Star-free and first-order definable languages

Angelo Montanari

Department of Mathematics, Computer Science, and Physics
Università degli Studi di Udine

The main issue

We know that *restricted regular expressions* (and general ones) are as expressive as *regular languages*, and, similarly, ω-*regular expressions* are as expressive as ω-*regular languages*.

What about **star-free regular expressions**?

We are going to prove that they are as expressive as the **first-order fragment of S1S**, where quantification is restricted to first-order variables (positions in sequences).

The infinite case

An **analogous result** holds for *star-free* and *first-order ω-languages* (see, for instance, W. Thomas, Automata on Infinite Objects, Handbook of Theoretical Computer Science, Vol B, Chapter 4, J. van Leeuwen (ed.), Elsevier, 1990).

As we will see, they are also as expressive as the propositional temporal logic of linear time.

For the sake of simplicity, we restrict ourselves to the finite case.

The main theorem (McNaughton and Papert)

Theorem

Let A be a finite alphabet. A language $W \subseteq A^*$ is star-free if and only if it is first-order definable in $S1S_A$ (with the ordering relation $<$ and unary predicates Q_a, for all $a \in A$).

Remark. While $+1$ is first-order definable in terms of $<$, the opposite is not, that is, $<$ is second-order, but not first-order, definable in terms of $+1$.

Hence, $S1S_A$ with the ordering relation $<$ and unary predicates Q_a, for all $a \in A$, is strictly more expressive than $S1S_A$ with the successor function $+1$ and unary predicates Q_a, for all $a \in A$.

Left-to-right direction

The left-to-right direction is straightforward.

It exploits the close correspondence between the operations $\cup$, $\sim$, and $\cdot$ and the connectives/quantifiers $\vee$, $\neg$, and $\exists$, respectively.

Example. The star-free expression

$$A^* \cdot (a \cup c) \cdot \sim (A^* \cdot b \cdot A^*)$$

over the alphabet $A = \{a, b, c\}$ is defined by the first-order sentence:

$$\exists x((x \in Q_a \vee x \in Q_c) \wedge \neg \exists y(x < y \wedge y \in Q_b))$$

Left-to-right direction (cont'd)

The **proof** is by **induction** on the structure of the star-free expression.

Example.
If U (resp., V) is defined by the first-order formula ϕ (resp., ψ), then

$$U \cdot V$$

is defined by the formula

$$\exists x(\phi'(x) \wedge \psi'(x))$$

where ϕ' (resp. ψ') is the relativitazion of ϕ (resp., ψ) to positions $\leq x$ (resp., $> x$), assuming, for simplicity, that $\epsilon \notin U$.

The detailed proof is left as an **exercise**.

Right-to-left direction

The **proof** is by **induction** on the quantifier depth of formulas (and definitely less straightforward).

We focus on the most interesting and difficult case of the inductive step: the **existential quantifier**.

We consider the formula $\exists x \phi(x)$ of quantifier depth $n + 1$ assuming (inductive hypothesis) that sentences of quantifier depth n define star-free languages. [The general case of formulas $\exists x \phi(x, \vec{y})$, with free variables $\vec{y}$, is technically more involved.]

The proof idea

To rewrite the formula $\exists x \phi(x)$ into formulas of the form

$$\exists x(\phi_{<x} \wedge x \in Q_a \wedge \phi_{>x})$$

where $\phi_{<x}$ (resp., $\phi_{>x}$) refers to positions $< x$ (resp., $> x$) only.

If $\phi_{<x}$ and $\phi_{>x}$ are of quantifier depth n, then such a formula describes a language

$$U \cdot a \cdot V$$

where U and V are star-free by inductive hypoyhesis.

n-equivalence and (n,1)-equivalence relations

We introduce two fundamental equivalence relations: $\equiv_n$ (n-equivalence) and $\equiv_{(n,1)}$ ((n,1)-equivalence).

The **equivalence relation** $\equiv_n$ over A^* is defined as follows:
for all $u, v \in A^*$,

$u \equiv_n v$ iff u and v satisfy the same sentences of quantifier depth n

To deal with formulas with free variables (one in our case), we consider words with one distinguished position, that is, pairs (u, r), with $u \in A^*$ and $1 \leq r \leq |u|$.

The **equivalence relation** $\equiv_{(n,1)}$ over $A^* \times \mathbb{N}$ is defined as follows:
for all $u, v \in A^*$ and $r, s \in \mathbb{N}$,

$(u, r) \equiv_{(n,1)} (v, s)$ iff (u, r) and (v, s) satisfy
the same formulas $\varphi(x)$ of quantifier depth n

Basic results

Fact 1. For any $n \geq 1$, both $\equiv_n$ and $\equiv_{(n,1)}$ are of finite index, that is, there exist finitely many equivalence classes of $\equiv_n$ and $\equiv_{(n,1)}$.

Fact 2. For any $n \geq 1$, any class W of words u of $\equiv_n$ (resp., any class $\underline{W}$ of word models (u, r) of $\equiv_{(n,1)}$) can be defined by a sentence φ_W (resp., a formula $\varphi_{\underline{W}}(x)$) of quantifier depth n.

Both facts can be proved by induction on the quantifier depth n.

The next proposition easily follows.

Proposition 1. *Any formula $\varphi(x)$ of quantifier depth n is equivalent to a finite disjunction of formulas $\varphi_{\underline{W}}(x)$ (those formulas $\varphi_{\underline{W}}(x)$ such that $\underline{W}$ contains some (u, r) satisfying $\varphi(x)$).*

n- and (n,1)-equivalence relations are congruences

Proposition 2. *The equivalence relations* $\equiv_n$ *and* $\equiv_{(n,1)}$ *are congruences, that is:*

$$\textit{if } u \equiv_n v \textit{ and } u' \equiv_n v', \textit{ then } uu' \equiv_n vv'$$

$$\textit{if } u \equiv_n v, a \in A, \textit{ and } u' \equiv_n v', \textit{ then } (uau', |u|+1) \equiv_{(n,1)} (vav', |v|+1)$$

Proposition 2 can be easily proved by exploiting the (Ehrenfeucht-Fraissé) game-theoretic characterization of $\equiv_n$ and $\equiv_{(n,1)}$.

Remark 1. Unlike Proposition 1, Proposition 2 **depends on** the signature of $S1S_A$ (binary ordering relation and unary predicates only).

Remark 2. The validity of Proposition 2 is not restricted to finite words: it holds for **any linear order** expanded by unary predicates.

From the formula to the star-free expression

As a preliminary step, we observe that, in order to find the star-free counterpart of the formula $\exists x \phi(x)$ (of quantifier depth $n+1$), we can restrict our attention to the formula $\exists x \phi_{\underline{W}}(x)$ as

$$\exists x \phi(x) \leftrightarrow \exists x \bigvee_{\underline{W}} \phi_{\underline{W}}(x) \leftrightarrow \bigvee_{\underline{W}} \exists x \phi_{\underline{W}}(x)$$

Let us consider a triple (U, a, V), with $U, V \equiv_n$-classes and $a \in A$.

If there exist $u_0 \in U$ and $v_0 \in V$ such that $(u_0 a v_0, |u_0| + 1) \in \underline{W}$, then, by Proposition 2, it holds that $(uav, |u| + 1) \in \underline{W}$, for all $u \in U$ and $v \in V$.

This allows us to conclude that all words in $U \cdot a \cdot V$ satisfy $\exists x \phi_{\underline{W}}(x)$.

From the formula to the star-free expression (cont'd)

Hence, $\exists x \phi_{\underline{W}}(x)$ defines (can be mapped into) the (finite) union of the sets/languages $U \cdot a \cdot V$, selected from the set of all triples (U, a, V), such that U, V are $\equiv_n$-classes that contain elements u_0, v_0 as above.

Since, by the inductive hypothesis, U, V are star-free sets/languages, it immediately follows that $\exists x \phi_{\underline{W}}(x)$ defines a star-free set/language.

A characterization of star-free regular languages

The relationships between star-free expressions and formulas of the first-order fragment of $S1S_A$ is tighter than expressed in McNaughton-Papert theorem.

It is indeed possible to show that

the classification of star-free regular languages by **dot-depth**, that is, the number of alternations between concatenation and Boolean operations in the expressions that define them

coincide with

the classification of languages definable in the first-order fragment of $S1S_A$ in terms of **quantifier alternation depth**.

The infinite case (Ladner)

Theorem

Let A be a finite alphabet. An ω-language $L \subseteq A^\omega$ is first-order definable in $S1S_A$ if and only if L is obtained from A^ω by repeated application of Boolean operations and concatenation with star-free sets/languages $W \subseteq A^*$ on the left.

ω-languages satisfying the conditions of the above theorem are called **star-free ω-languages**.

Ladner showed that star-free ω-languages are a proper subclass of ω-languages.