorem**: All regular languages can be recognized by clearing restarting automata using only instructions with left contexts starting with $¢$.
- **Theorem**: If $M = (\Sigma, I)$ is a k-cl-RA-automaton such that for each $(x, z, y) \in I$: $¢$ is a prefix of x **or** $\$$ is a suffix of y then $L(M)$ is a regular language.

Non-Context-Free Languages

- **Theorem**: The family of languages recognized by *1-cl-RA*-automata is strictly included in the family of context-free languages.
- **Theorem**: *2-cl-RA*-automata can recognize some **non**-context-free languages.

Problem with *cl-RA*-automata

- **Theorem**: The language:

$$L_1 = \{a^n c b^n \mid n \geq 0\} \cup \{\lambda\}$$

is not recognized by any *cl-RA*-automaton.

Extended Models

- *Δ*- clearing restarting automata
  - Can leave a mark – a symbol *Δ* – at the place of deleting besides rewriting into the empty word.
  - Can recognize Greibach's hardest-context-free language.
- *Δ**- clearing restarting automata
  - Can rewrite a subword w into Δ^k where $k \leq |w|$.
  - Can recognize all context-free languages.

Conclusion

- The main goal of the thesis was successfully achieved.
- The results of the thesis were presented in:
  - **ABCD workshop**, Prague, March 2009
  - **NCMA workshop**, Wroclaw, August 2009
  - An extended version of the paper from the NCMA workshop was accepted for publication in **Fundamenta Informaticae**.
- Many interesting theoretical questions and problems are still open or under investigation.

Thank You

http://www.petercerno.wz.cz/ra.html