From Regular to Strictly Locally Testable Languages

Stefano Crespi Reghizzi **Pierluigi San Pietro**[1]

[1]DEI-Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy

WORDS 2011, Prague

Regular languages = hom. images of local languages

- A language L is *local* if $\exists$ three finite sets: $I, T \subseteq A$, $F \subseteq A \times A$, such that $x \in L \iff$

  the first (resp. last) symbol of x is in I (resp. in T)

  and the factors of length 2 of x are in F.

- Local languages important as generators of language families: context-free, and more to the point, regular.
- Classical result (Y. Medvedev 1964, Eilenberg 1974): every regular language $R \subseteq A^*$ is the homomorphic image of a local language $L \subseteq B^*$. Alphabet B is called *local*.
- In the original construction, alphabet B is much larger: it is the set $E \subseteq Q \times A \times Q$ of labelled edges of a NFA (Q, A, E, q_0, F) accepting language R.

Problems we want to study

- Define the *alphabetic ratio* $|B|/|A|$, which in Medvedev and Eilenberg is $O(|Q|^2)$.
- How small can the ratio be?
- Local languages are a member of McNaughton and Papert's infinite hierarchy of *k-strictly locally testable* (k-slt), languages, where $k \geq 2$ is the *width*.
  What is the minimum alphabetic ratio such that, for some finite k, every regular language is the alphabetic homomorphism of a k-slt language?

An easy reduction of Medvedev's ratio

The local alphabet size can be reduced from quadratic to *linear* in the number of states. Let $M = (Q, A, E, q_0, F)$ be an NFA and $R = L(M)$.

Proposition

Language R is the hom. image of a local language L' on an alphabet B of size $|Q| \cdot |A|$.

Proof: the following sets define a local language $L' \subseteq (Q \times A)^+$.

$$I_1 = \{\langle q_0, a\rangle \mid a \in A\};$$
$$F_2 = \{\langle q, a\rangle\langle q', b\rangle \mid a, b \in A, q, q' \in Q, (q, a, q') \in E\};$$
$$T_1 = \{\langle q, a\rangle \mid a \in A, \exists q' \in F : (q, a, q') \in E\}.$$

Can we do better? We study a more general problem, using as generators k-slt instead of local languages.

Strictly Locally Testable Languages

For a word $w \in A^k \cdot A^*$, $k \geq 2$, $i_k(w)$ and $t_k(w)$ are the prefix and, resp., the suffix of w of length k, and $f_k(w)$ the set of factors of w of length k.

Definition

A language L is *k-strictly locally testable*, (k-slt) $\iff$ exist finite sets $I_{k-1}, T_{k-1} \subseteq A^{k-1}$ and $F_k \subseteq A^k$ such that, for every $x \in A^k \cdot A^*$:

$$x \in L \iff i_{k-1}(x) \in I_{k-1} \wedge t_{k-1}(x) \in T_{k-1} \wedge f_k(x) \subseteq F_k$$

A language is *slt* if it is k-slt for some k (called the *width*).

For $k = 2$ we obtain local languages.

(h,k)-*homomorphic* languages, a new concept

Definition

A language $R \subseteq A^+$ is $(\overbrace{h}^{\geq 1}, \overbrace{k}^{\geq 2})$-*homomorphic* if there exist an alphabet B of size h, a k-slt language $L \subseteq B^+$, and a homomorphism $\pi : B \to A$ such that $R = \pi(L)$.

- If R is k-slt then it is trivially $(|A|, k)$-homomorphic
- Otherwise, a local alphabet larger than A may be needed
- Medvedev (improved) result restated: every language accepted by an NFA with n states is $(n \cdot |A|, 2)$-homomorphic.

Example: trade-off of alph. ratio vs. width

$$R = (aaa)^+$$

$$(3,2) - hom. \quad R = \pi(L') \qquad L' = (a_1 a_2 a_3)^+$$

$$(2,3) - hom. \quad R = \pi(L'') \qquad L'' = (a_1 a_1 a_2)^+$$

$$\pi(a_1) = \pi(a_2) = \pi(a_3) = a$$

E.g., L'' is defined by:

$$I_2 = \{a_1 a_1\}$$

$$T_2 = \{a_1 a_2\}$$

$$F_3 = \text{circ. permutations of } a_1 a_1 a_2$$

A simple yet perhaps surprising result

A natural question

By allowing the width k to be larger than 2, one can often reduce the alph. ratio to less than $n = |Q|$: are there any lower bounds on the alph. ratio?

In general the local alphabet cannot be smaller than twice the size of the original alphabet:

Theorem

For every alphabet A, there exists a regular language $R \subseteq A^+$ that is not $(2 \cdot |A| - 1, k)$-homomorphic, for every $k \geq 2$.

Proof: $L = \bigcup_{a \in A}(aa)^*$ is not $(2 \cdot |A| - 1, k)$-homomorphic

By contradiction, R is $(2 \cdot |A| - 1, k)$-homomorphic: $\exists$ local alphabet B of size $2 \cdot |A| - 1$, a k-slt language $L \subseteq B^+$ and hom. $\pi : B \to A$ such that $R = \pi(L)$.
Since $|B| = 2 \cdot |A| - 1$, there exists a symbol, say, $a \in A$ having exactly one pre-image $b \in B$, i.e., $\pi^{-1}(a) = \{b\}$.
Word $a^{2k} \in R$ implies $\exists x \in L$ such that $\pi(x) = a^{2k}$, and $x = b^{2k}$.
Consider $xb = b^{2k+1}$. Clearly, $\pi(xb) = a^{2k+1} \notin R$, hence $xb \notin L$.
But x and xb have the same factors, prefix and suffix: a contradiction to the Def. of k-slt.

Main result

relates the language complexity in terms of number of states, the alphabetic ratio, and the width of the slt language.

Theorem

Every $R \subseteq A^*$ accepted by a NFA with $n > 1$ states is $(2|A|, O(\lg n))$-homomorphic.

Theorem is generalized at the end also allowing a larger alphabet in order to decrease width.

Idea of the proof: binary encoding of states

We want to encode the states of the original automaton into words of fixed length of the local alphabet.
Given $m \geq \lceil \lg_2 n \rceil$, $\forall q \in Q$ let $[q]$ be an m-bit encoding of q.
Local alphabet $B = A \times \{0, 1\}$.
Let $\pi_{0,1} : A \times \{0, 1\}$ such that $\forall a \in A, i \in \{0, 1\}$, $\pi_{0,1}(\langle a, i \rangle) = i$.
If $w \in B^m$, $\pi_{0,1}(w)$ may be the encoding $[q]$ of a state q.

Idea of the proof: encoding paths

For simplicity, consider words of length multiple of m:

$$x = x_1 x_2 \dots x_j, \quad |x_i| = m, j \geq 1$$

Assume the transition relation of the NFA accepting R is total. Then, $\exists$ a path in the automaton of the form:

$q_0 \xrightarrow{x_1} q_1 \xrightarrow{x_2} q_2 \cdots \xrightarrow{x_j} q_j$, with q_j final iff $x \in R$.

Define $w = w_1 \dots w_j$ such that for every i, $1 \leq i \leq m$:

- $\pi(w_i) = x_i$;
- $\pi_{0,1}(w_i) = [q_i]$;

We want to define a $2m$-slt lang. L with $\pi(L) = R$ s.t. $w \in L$ has the above property of "encoding a path".

Encoding of a path

Valid factor

A factor w_1w_2 is *valid* if there are $q_1, q_2 \in Q$ such that

- $[q_1] = \pi_{0,1}(w_1)$, $[q_2] = \pi_{0,1}(w_2)$, and
- $q_1 \xrightarrow{\pi(w_2)} q_2$

Hence, $\pi_{0,1}(w_1w_2) = [q_1][q_2]$.
A path for the original automaton can be decomposed in valid factors at distance m.
Idea is to define a $2m$-slt language allowing only valid factors and their shifts.

Not all encodings are good

Example

For $Q = \{q_0, q_1, q_2\}$ the binary encoding $[q_0] = 01, [q_1] = 10, [q_2] = 11$ is not adequate: factor 0110 can be interpreted as either:

$$[q_0][q_1]$$

$$0[q_2]1$$

The traditional notion of decodability (for every $x, y \in Q^+$, if $[x] = [y]$ then $x = y$) is not adequate: it assumes that the word to be decoded is a string in $[q_0][Q^*]$, while we need to consider *any factor of length 2m* of $[Q^+]$.

Idea of the proof: Factor decodability

Definition

A word $x \in \{0,1\}^{2m-1}$ is *factor-decodable* if there exists one, and only one, position j, $1 \leq j \leq m-1$, such that for some $q \in Q$: $s_{j,j+m}(x) = [q]$.

A code $[\ ] : Q \to \{0,1\}^m$ is *factor-decodable* if every word in $f_{2m-1}([Q^+])$ is factor-decodable.

An implementation

Let code $[\ ]$ be such that for every $q \in Q$, $[q]$ ends with 00, i.e., $s_{m-1,m}([q]) = 00$ and there is no other occurrence of 00 in $[q]$.

Main Lemma

The number of binary strings of length $p > 1$ without an occurrence of 00 is well-known to be $F(p+2)$, where $F(p)$ is the p-th Fibonacci number. It then follows:

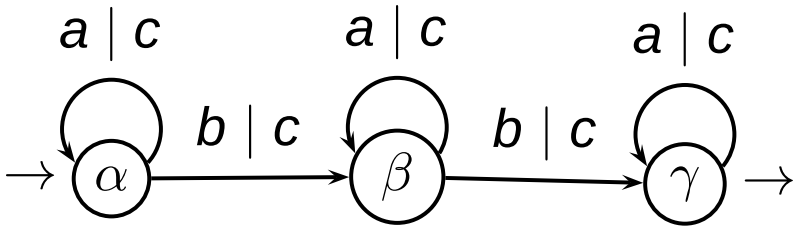
Lemma

Let $\phi = \frac{1+\sqrt{5}}{2}$. For all finite alphabets Q of size $n = |Q| \geq 2$, there exists a factor-decodable binary code of length $m = \lceil a + b \lg_2 n \rceil \geq 4$, with:

$$a = 1 + \frac{\lg_2 \sqrt{5}}{\lg_2 \phi} \approx 2.67$$

$$b = \frac{1}{\lg_2 \phi} \approx 1.44.$$

Example / 1 Medvedev's classical 2-slt language



$|Q| = 3$

$$\textit{Local alph.} = B = \{\langle \alpha, a, \alpha\rangle, \langle \alpha, c, \alpha\rangle, \langle \alpha, b, \beta\rangle, \ldots, \langle \gamma, a, \gamma\rangle\}$$

$$I_1 = \{\langle \alpha, a, \alpha\rangle, \langle \alpha, c, \alpha\rangle, \langle \alpha, b, \beta\rangle, \langle \alpha, c, \beta\rangle\}$$

$$T_1 = \{\langle \beta, b, \gamma\rangle \langle \beta, c, \gamma\rangle, \langle \gamma, a, \gamma\rangle, \langle \gamma, c, \gamma\rangle\}$$

Projection:

$$\begin{array}{lcccc} x' = & \langle \alpha, a, \alpha\rangle & \langle \alpha, b, \beta\rangle & \langle \beta, c, \beta\rangle & \cdots \\ x = & a & b & c & \cdots \end{array}$$

Size of local alphabet $= 10$. Alph. ratio = 10/3

Example *I* 2 $(4, 8)$-homom. language

alph. ratio $= 2$, binary state encoding:

α	β	γ
01	10	11

separator field 00, bit count per state $= 4$
width of local lang. $= 2 \times 4 = 8$

Projections:

$x' =$	a_0	a_1	b_0	a_0		a_1	a_0	a_0	a_0		a_1	c_0
$x =$	a	a	b	a		a	a	a	a		a	c

Each factor of length 7 contains exactly one code:

$$a_0 a_1 \underbrace{b_0 a_0 a_1 a_0 a_0 a_0 a_1}_{\text{code } 10 \to \beta} b_0$$

Sets I_7, T_7 are straightforward, F_8 includes all and only valid factors and their shifts.

Generalization to non-binary encodings 1/2

Encode states with alphabet D, $|D| \geq 2$, to decrease width.

Lemma

For all alphabets Q, D with $n = |Q| \geq 2$ and $h = |D|$, $2 \leq h < n$, $\exists$ a code of Q into D of length $m = \lceil g(h) + f(h) \lg_2 n \rceil$:

$$f(h) = \lg_2^{-1}\left(h - 1 + \sqrt{(h-1)(h+3)}\right) - 1 \lessapprox 1.44$$

$$g(h) = 1 + \frac{f(h)}{2}\left(\lg_2(h-1) + \lg_2(h+3)\right) \lessapprox 2.67.$$

Result is asymptotically optimal.

Generalization to non-binary encodings 2/2

Theorem

A language $R \subseteq A^*$ accepted by a NFA with $n > 1$ states, is $\left(h|A|, O(\frac{\lg n}{\lg h})\right)$-homomorphic for every $h \geq 2$.

Open Problems

Open questions and related problems:

- Question 1: with given alph. ratio, say, 2, what is the *minimal* slt *width* that suffices for any regular lang.?
- Question 2: do sub-families of regular langues (e.g., *aperiodic*) languages admit lower alph. ratios and slt widths?
- Application to *consensual languages*, a recent [S.C.R & P.S.P., RAIRO-Th. Inf. Appl. 2011] computational model based on concurrent operations of a DFA.
- 2-dim. or *picture lang.* homomorphically defined by *tiling systems* [Giammarresi and Restivo]: does our result hold?