Parikh's theorem

Canonical name	ParikhsTheorem
Date of creation	2013-03-22 18:55:49
Last modified on	2013-03-22 18:55:49
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	9
Author	CWoo (3771)
Entry type	Theorem
Classification	msc 68Q42
Classification	msc 68Q45
Defines	Parikh mapping
Defines	linear set
Defines	semilinear set
Defines	letter-equivalent
Defines	slip-language

Let Σ be an alphabet. Fix an order on elements of Σ, so that they are $a_1, \ldots, a_n$. For each word w over Σ, we can form an n-tuple $(w_1, \ldots, w_n)$ so that each w_i is the number of occurrences of a_i in w. The function $\Psi : \Sigma^* \to \mathbb{N}^n$ such that

$$\Psi(w) = (w_1, \ldots, w_n)$$

is called the *Parikh mapping* (over the alphabet Σ). Here, $\mathbb{N}$ is the set of non-negative integers.

Two languages L_1 and L_2 over some Σ are called *letter-equivalent* if $\Psi(L_1) = \Psi(L_2)$.

For example, $L_1 := \{a^n b^n \mid n \geq 0\}$ is letter-equivalent to the $L_2 := \{(ab)^n \mid n \geq 0\}$, as both are mapped to the set $\{(n, n) \mid n \geq 0\}$ by Ψ. Note that L_1 is context-free, while L_2 is regular. In fact, we have the following:

Theorem 1 (Parikh)**.** *Every context-free language is letter-equivalent to a regular language.*

Like the pumping lemma, Parikh's theorem can be used to show that certain languages are not context-free. But in order to apply Parikh's theorem above, one needs to know more about the structure of the set $\Psi(L)$ for a given context-free language L.

First, note that $\mathbb{N}^n$ is a commutative monoid under addition. For any element x and subset S of $\mathbb{N}^n$, define $x + S$ in the obvious manner. Next, let $x_1, \ldots, x_m$ be elements of $\mathbb{N}^n$. Denote

$$\langle x_1, \ldots, x_m \rangle$$

the submonoid generated by $x_1, \ldots, x_m$.

A subset S of $\mathbb{N}^n$ is said to be *linear* if it has the form

$$x_0 + \langle x_1, \ldots, x_m \rangle,$$

for some $x_0, \ldots, x_m \in \mathbb{N}^n$. It is not hard to see that every linear set is the image of some regular language under Ψ. For example,

$$(1, 0, 0) + \langle (2, 1, 0), (0, 3, 2) \rangle = \Psi(\{a(a^2b)^m(b^3c^2)^n \mid m, n \geq 0\}).$$

The example can be easily generalized.

A subset of $\mathbb{N}^n$ is said to be *semilinear* if it is the union of finitely many linear sets. Since regular languages are closed under union, every semilinear set is the image of a regular language under Ψ. As a result, Parikh's theorem can be restated as

Theorem 2 (Theorem 1 restated)**.** *If L is context-free, then $\Psi(L)$ is semi-linear.*

Using this theorem, one easily sees that the language $\{a^p \mid p \text{ is a prime number }\}$ is not context-free, and neither is the language

$$\{a^m b^n \mid \text{ either } m \text{ is prime, and } n \geq m, \text{ or } n \text{ is prime, and } m \geq n\},$$

which can not be proved by the pumping lemma in a straightforward manner.

Remarks.

- The converse of Theorem 2 above is false. For example, let $L = \{a^n b^n c^n \mid n \geq 0\}$. Then L is not context-free by the pumping lemma. However, $\Psi(L) = \langle (1,1,1) \rangle$, which is clearly linear.
- In the literature, a lanugage L such that $\Psi(L)$ is semilinear is often called a *slip-language*. It can be shown that for a slip-language L over an alphabet consisting of two symbols, the language $\Psi^{-1} \circ \Psi(L)$ is context-free.

References

[1] H.R. Lewis, C.H. Papadimitriou, *Elements of the Theory of Computation.* Prentice-Hall, Englewood Cliffs, New Jersey (1981).

[2] D. C. Kozen, *Automata and Computability*, Springer, New York (1997).

[3] R.J. Parikh, *On Context-Free Languages, Journal of the Association for Computing Machinery, 4* (1966), 570-581.

[4] M. Latteux, *Cônes rationnels commutatifs*, J. Comput. Systems Sci. 18 (1979), 307-333.