

Generalized context-freeness

Joost Winter

Centrum Wiskunde & Informatica

January 23, 2012

Overview

- ▶ In a paper published at CALCO 2011, we presented a coalgebraic treatment of context-free languages.
- ▶ Formal languages can be seen as a specific instance of *formal power series* over a semiring, here over the Boolean semiring \mathbb{B} .
- ▶ Two different definitions of *algebraic* power series exist. We can easily generalize our approach to context-free languages to obtain these algebraic power series.

Formal power series, formal languages, and streams

- ▶ Given a finite set A called the *alphabet*, and a semiring R , a *formal power series* on A with coefficients in R is a function

$$A^* \rightarrow R.$$

- ▶ When R is the Boolean semiring $\{0, 1\}$ (with $1 + 1 = 1$), formal power series on A with coefficients in R correspond to formal languages over the alphabet A .
- ▶ When A is a singleton set, formal power series on A with coefficients on R correspond to *streams* over R .

Coalgebras representing formal power series (1)

In the this talk, we will represent these power series with coalgebras over functors of the type $R \times (-)^A$.

Coalgebras representing formal power series (2)

Given a coalgebra (X, f) over the functor $R \times (-)^A$, we will use the following notation:

- ▶ Given $x \in X$, $o(x)$ (called the *output value of x*) will be the first component of $f(x)$.
- ▶ Given $x \in X$, x_a (called the *a -derivative of x*) will be the second component of $f(x)$, applied to a .

So: for every x , $o(x)$ is an element of the semiring R , and for every x and a , x_a is an element of X again.

When (in the case of streams) the alphabet is a singleton set $\{a\}$, we usually write x' instead of x_a .

Coalgebras representing formal power series (3)

We can extend the notion of derivatives from alphabet symbols (i.e. elements of A) to words (i.e. elements of A^*) inductively:

- ▶ $x_\lambda = x$
- ▶ $x_{a \cdot w} = (x_a)_w$.

Homomorphisms and bisimulations (1)

Given two $R \times (-)^A$ -coalgebras (X, f) and (Y, g) , a function $h : X \rightarrow Y$ is a *homomorphism* if the following hold:

1. For every $x \in X$, $o(x) = o(h(x))$.
2. For every $x \in X$ and $a \in A$, $h(x_a) = (h(x))_a$.

Homomorphisms and bisimulations (2)

Given two $R \times (-)^A$ -coalgebras (X, f) and (Y, g) , a relation $R \subseteq X \times Y$ is a *bisimulation* if the following hold:

1. If $(x, y) \in R$, then $o(x) = o(y)$.
2. If $(x, y) \in R$, then for all $a \in A$, $(x_a, y_a) \in R$.

Final coalgebras (1)

A final coalgebra for the functor $2 \times (-)^A$ is the coalgebra (\mathcal{L}, l) , defined as follows:

- ▶ \mathcal{L} is the set of all languages on the alphabet A .
- ▶ For any $L \in \mathcal{L}$:
 - ▶ $o(L)$ is 1 iff the empty word is in L .
 - ▶ $F_a = \{w \mid a \cdot w \in L\}$.

Given a $2 \times (-)^A$ -coalgebra (X, f) , and an element $x \in X$, we let $\llbracket x \rrbracket$ denote the value of x under the unique homomorphism from (X, f) to the final coalgebra.

Final coalgebras (2)

Generalizing the picture from the previous slide to arbitrary semirings R , the final coalgebras will be sets of formal power series:

- ▶ \mathcal{S} is the set of all formal power series on A with coefficients in R , i.e. the function space from A^* to R .
- ▶ For any $F \in \mathcal{S}$:
 - ▶ $o(F) = F(\lambda)$.
 - ▶ $F_a = G : G(x) = F(a \cdot x)$.

Again, we let $\llbracket x \rrbracket$ denote the value of x under the final homomorphism.

Languages and streams

The presentation above is a generalization of both languages and streams:

- ▶ When we choose the Boolean semiring $\{0, 1\}$, we get coalgebras representing languages over the alphabet A .
- ▶ When we choose a singleton alphabet A , we get coalgebras representing streams over the semiring R .

The coalgebra of rational series (1)

The set \mathcal{E} of *rational series* over a finite alphabet A and a semiring $(R, +, \cdot, 0, 1)$ can be defined as follows:

$$t ::= a \in A \mid r \in R \mid t + t \mid t \cdot t \mid t^*$$

We can assign a \mathcal{D} -coalgebra structure to this set of regular expressions by specifying the output values and derivatives for each expression, giving us a \mathcal{D} -coalgebra (\mathcal{E}, e) :

Coalgebras of rational series (2)

t	$o(t)$	t_a
$r \in R$	r	0
$b \in A$	0	if $b = a$ then 1 else 0
$u + v$	$o(u) + o(v)$	$u_a + v_a$
$u \cdot v$	$o(u) \cdot o(v)$	$u_a \cdot v + o(u) \cdot v_a$
u^*	1	$u_a \cdot u^*$

When the underlying semiring R is the Boolean semiring 2 , these rational series are exactly the regular languages.

Introduction: context-free languages and systems

- ▶ The 'next step up' from regular expressions and languages, and finite automata, in the Chomsky hierarchy, are the context-free languages and grammars, and pushdown automata.
- ▶ We will present a format of coinductively defined systems of equations.
- ▶ When using \mathbb{B} , it turns out that these systems of equations characterize precisely the context-free languages.
- ▶ For arbitrary semirings, these systems of equations correspond to so-called *algebraic power series*

Systems of equations (1)

We will use terms t specified as follows:

$$t ::= a \in A \mid x \in X \mid r \in R \mid t + t \mid t \cdot t$$

where X is a finite set of variables, R is the carrier of the underlying semiring, and A , as before, is a finite alphabet. Given X , we let TX denote the set of terms over X .

A well-formed system of equations, for a set of variables X , consists of:

1. For every $x \in X$, exactly one equation of the form $o(x) = v$, where $v \in \{0, 1\}$.
2. For every $x \in X$ and $a \in A$, exactly one equation of the form $x_a = t$, where $t \in TX$.

Systems of equations (2)

Alternatively, we can consider a well-formed system of equations as a mapping

$$f : X \rightarrow R \times TX^A$$

We can extend such a mapping f to the \mathcal{D} -coalgebra (TX, \bar{f}) generated by (X, f) as follows:

t	$o(t)$	t_a
$x \in X$	$o(x)$	x_a (as specified by f)
$r \in R$	r	0
$b \in A$	0	if $b = a$ then 1 else 0
$u + v$	$o(u) + o(v)$	$u_a + v_a$
$u \cdot v$	$o(u) \cdot o(v)$	$u_a \cdot v + o(u) \cdot v_a$

Systems of equations (3)

The construction can be summarized as follows:

$$\begin{array}{ccccc}
 X \subset & \xrightarrow{i} & TX & \xrightarrow{\quad} & A^* \rightarrow R \\
 \downarrow f & & \swarrow \text{!} & \begin{array}{c} \llbracket \\ \rrbracket \end{array} & \downarrow \\
 R \times TX^A & \xrightarrow{\quad} & & & R \times (A^* \rightarrow R)^A
 \end{array}$$

Systems of equations (3)

When taking \mathbb{B} as underlying semiring, we obtain:

Proposition: *A language L is context-free iff there is a well-formed system of equations (X, f) and an $x \in X$, such that $\llbracket x \rrbracket = L$ w.r.t. the coalgebra (TX, \bar{f}) generated by it.*

First definition (1)

(The following definition can be found in work by Salomaa, Droste, Kuich et al., and seems to originate with Michel Fließ in the 70s.)
An R -algebraic system is a set of equations of the form

$$x_i = p_i \quad (i \in \{1, \dots, n\})$$

where $p_i \in R\langle A \cup X \rangle$. Such a system is called *proper* if, for all i , $(p_i, \lambda) = 0$, and for all i, j , $(p_i, y_j) = 0$. Furthermore, such a system is said to be in *Greibach normal form* if its support is contained in the set $A \cup AX \cup AXX$.

First definition (2)

A *strong solution to an R -algebraic system* is an assignment of power series to variables x_i , such that for each x_i , $o(x_i) = 0$. We call a formal power series s *constructively R -algebraic* if

$$s = r + \bar{s},$$

where $r \in R$ and \bar{s} is a strong solution to an R -algebraic system.

Results w.r.t. the first definition

- ▶ Every proper R -algebraic system has exactly one strong solution.
- ▶ Any component of a strong solution to a proper R -algebraic system also occurs as a strong solution to such a system in Greibach normal form.
- ▶ A formal power series is R -algebraic iff it occurs as a solution to a context-free system of equations.

Second definition

For a field F , a stream σ over F is F -algebraic if there are polynomial streams a_0, \dots, a_n such that

$$\sum_{0 \leq i \leq n} a_i \sigma^i = 0$$

This definition occurs in the world of *automatic sequences*: it is well-known that, over a finite field \mathbb{F}_p , a sequence is p -automatic iff it is \mathbb{F}_p -algebraic.

Correspondence between the two definitions

- ▶ Proposition: *Given a stream σ over a perfect field F , σ is F -algebraic iff it is constructively F -algebraic (Fliess 1971)*
- ▶ Proposition: *A power series s is constructively R -algebraic iff there is a well-formed system of equations (X, f) and an $x \in X$, such that $\llbracket x \rrbracket = s$ w.r.t. the coalgebra (TX, \bar{f}) generated by it.*

Examples: context-free streams (1)

When we take the semiring of natural numbers as underlying semiring, the Catalan numbers

$$1, 1, 2, 5, 14, 42, 132, 429, 1430, \dots$$

are context-free, and are generated by the following system of equations:

$$o(x) = 1 \quad x' = x \cdot x$$

Examples: context-free streams (2)

When we take the Boolean *field* \mathbb{F}_2 (with $1 + 1 = 0$) as underlying semiring, the Prouhet-Thue-Morse sequence

1001011001101001...

is context-free, and is generated by the following system of equations:

$$\begin{aligned}o(x) &= 1 & x' &= y \\o(y) &= 0 & y' &= z \\o(z) &= 0 & z' &= x + y + z + z \cdot w \\o(w) &= 1 & w' &= y \cdot w\end{aligned}$$

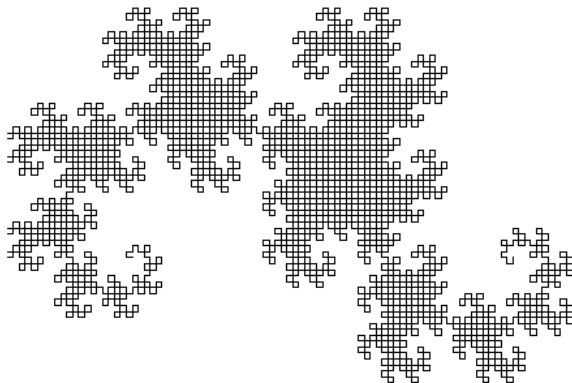
Examples: context-free streams (3)

Still using \mathbb{F}_2 as underlying semiring, the context-free system of equations

$$\begin{aligned}o(x) &= 1 & x' &= y \\o(y) &= 1 & y' &= z \\o(z) &= 0 & z' &= w \\o(w) &= 1 & w' &= v \\o(v) &= 1 & v' &= v + w + v \cdot x + x \cdot x\end{aligned}$$

gives us the paperfolding sequence, or Dragon curve sequence...

Examples: context-free streams (4)



Conclusions and further work

- ▶ We have a neat correspondence...
- ▶ ... but still want to gain a better understanding of Fließ's 1971 proof...
- ▶ ... and find a constructive way of transforming the two definitions of algebraicity into each other.
- ▶ Can we say anything about algebraic reals? (See: coalgebraic work on exact real arithmetic.)