Regular Languages Closed Under Kleene Plus

Yuval Filmus

January 2011

1 Introduction

Vincenzo Ciancia defined the following class of regular languages, which he called *circular languages*.

Definition 1.1. A regular language L is $+$-*closed* if whenever $w \in L$ then $w^+ \in L$.

In this note we lay some of the theory of regular $+$-closed languages.

2 Normal Form

At first glance, it might seem that a $+$-closed language is always of the form L^+, what we call a $+$-language.

Definition 2.1. A regular language L is a $+$-*language* if $L = M^+$ for some regular language M.

However, the language $a^+ + b^+$ is $+$-closed but not a $+$-language. Our goal in this section is to prove the following theorem, which shows that every $+$-closed regular language is a finite union of regular $+$-languages.

Theorem 2.2. *A regular language is $+$-closed if and only if it is the finite union of regular $+$-languages.*

Proof. Any union of $+$-languages is clearly $+$-closed. To prove the converse, let L be a regular $+$-closed language over some alphabet Σ given by some DFA with state-set S, accepting states A and starting state s, and let $q\colon S \times \Sigma^* \to S$ be the transition function. Denote the number of states by $n = |S|$.

For any word w in the language, define its trace $\tau(w)\colon \mathbb{N}_+ \to S$ by $\tau(w)(k) = q(s, w^k)$, i.e. $\tau(w)(k)$ is the state the DFA is after reading w^k. Since L is $+$-closed, $w \in L$ iff $\operatorname{ran} \tau(w) \subset A$. Furthermore, for any w the trace $\tau(w)$ is eventually periodic, so there are only finitely many traces.

Define $T = \{\tau(w) : w \in L\}$. Note that T is a finite set. For $\tau \in T$ define $L(\tau) = \{w : \tau(w) = \tau\}$. We claim that $L(\tau)$ is regular. Indeed, let m be the minimal position in τ which is repeated, i.e. $\tau(m) = \tau(r)$ for some $r < m$.

Thus $w \in L$ iff $\tau(w)(k) = \tau(k)$ for all $k \leq m$. In other words, $w \in L$ iff $q(\tau(k-1), w) = \tau(k)$ for $k \leq m$, where $\tau(0) = s$. We can check all these finitely many conditions in parallel using a single DFA.

Clearly $L = \bigcup_{\tau \in T} L(\tau)$. Since L is $+$-closed, moreover $L = \bigcup_{\tau \in T} L(\tau)^+$. Since T is finite, this is the required representation. $\square$

3 Inherent Ambiguity

In the previous section we have shown that every regular $+$-closed language is the finite union of regular $+$-languages. Can the union be disjoint? Consider, for example, the language

$$(a+b)^+ + (a+c)^+ + (b+c)^+.$$

Words which contain only one of a, b, c will belong to two summands. We call this phenomenon *ambiguity*.

Definition 3.1. A union of regular $+$-languages is *ambiguous* if the union is not disjoint. A $+$-closed regular language is *inherently ambiguous* if it cannot be written as a disjoint union of regular $+$-languages.

The language considered above has an unambiguous representation:

$$a^+ + b^+ + c^+ + (a^+b^+a^*)^+ + (b^+a^+b^*)^+ + (a^+c^+a^*)^+ + (c^+a^+c^*)^+ + (b^+c^+b^*)^+ + (c^+b^+c^*)^+.$$

However, other languages are inherently ambiguous.

Lemma 3.2. *A language L is a $+$-language if and only if whenever $a, b \in L$ then $ab \in L$, i.e. L is closed under concatenation.*

Proof. If L is closed under concatenation then it is certainly closed under taking positive powers. Conversely, let $L = M^+$. If $a, b \in L$ then $a = \alpha^i$ and $b = \beta^j$ for some $\alpha, \beta \in M$. Thus $ab = \alpha^i \beta^j \in M^{i+j} \subset M^+$. $\square$

Theorem 3.3. *The language L of words over $\{a, b\}$ containing either an even number of as or an even number of bs (or both) is an inherently ambiguous regular $+$-closed language.*

Proof. The language L is clearly regular. It is $+$-closed since if a word contains an even number of as then all its powers will also contain an even number of as.

Suppose that $L = \bigcup_i L_i^+$ is an unambiguous representation of the language. Choose an odd number n larger than the sizes of all DFAs for all L_i^+. Since $a^n b^{n!} \in L$, it must be generated by some L_i^+. Using the pumping lemma, we see that L_i^+ also generates $a^{n!} b^{n!}$. Similarly, $a^{n!} b^n$ is generated by some L_j^+ which also generates $a^{n!} b^{n!}$. Finally, if $i = j$ then $a^n b^{n! + n} b^{n!} \in L_i^+$, since L_i^+ is closed under concatenation. However, since $n + n!$ is odd this word doesn't belong to L. So $i \neq j$, and $a^{n!} b^{n!} \in L_i^+ \cap L_j^+$. $\square$

Open Question 1. *When is a $+$-closed language inherently ambiguous?*

4 Decidability

In this section we show how to decide whether a regular language is +-closed. On the negative side, we show that this problem is coNP-hard.

Theorem 4.1. *Given a DFA for a language L with n states, one can decide whether L is +-closed in time $n^{O(n)}$.*

Proof. The following uses some notations defined during the proof of Theorem 2.2.

Construct a new DFA which is the p'th power of the DFA for L, where $p = |A| + 1$. Denote the transition function of this new DFA by Q. For any word w with trace $\tau(w)$ we have

$$Q(s\tau(w)(1)\cdots\tau(w)(p-1), w) = \tau(w)(1)\cdots\tau(w)(p).$$

The language is not +-closed iff there is a word w whose trace $\tau(w)$ satisfies $\tau(w)(1) \in A$ but $\tau(w)(k) \notin A$ for some $k > 1$. By the pigeon-hole principle, the minimal such k satisfies $k \le p$. There are at most n^p such "illegal" traces, and for each such trace it is straightforward to check whether the state $\tau(w)(1)\cdots\tau(w)(p)$ is reachable from the state $s\tau(w)(1)\cdots\tau(w)(p-1)$. □

The proof shows that if a language with DFA size n is not +-closed, then there is a witness of size $n^{O(n)}$.

Open Question 2. *What is the best upper bound on the size of the smallest witness for a language not being +-closed?*

Given a DFA, it is coNP-hard to decide whether the corresponding language is +-closed.

Theorem 4.2. *It is coNP-hard to decide whether a regular language is +-closed, given its DFA.*

Proof. The reduction is from SAT. Let us be given a SAT instance. We can assume that the instance has p variables and clauses, for some prime p (this results in at most quadratic blowup over the original instance). Define a regular language L over $\{0, 1\}$ as follows. An input $\vec{x}_1 \cdots \vec{x}_p$, where $\vec{x}_i \in \{0, 1\}^p$, is *not* in L if $\vec{x}_i$ is a satisfying assignment for clause i. One can easily construct a DFA for L with $p^2 + 2$ states.

We claim that the SAT instance is satisfiable if and only if L is not +-closed. Indeed, suppose that the instance is satisfied by some assignment $\vec{x}$. Then $\vec{x} \in L$ whereas $\vec{x}^p \notin L$. Conversely, suppose that L is not +-closed. Then there is some $w \in L$ such that $w^n \notin L$ for some $n > 1$. Thus $|w^n| = p^2$, and so either $|w| = 1$ or $|w| = p$; in the former case, replace w by w^p, which is also a witness. The word w then represents a satisfying assignment for the SAT instance. □

Open Question 3. *Determine the complexity of deciding +-closedness.*