Some more regular languages that are Church Rosser congruential

Klaus Reinhardt
Institut für Informatik, University of Tübingen
e-mail: reinhard@informatik.uni-tuebingen.de

and

Denis Thérien
School of Computer Science, McGill University
e-mail: denis@cs.mcgill.ca

ABSTRACT

We show that those languages, where the syntactic monoid is a finite group, are Church Rosser Congruential Languages (CRCL). We then extend this result to the class of regular languages whose syntactic monoid lies in the variety DO.

1. Introduction

In [1] McNaughton et al considered finite, length-reducing and confluent string-rewriting systems, which allow to find a unique irreducible shortest representing word for any given word: the Church-Rosser congruential languages (CRCL) are defined as the set of languages, which are the union of finitely many such equivalence classes. This is defined formally as follows:

A language $L \in A^*$ is in CRCL if there exists a rewriting system $R \subseteq \{l \mapsto r \mid l, r \in A^*, |l| > |r|\}$ allowing derivations $\alpha l \beta \underset{R}{\Longrightarrow} \alpha r \beta$ with the property that if $w \underset{R}{\overset{*}{\Longrightarrow}} w'$ and $w \underset{R}{\overset{*}{\Longrightarrow}} w''$ then there is a $v \in A^*$ with $w' \underset{R}{\overset{*}{\Longrightarrow}} v$ and $w'' \underset{R}{\overset{*}{\Longrightarrow}} v$ and L is the finite union of sets $[w_e]_R = \{w \mid w \underset{R}{\overset{*}{\Longrightarrow}} w_e\}$ for finitely many w_e.

It is an open problem weather all regular languages are in CRCL.

It was shown in [2] that some shuffle languages as well as Level 1 of the Straubing-Thérien hierarchy are in CRCL. Furthermore [2] describes a solution for the language $(A^2)^*$ (words of even length on alphabet $A = \{a, b\}$), which is the smallest nontrivial example for our result since the syntactic monoid for $(A^2)^*$ is the cyclic group $\mathbb{Z}_2$.

Our approach is as follows: We refine the *syntactic monoid* $M(L) := \Sigma^*/\equiv_L$ for the congruence relation $\equiv_L$ defined by $w \equiv_L v$ iff $\forall u, x \in \Sigma^* : uwx \in L \leftrightarrow uvx \in L$. The refinement is defined by the construction of a confluent length-reducing string-rewriting system $R \subset\equiv_L$ with the property that words exceeding some length contain an infix l with $l \mapsto r \in R$, which leads to a finite $\equiv_R\supseteq\equiv_L$ having the property that every congruence class has a unique shortest representing word.

For the example $(A^2)^*$ the string-rewriting system $R = \{aa \mapsto \lambda, bab \mapsto b, bbb \mapsto b\}$ leads to the syntactic monoid consisting of the 10 elements $[\lambda], [ab], [ba], [bb], [abba], [a], [b], [aba], [abb], [bba]$, where two words are equivalent, if their length, the number of a's before the first b and the number of a's after the last b differ by an even number and both contain or do not contain a b. The first 5 refine $[\lambda]$ and the last 5 refine $[a] = [b]$ (two minimal representing strings since $|a| = |b|$, which was the reason which made the refinement necessary).

2. The construction for a group

Theorem *If the syntactic monoid $M(L)$ is a group, then $L \in CRCL$. Furthermore the equivalence relation $\equiv_R \supseteq \equiv_L$ is finite but for any length we can choose R such that all words up to that length are irreducible.*

Proof. For $L = \emptyset$ or $L = \{\lambda\}$ this is trivial, the following is an induction over the size of the alphabet of L. Let L be a language over the alphabet $\Sigma = \{b, a_0, ..., a_{s-1}\}$, $G = M(L)$ and n be a multiple of the order of all elements $e \in G$ (Note here that n can be chosen arbitrarily large). We use sequences of consecutive powers of words of the form $\gamma_i := ba_{i \,\mathrm{Mod}\, s}^{n+(i \,\mathrm{Div}\, s)}$ or in other words $\gamma_{i+sj} = ba_i^{n+j}$ to build up representing words for each element in G. For example $[a_0bba_1a_2]$ can be represented as

$$(ba_0^n)^{n-1}(ba_0^{n+1})(ba_0^{2n})(ba_1^{2n+1})(ba_2^{3n})^{n-1}(ba_2^{3n+1}) = \gamma_0^{n-1}\gamma_s\gamma_{sn}\gamma_{s(n+1)+1}\gamma_{s2n+2}^{n-1}\gamma_{s(2n+1)+2}.$$

Since G is finite, we can find an m such that for every $e \in G$ there is a representing word w_e with $[w_e] = e$ and $bw_e = \gamma_0^{i_0}\gamma_1^{i_1}...\gamma_m^{i_m}$. Furthermore since $|\gamma_0| + 1 = |\gamma_s|$, we can pump i_0 and i_s by some multiples of n for each e in a way such that $l \leq |w_e| < l + n$ for some l chosen big enough.

By induction over the alphabet size we have a confluent, strictly length-reducing rewriting system R_a, which reduces each word $w \in \{a_0, ..., a_{s-1}\}^*$ to an irreducible word $w_w \in \mathrm{Irr}_a$ with $[w_w] = [w]$. Let $l_a := max\{|w| \mid w \in \mathrm{Irr}_a\}$. Furthermore we may assume w.l.o.g. that $\{\gamma_0, ..., \gamma_m\} \subseteq b\mathrm{Irr}_a$ (Choosing the n in the induction large enough). Let

$$\{\alpha_1, ...\alpha_p\} := \{w \in (b\mathrm{Irr}_a)^+ \mid w \text{ primitive and } |w| \leq n \text{ or } w \in bIrr_a\}.$$

In order to deal with all long cyclic repetitions of these words, we use the rewriting-rules $R_c = \{\alpha_i^{l+n}b \mapsto \alpha_i^l b \mid i \leq p\}$. Relevant words not being a part of such a cycle are in $L_{nc} := \{w \in (bIrr_a)^+b \mid \neg\exists i \leq p, x, y \in \Sigma^* \; xwy \in \alpha_i^+\}$. As marker-words we use

$$\{\beta_1, ..., \beta_q\} := L_{nc} \setminus \Sigma^+ L_{nc} \setminus L_{nc}\Sigma^+ \setminus \{w\gamma_j b \mid j \leq p \; \neg\exists i > j \; w = \gamma_i\}.$$

Observe here that each of the β's has a length of at least $n + 1$. Furthermore we assume an ordering of the β's such that if $\beta_j = \gamma_k x$ and $\beta_i \neq \gamma_h y$ for any $h \leq k$ then $i > j$. Now we define the rewrite-rules

$$R_n' := \{\beta_i w \beta_i \mapsto \beta_i w_{[w]} \beta_i \mid |w_{[w]}| < |w| \leq l_{i-1}, \beta_i w \beta_i \notin \Sigma^* \beta_j \Sigma^* \text{ for any } j > i, \}$$

where the length restriction of l_{i-1} for w is constructed in the proof of Claim 2 showing that a longer w would not be reducible. To make it easier to prove the confluence, we eliminate rules, where the left side contains the left side of another rule, which does not change the reducibility of a word:

$$R_n := \{(l \mapsto r) \in R_n' \mid \neg\exists(l' \mapsto r') \in R_n' \; l' \text{ is Infix of } l\}.$$

It is clear that the rewriting system $R = R_a \cup R_c \cup R_n$ is strictly length-reducing and it's congruence refines G since left and right side of a rule are always congruent in G. It remains to show the following two claims:

□

Example: Let $G = S_3$, the group of permutations of $\{1, 2, 3\}$ be generated by the mirrorings $[a] = (12)$ and $[b] = (23)$ represented by the symbols a, b. (Choose $L \subseteq \{a, b\}^*$ to be any language accepted by this group.) The example is not trivial because of $[aba] = [bab]$. Choosing $n = 6$ according to the construction would make the example very big but the following 'handmade' construction of the representing words works already with $n = 2$: By recursion for the alphabet $\{a\}$ we obtain $R_a = \{a^5 \mapsto a^3\}$ leaving $\mathrm{Irr}_a = \{a^i \mid i < 5\}$ and $l_a = 4$. Using $\gamma_0 = baa, \gamma_1 = baaa, \gamma_2 = baaaa$, we can find the representing words

$w_{()} = aa(baa)^2(baaaa)^2$, $w_{(12)} = aa(baa)^1(baaa)^1(baaaa)^2$, $w_{(23)} = aaaa(baaaa)^3$,
$w_{(123)} = aa(baa)^4(baaa)^1$, $w_{(132)} = aaa(baaaa)^3$, $w_{(23)} = aa(baa)^3(baaa)^2$,
all having length l =18 or 19. We get $\alpha_i = ba^{i-1}$ for $i \leq p = 5$ and obtain the rules $R_c = \{(ba^i)^{20} \mapsto (ba^i)^{18} \mid i \leq 5\}$. We get $\{\beta_1, ..., \beta_{11}\} =$

$$\{\gamma_1\gamma_0 b, \gamma_2\gamma_0 b, \gamma_2\gamma_1 b, \gamma_0 bb, \gamma_0 bab, \gamma_1 bb, \gamma_1 bab, \gamma_2 bb, \gamma_2 bab, babb, bbab\}$$

and obtain the rules $R_n := \{\beta_i w \beta_i \mapsto \beta_i w_{[w]} \beta_i \mid |w_{[w]}| < |w| \leq 338210, \beta_i w \beta_i \notin \Sigma^* \beta_j \Sigma^*$ for any $j > i, \}$. The length of an irreducible word can be at most $l_{11} = 676450$ which bounds the number of equivalence classes to $< 2^{676450}$.

Claim 1: R is confluent.

Proof. We have to show that for any pair of rules $l_1 \mapsto r_1, l_2 \mapsto r_2 \in R$ that if $ul_1v = xl_2y$ then there exists a $w \in \Sigma^*$ with $ur_1v \overset{*}{\underset{R}{\Longrightarrow}} w$ and $xr_2y \overset{*}{\underset{R}{\Longrightarrow}} w$.

- If both rules are in R_a this holds by induction.
- If $l_1 \mapsto r_1 \in R_a$ and $l_2 \mapsto r_2 \in R_c \cup R_n$ this holds since $l_1 \in \{a_0, ..., a_{s-1}\}^*$ and $l_2 \in (bIrr_a)^+ b$ can not overlap.
- If both rules are in R_c this holds since either l_1 and l_2 overlap only in a small part z, which is not changed by the rules, that means w.l.o.g. $l_1 = \alpha_{i_1}^{l+n} b = \alpha_{i_1}^{l+n-1} z' z$ and $l_2 = \alpha_{i_2}^{l+n} b = zz'' \alpha_{i_2}^{l+n-1}$ and thus $w = u\alpha_{i_1}^{l-1} z' z z'' \alpha_{i_2}^{l-1} y$ or otherwise α_{i_1} is a rotation of α_{i_2} and thus $w = ur_1v = xr_2y$.
- If $l_1 \mapsto r_1 \in R_c$ and $l_2 \mapsto r_2 \in R_n$, again there can be only a short overlapping of the left sides inside a β_i, which is not changed since the appearance of α_k^{n+l} is not allowed in l_2 and β_i cannot appear in α_k^* by definition.
- If both rules are in R_n then either both rules have the same β_i in which case $w = u\beta_i w_{[z]} \beta_i y$ for $ul_1v = xl_2y = u\beta_i z \beta_i y$ since positions of repeated occurances of β_i as infix in a word must differ at least n, which garanties that w is shorter than ur_1v and xr_2y or otherwise with different β_{i_1}, β_{i_2} with $i_1 < i_2$ there is again a short unchanged overlap since in the case that $\beta_{i_2} w' \beta_{i_2} \in \Sigma^* \beta_{i_1} \Sigma^*$, we would have $\beta_{i_2} w' \beta_{i_2} \in \Sigma^* \beta_{i_1} \Sigma^* \beta_{i_1} \Sigma^*$, since no β_{i_2} can occur between two β_{i_1}'s but then l_1 would be an infix of l_2, which is not possible according to the definition of R_n.

□

Claim 2: The number of congruence classes defined by R is finite.

Proof. We show by induction on h that any corresponding to R irreducible word w, which does not contain any β_j with $j > h$ as infix, can only have some bounded length.

The case $h = 0$: Assume there would be an arbitrarily long irreducible word. After reading a prefix in Irr_a, we see the first b, which is the beginning of an infix $\alpha_{i'}$. If we continue reading, we can find $k < l + n$ consecutive repetitions of $\alpha_{i'}$, before we find a different $\alpha_{j'}$, which is not a prefix of a word in $\alpha_{i'}\Sigma^*$. Now this infix $\alpha_{i'}^k \alpha_{j'}$ must have an infix in L_{nc} and thus a minimal one in $L_{nc} \setminus \Sigma^+ L_{nc} \setminus L_{nc}\Sigma^+$. Therfore the only possibility not to get a β as infix is that $\alpha_{j'} = \gamma_j$ for a $j \leq m$ and $\neg \exists i > j\ \alpha_{i'} = \gamma_i$, which can occur at most m times continuing to read the word. This restricts the length of the word to $\leq l_0 = (m + 1)(l + n)(l_a + 1)$.

Step from $h - 1$ to h: By induction, the length of the prefix before the first occurance of of β_h and the postfix after the last occurance of β_h is bounded by l_h. The same holds for the distance between two occurances of β_h, but then the reduction-rules for β_h ensure that it is even at most $l + n$ and thus this even holds for the distance between the first and the last occurance of β_h, which means we can bound the length of the word by $l_h := 2l_{h-1} + l + n + 2(l_a + 1)$.

Thus no irreducible word can be longer than l_q. □

3. Extension to DO

The variety DO consists of the closure of group and letter-testing languages under unambiguous concatenation.

Theorem *If the syntactic monoid $M(L)$ of a regular language L is in the variety DO, then $L \in CRCL$.*

References

[1] R. McNaughton, P. Narendran, F.Otto. Church-Rosser Thue systems and formal languages. Journal of the ACM, 35:324-344, 1988.

[2] G. Niemann, J. Waldmann. Some regular languages that are Church-Rosser congruential. DLT 2001.